Christopher C. Heyde

Quasi-Likelihood And Its Application

A General Approach to Optimal Parameter Estimation

Springer

Christopher C. Heyde
Australian National University
Stochastic Analysis Program, SMS
Canberra, ACT 0200
Australia
and
Department of Statistics
Columbia University
New York, NY 10027
USA

Library of Congress Cataloging in Publication Data
Heyde, C. C.
Quasi-likelihood and its application : a general approach to optimal parameter estimation / Christopher C. Heyde.
p. cm. — (Springer series in statistics)
Includes bibliographical references and index.
1. Parameter estimation. I. Title. II. Series.
QA276.8.H475 1997
519.5′44—dc21 97-11168

ISBN 0-387-98225-6 Springer-Verlag New York Berlin Heidelberg SPIN 10573429

Preface

This book is concerned with the general theory of optimal estimation of parameters in systems subject to random effects and with the application of this theory. The focus is on choice of families of estimating functions, rather than the estimators derived therefrom, and on optimization within these families. Only assumptions about means and covariances are required for an initial discussion. Nevertheless, the theory that is developed mimics that of maximum likelihood, at least to the first order of asymptotics.

The term quasi-likelihood has often had a narrow interpretation, associated with its application to generalized linear model type contexts, while that of optimal estimating functions has embraced a broader concept. There is, however, no essential distinction between the underlying ideas and the term quasi-likelihood has herein been adopted as the general label. This emphasizes its role in extension of likelihood based theory. The idea throughout involves finding quasi-scores from families of estimating functions. Then, the quasi-likelihood estimator is derived from the quasi-score by equating to zero and solving, just as the maximum likelihood estimator is derived from the likelihood score.

This book had its origins in a set of lectures given in September 1991 at the 7th Summer School on Probability and Mathematical Statistics held in Varna, Bulgaria, the notes of which were published as Heyde (1993). Subsets of the material were also covered in advanced graduate courses at Columbia University in the Fall Semesters of 1992 and 1996. The work originally had a quite strong emphasis on inference for stochastic processes but the focus gradually broadened over time. Discussions with V.P. Godambe and with R. Morton have been particularly influential in helping to form my views.

The subject of estimating functions has evolved quite rapidly over the period during which the book was written and important developments have been emerging so fast as to preclude any attempt at exhaustive coverage. Among the topics omitted is that of quasi- likelihood in survey sampling, which has generated quite an extensive literature (see the edited volume Godambe (1991), Part 4 and references therein) and also the emergent linkage with Bayesian statistics (e.g., Godambe (1994)). It became quite evident at the Conference on Estimating Functions held at the University of Georgia in March 1996 that a book in the area was much needed as many known ideas were being rediscovered. This realization provided the impetus to round off the project rather earlier than would otherwise have been the case.

The emphasis in the monograph is on concepts rather than on mathematical theory. Indeed, formalities have been suppressed to avoid obscuring "typical" results with the phalanx of regularity conditions and qualifiers necessary to avoid the usual uninformative types of counterexamples which detract from most statistical paradigms. In discussing theory which holds to the first order of asymptotics the treatment is especially informal, as befits the context. Sufficient conditions which ensure the behaviour described are not difficult to furnish but are fundamentally uninlightening.

A collection of complements and exercises has been included to make the material more useful in a teaching environment and the book should be suitable for advanced courses and seminars. Prerequisites are sound basic courses in measure theoretic probability and in statistical inference.

Comments and advice from students and other colleagues has also contributed much to the final form of the book. In addition to V.P. Godambe and R. Morton mentioned above, grateful thanks are due in particular to Y.-X. Lin, A. Thavaneswaran, I.V. Basawa, E. Saavendra and T. Zajic for suggesting corrections and other improvements and to my wife Beth for her encouragement.

C.C. Heyde

Canberra, Australia
February 1997

Contents

Preface **v**

1 Introduction **1**
1.1 The Brief 1
1.2 Preliminaries 1
1.3 The Gauss-Markov Theorem 3
1.4 Relationship with the Score Function 6
1.5 The Road Ahead 7
1.6 The Message of the Book 10
1.7 Exercise 10

2 The General Framework **11**
2.1 Introduction 11
2.2 Fixed Sample Criteria 11
2.3 Scalar Equivalences and Associated Results 19
2.4 Wedderburn's Quasi-Likelihood 21
2.4.1 The Framework 21
2.4.2 Limitations 23
2.4.3 Generalized Estimating Equations 25
2.5 Asymptotic Criteria 26
2.6 A Semimartingale Model for Applications 30
2.7 Some Problem Cases for the Methodology 35
2.8 Complements and Exercises 38

3 An Alternative Approach: E-Sufficiency **43**
3.1 Introduction 43
3.2 Definitions and Notation 43
3.3 Results 46
3.4 Complement and Exercise 51

4 Asymptotic Confidence Zones of Minimum Size **53**
4.1 Introduction 53
4.2 The Formulation 54
4.3 Confidence Zones: Theory 56

4.4 Confidence Zones: Practice 60
4.5 On Best Asymptotic Confidence Intervals 62
4.5.1 Introduction and Results 62
4.5.2 Proof of Theorem 4.1 64
4.6 Exercises 67

5 Asymptotic Quasi-Likelihood 69
5.1 Introduction 69
5.2 The Formulation 71
5.3 Examples 79
5.3.1 Generalized Linear Model 79
5.3.2 Heteroscedastic Autoregressive Model 79
5.3.3 Whittle Estimation Procedure 82
5.3.4 Addendum to the Example of Section 5.1 87
5.4 Bibliographic Notes 88
5.5 Exercises 88

6 Combining Estimating Functions 91
6.1 Introduction 91
6.2 Composite Quasi-Likelihoods 92
6.3 Combining Martingale Estimating Functions 93
6.3.1 An Example 98
6.4 Application. Nested Strata of Variation 99
6.5 State-Estimation in Time Series 103
6.6 Exercises 104

7 Projected Quasi-Likelihood 107
7.1 Introduction 107
7.2 Constrained Parameter Estimation 107
7.2.1 Main Results 109
7.2.2 Examples 111
7.2.3 Discussion 112
7.3 Nuisance Parameters 113
7.4 Generalizing the E-M Algorithm: The P-S Method 116
7.4.1 From Log-Likelihood to Score Function 117
7.4.2 From Score to Quasi-Score 118
7.4.3 Key Applications 121
7.4.4 Examples 122
7.5 Exercises 127

8 Bypassing the Likelihood 129
8.1 Introduction 129
8.2 The REML Estimating Equations 129
8.3 Parameters in Diffusion Type Processes 131
8.4 Estimation in Hidden Markov Random Fields 136
8.5 Exercise 139

9 Hypothesis Testing 141
9.1 Introduction 141
9.2 The Details 142
9.3 Exercise 145

10 Infinite Dimensional Problems 147
10.1 Introduction 147
10.2 Sieves 147
10.3 Semimartingale Models 148

11 Miscellaneous Applications 153
11.1 Estimating the Mean of a Stationary Process 153
11.2 Estimation for a Heteroscedastic Regression 159
11.3 Estimating the Infection Rate in an Epidemic 162
11.4 Estimating Population Size 164
11.5 Robust Estimation 169
11.5.1 Optimal Robust Estimating Functions 170
11.5.2 Example 173
11.6 Recursive Estimation 176

12 Consistency and Asymptotic Normality for Estimating Functions 179
12.1 Introduction 179
12.2 Consistency 180
12.3 The SLLN for Martingales 186
12.4 The CLT for Martingales 190
12.5 Exercises 195

13 Complements and Strategies for Application 199
13.1 Some Useful Families of Estimating Functions 199
13.1.1 Introduction 199
13.1.2 Transform Martingale Families 199
13.1.3 Use of the Infinitesimal Generator of a Markov Process 200
13.2 Solution of Estimating Equations 201
13.3 Multiple Roots 202
13.3.1 Introduction 202
13.3.2 Examples 204
13.3.3 Theory 208
13.4 Resampling Methods 210

References 211

Index 227

Chapter 1

Introduction

1.1 The Brief

This monograph is primarily concerned with parameter estimation for a random process $\{\boldsymbol{X}_t\}$ taking values in r-dimensional Euclidean space. The distribution of $\boldsymbol{X}_t$ depends on a characteristic $\boldsymbol{\theta}$ taking values in a open subset Θ of p-dimensional Euclidean space. The framework may be parametric or semiparametric; $\boldsymbol{\theta}$ may be, for example, the mean of a stationary process. The object will be the "efficient" estimation of $\boldsymbol{\theta}$ based on a sample $\{\boldsymbol{X}_t,\ t \in T\}$.

1.2 Preliminaries

Historically there are two principal themes in statistical parameter estimation theory:

least squares (LS) - introduced by Gauss and Legendre and founded on finite sample considerations (minimum distance interpretation)

maximum likelihood (ML) - introduced by Fisher and with a justification that is primarily asymptotic (minimum size asymptotic confidence intervals, ideas of which date back to Laplace)

It is now possible to unify these approaches under the general description of quasi-likelihood and to develop the theory of parameter estimation in a very general setting. The fixed sample optimality ideas that underly quasi-likelihood date back to Godambe (1960) and Durbin (1960) and were put into a stochastic process setting in Godambe (1985). The asymptotic justification is due to Heyde (1986). The ideas were combined in Godambe and Heyde (1987).

It turns out that the theory needs to be developed in terms of estimating functions (functions of both the data and the parameter) rather than the estimators themselves. Thus, our focus will be on functions that have the value of the parameter as a root rather than the parameter itself.

The use of estimating functions dates back at least to K. Pearson's introduction of the method of moments (1894) although the term "estimating function" may have been coined by Kimball (1946). Furthermore, all the standard methods of estimation, such as maximum likelihood, least-squares, conditional least-squares, minimum chi-squared, and M-estimation, are included under minor regularity conditions. The subject has now developed to the stage where books are being devoted to it, e.g., Godambe (1991), McLeish and Small (1988).

The rationale for the use of the estimating function rather than the estimator derived therefrom lies in its more fundamental character. The following dot points illustrate the principle.

- Estimating functions have the property of invariance under one-to-one transformations of the parameter θ.

- Under minor regularity conditions the score function (derivative of the log-likelihood with respect to the parameter), which is an estimating function, provides a minimal sufficient partitioning of the sample space. However, there is often no single sufficient statistic.

  For example, suppose that $\{Z_t\}$ is a Galton-Watson process with offspring mean $E(Z_1 \mid Z_0 = 1) = \theta$. Suppose that the offspring distribution belongs to the power series family (which is the discrete exponential family). Then, the score function is

  $$U_T(\theta) = c \sum_{t=1}^{T} (Z_t - \theta\, Z_{t-1}),$$

  where c is a constant and the maximum likelihood estimator

  $$\hat{\theta}_T = \sum_{t=1}^{T} Z_t \bigg/ \sum_{t=1}^{T} Z_{t-1}$$

  is not a sufficient statistic. Details are given in Chapter 2.

- Fisher's information is an estimating function property (namely, the variance of the score function) rather than that of the maximum likelihood estimator (MLE).

- The Cramér-Rao inequality is an estimating function property rather than a property of estimators. It gives the variance of the score function as a bound on the variances of standardized estimating functions.

- The asymptotic properties of an estimator are almost invariably obtained, as in the case of the MLE, via the asymptotics of the estimating function and then transferred to the parameter space via local linearity.

- Separate estimating functions, each with information to offer about an unknown parameter, can be combined much more readily than the estimators therefrom.

We shall begin our discussion by examining the minimum variance ideas that underly least squares and then see how optimality is conveniently phrased in terms of estimating functions. Subsequently, we shall show how the score function and maximum likelihood ideas mesh with this. The approach is along the general lines of the brief overviews that appear in Godambe and Heyde (1987), Heyde (1989b), Desmond (1991), Godambe and Kale (1991). An earlier version appeared in the lecture notes Heyde (1993). Another approach to the subject of optimal estimation, which also uses estimating functions but is based on extension of the idea of sufficiency, appears in McLeish and Small (1988); the theories do substantially overlap, although this is not immediately transparent. Details are provided in Chapter 3.

1.3 Estimating Functions and the Gauss-Markov Theorem

To indicate the basic LS ideas that we wish to incorporate, we consider the simplest case of independent random variables (rv's) and a one-dimensional parameter θ. Suppose that $X_1, \ldots, X_T$ are independent rv's with $EX_t = \theta$, $\operatorname{var} X_t = \sigma^2$. In this context the Gauss-Markov theorem has the following form.

GM Theorem: Let the estimator $S_T = \sum_{t=1}^T a_t X_t$ be unbiased for θ, the a_t being constants. Then, the variance, $\operatorname{var} S_T$, is minimized for $a_t = 1/T$, $t = 1, \ldots, T$. That is, the sample mean $\bar{X} = T^{-1} \sum_{t=1}^T X_t$ is the linear unbiased minimum variance estimator of θ.

The proof is very simple; we have to minimize $\operatorname{var} S_T = \sigma^2 \sum_{t=1}^T a_t^2$ subject to $\sum_{t=1}^T a_t = 1$ and

$$\begin{aligned}\operatorname{var} S_T &= \sigma^2 \sum_{t=1}^T \left(a_t^2 - \frac{2a_t}{T} + \frac{1}{T^2} \right) + \frac{\sigma^2}{T} \\ &= \sigma^2 \sum_{t=1}^T \left(a_t - \frac{1}{T} \right)^2 + \frac{\sigma^2}{T} \geq \frac{\sigma^2}{T}.\end{aligned}$$

Now we can restate the GM theorem in terms of estimating functions. Consider the set $\mathcal{G}_0$ of unbiased estimating functions $G = G(X_1, \ldots, X_T, \theta)$ of the form $G = \sum_{t=1}^T b_t(X_t - \theta)$, the b_t's being constants with $\sum_{t=1}^T b_t \neq 0$.

Note that the estimating functions kG, k constant and G produce the same estimator, namely $\sum_{t=1}^T b_t X_t / \sum_{t=1}^T b_t$, so some standardization is necessary if variances are to be compared.

One possible standardization is to define the standardized version of G as

$$G^{(s)} = \left(\sum_{t=1}^T b_t \right) \sum_{t=1}^T b_t \left(X_t - \theta \right) \left(\sigma^2 \sum_{t=1}^T b_t^2 \right)^{-1} .$$

The estimator of θ is unchanged and, of course, kG and G have the same standardized form. Let us now motivate this standardization.

(1) In order to be used as an estimating equation, the estimating function G

needs to be as close to zero as possible when θ is the true value. Thus we want $\operatorname{var} G = \sigma^2 \sum_{t=1}^{T} b_t^2$ to be as small as possible. On the other hand, we want $G(\theta + \delta\theta)$, $\delta > 0$, to differ as much as possible from $G(\theta)$ when θ is the true value. That is, we want $(E\dot{G}(\theta))^2 = \left(\sum_{t=1}^{T} b_t\right)^2$, the dot denoting derivative with respect to θ, to be as large as possible. These requirements can be combined by maximizing $\operatorname{var} G^{(s)} = (E\dot{G})^2/EG^2$.

(2) Also, if $\max_{1\le t\le T} b_t / \sum_{t=1}^{T} b_t \to 0$ as $T \to \infty$, then

$$\sum_{t=1}^{T} b_t\,(X_t - \theta) \Big/ \left(\sigma^2 \sum_{t=1}^{T} b_t^2\right)^{1/2} \xrightarrow{\mathrm{d}} N(0,1)$$

using the Lindeberg-Feller central limit theorem. Thus, noting that our estimator for θ is

$$\hat{\theta}_T = \sum_{t=1}^{T} b_t\, X_t \Big/ \sum_{t=1}^{T} b_t,$$

we have

$$\left(\operatorname{var} G_T^{(s)}\right)^{1/2} (\hat{\theta}_T - \theta) \xrightarrow{\mathrm{d}} N(0,1),$$

i.e.,

$$\hat{\theta}_T - \theta \overset{\mathrm{d}}{\simeq} N\left(0, \left(\operatorname{var} G_T^{(s)}\right)^{-1}\right).$$

We would wish to choose the best asymptotic confidence intervals for θ and hence to maximize $\operatorname{var} G_T^{(s)}$.

(3) For the standardized version $G^{(s)}$ of G we have

$$\operatorname{var} G^{(s)} = \left(\sum_{t=1}^{T} b_t\right)^2 \Big/ \sigma^2 \sum_{t=1}^{T} b_t^2 = -E\left(\dot{G}^{(s)}\right),$$

i.e., $G^{(s)}$ possesses the standard likelihood score property.

Having introduced standardization we can say that $G^* \in \mathcal{G}_0$ is *an optimal estimating function within* $\mathcal{G}_0$ if $\operatorname{var} G^{*(s)} \ge \operatorname{var} G^{(s)}$, $\forall G \in \mathcal{G}_0$. This leads to the following result.

GM Reformation The estimating function $G^* = \sum_{t=1}^{T}(X_t - \theta)$ is an optimal estimating function within $\mathcal{G}_0$. The estimating equation $G^* = 0$ provides the sample mean as an optimal estimator of θ.

The proof follows immediately from the Cauchy-Schwarz inequality. For $G \in \mathcal{G}_0$ we have

$$\operatorname{var} G^{(s)} = \left(\sum_{t=1}^{T} b_t \right)^2 \Big/ \sigma^2 \sum_{t=1}^{T} b_t^2 \leq T/\sigma^2 = \operatorname{var} G^{*(s)}$$

and the argument holds even if the b_t's are functions of θ.

Now the formulation that we adapted can be extended to estimating functions G in general by defining the standardized version of G as

$$G^{(s)} = -(E\dot{G})\,(EG^2)^{-1}\,G.$$

Optimality based on maximization of $\operatorname{var} G^{(s)}$ leads us to define G^* to be optimal within a class $\mathcal{H}$ if

$$\operatorname{var} G^{*(s)} \geq \operatorname{var} G^{(s)}, \qquad \forall G \in \mathcal{H}.$$

That this concept does differ from least squares in some important respects is illustrated in the following example.

We now suppose that X_t, $t = 1, 2, \ldots, T$ are independent rv's with $EX_t = \alpha_t(\theta)$, $\operatorname{var} X_t = \sigma_t^2(\theta)$, the α_t's, σ_t^2's being specified differentiable functions. Then, for the class of estimating functions

$$\mathcal{H} = \left\{ H : \; H = \sum_{t=1}^{T} b_t(\theta)\,(X_t - \alpha_t(\theta)) \right\},$$

we have

$$\operatorname{var} H^{(s)} = \left(\sum_{t=1}^{T} b_t(\theta)\,\dot{\alpha}_t(\theta) \right)^2 \Big/ \sum_{t=1}^{T} b_t^2(\theta)\,\sigma_t^2(\theta),$$

which is maximized (again using the Cauchy-Schwarz inequality) if

$$b_t(\theta) = k(\theta)\,\dot{\alpha}_t(\theta)\,\sigma_t^{-2}(\theta), \qquad t = 1, 2, \ldots, T,$$

$k(\theta)$ being an undetermined multiplier. Thus, an optimal estimating function is

$$H^* = \sum_{t=1}^{T} \dot{\alpha}_t(\theta)\,\sigma_t^{-2}(\theta)\,(X_t - \alpha_t(\theta)).$$

Note that this result is not what one gets from least squares (LS). If we applied LS, we would minimize

$$\sum_{t=1}^{T} (X_t - \alpha_t(\theta))^2\,\sigma_t^{-2}(\theta),$$

which leads to the estimating equation

$$\sum_{t=1}^{T} \dot{\alpha}_t(\theta)\,\sigma_t^{-2}(\theta)\,(X_t - \alpha_t(\theta)) + \sum_{t=1}^{T} (X_t - \alpha_t(\theta))^2\,\sigma_t^{-3}(\theta)\,\dot{\sigma}_t(\theta) = 0.$$

This estimating equation will generally not be unbiased, and it may behave very badly depending on the σ_t's. It will not in general provide a consistent estimator.

1.4 Relationship with the Score Function

Now suppose that $\{X_t,\ t = 1, 2, \ldots, T\}$ has likelihood function

$$L = \prod_{t=1}^{T} f_t(X_t; \theta).$$

The score function in this case is a sum of independent rv's with zero means,

$$U = \frac{\partial \log L}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial \log f_t(X_t; \theta)}{\partial \theta}$$

and, when $H = \sum_{t=1}^{T} b_t(\theta)\,(X_t - \alpha_t(\theta))$, we have

$$E(U\,H) = \sum_{t=1}^{T} b_t(\theta)\, E\left(\frac{\partial \log f_t(X_t; \theta)}{\partial \theta} (X_t - \alpha_t(\theta))\right).$$

If the f_t's are such that integration and differentiation can be interchanged,

$$E\,\frac{\partial \log f_t(X_t; \theta)}{\partial \theta}\, X_t = \frac{\partial}{\partial \theta}\, EX_t = \dot{\alpha}_t(\theta),$$

so that

$$E(U\,H) = \sum_{t=1}^{T} b_t(\theta)\,\dot{\alpha}_t(\theta) = -E\dot{H}.$$

Also, using corr to denote correlation,

$$\begin{aligned} \operatorname{corr}^2(U, H) &= (E(U\,H))^2/(EU^2)(EH^2) \\ &= (\operatorname{var} H^{(s)})/EU^2, \end{aligned}$$

which is maximized if $\operatorname{var} H^{(s)}$ is maximized. That is, the choice of an optimal estimating function $H^* \in \mathcal{H}$ is giving an element of $\mathcal{H}$ that has maximum correlation with the generally unknown score function.

Next, for the score function U and $H \in \mathcal{H}$ we find that

$$\begin{aligned} E(H^{(s)} - U^{(s)})^2 &= \operatorname{var} H^{(s)} + \operatorname{var} U^{(s)} - 2E(H^{(s)}\,U^{(s)}) \\ &= EU^2 - \operatorname{var} H^{(s)}, \end{aligned}$$

since

$$U^{(s)} = U$$

and

$$E H^{(s)} U^{(s)} = \operatorname{var} H^{(s)}$$

when differentiation and integration can be interchanged. Thus $E(H^{(s)} - U^{(s)})^2$ is minimized when an optimal estimating function $H^* \in \mathcal{H}$ is chosen. This gives an optimal estimating function the interpretation of having minimum expected distance from the score function. Note also that

$$\operatorname{var} H^{(s)} \leq E U^2,$$

which is the Cramér-Rao inequality.

Of course, if the score function $U \in \mathcal{H}$, the methodology picks out U as optimal. In the case in question $U \in \mathcal{H}$ if and only if U is of the form

$$U = \sum_{t=1}^{T} b_t(\theta)\,(X_t - \alpha_t(\theta)),$$

that is,

$$\frac{\partial \log f(X_t;\theta)}{\partial \theta} = b_t(\theta)\,(X_t - \alpha_t(\theta)),$$

so that the X_t's are from an exponential family in linear form.

Classical quasi-likelihood was introduced in the setting discussed above by Wedderburn (1974). It was noted by Bradley (1973) and Wedderburn (1974) that if the X_t's have exponential family distributions in which the canonical statistics are linear in the data, then the score function depends on the parameters only through the means and variances. They also noted that the score function could be written as a weighted least squares estimating function. Wedderburn suggested using the exponential family score function even when the underlying distribution was unspecified. In such a case the estimating function was called a *quasi-score estimating function* and the estimator derived therefore a *quasi-likelihood estimator.*

The concept of optimal estimating functions discussed above conveniently subsumes that of quasi-score estimating functions in the Wedderburn sense, as we shall discuss in vector form in Chapter 2. We shall, however, in our general theory, take the names *quasi-score* and *optimal* for estimating functions to be essentially synonymous.

1.5 The Road Ahead

In the above discussion we have concentrated on the simplest case of independent random variables and a scalar parameter, but the basis of a general formulation of the quasi-likelihood methodology is already evident.

In Chapter 2, quasi-likelihood is developed in its general framework of a (finite dimensional) vector valued parameter to be estimated from vector valued data. Quasi-likelihood estimators are derived from quasi-score estimating functions whose selection involves maximization of a matrix valued information criterion in the partial order of non-negative definite matrices. Both fixed sample and asymptotic formulations are considered and the conditions under which they hold are shown to be substantially overlapping. Also, since matrix valued criteria are not always easy to work with, some scalar equivalences are formulated. Here there is a strong link with the theory of optimal experimental design.

The original Wedderburn formulation of quasi-likelihood in an exponential family setting is then described together with the limitations of its direct extension. Also treated is the closely related methodology of generalized estimating equations, developed for longitudinal data sets and typically using approximate covariance matrices in the quasi-score estimating function.

The basic formulation having been provided, it is now shown how a semimartingale model leads to a convenient class of estimating functions of wide applicability. Various illustrations are provided showing how to use these ideas in practice, and some discussion of problem cases is also given.

Chapter 3 outlines an alternative approach to optimal estimation using estimating functions via the concepts of E-sufficiency and E-ancillarity. Here E refers to expectation. This approach, due to McLeish and Small, produces results that overlap substantially with those of quasi-likelihood, although this is not immediately apparent. The view is taken in this book that quasi-likelihood methodology is more transparent and easier to apply.

Chapter 4 is concerned with asymptotic confidence zones. Under the usual sort of regularity conditions, quasi-likelihood estimators are associated with minimum size asymptotic confidence intervals within their prespecified spaces of estimating functions. Attention is given to the subtle question of whether to normalize with random variables or constants in order to obtain the smallest intervals. Random normings have some important advantages.

Ordinary quasi-likelihood theory is concerned with the case where the maximum information criterion holds exactly for fixed T or for each T as $T \to \infty$. Chapter 5 deals with the case where optimality holds only in a certain asymptotic sense. This may happen, for example, when a nuisance parameter is replaced by a consistent estimator thereof. The discussion focuses on situations where the properties of regular quasi-likelihood of consistency and possession of minimum size asymptotic confidence zones are preserved for the estimator.

Estimating functions from different sources can conveniently be added, and the issue of their optimal combination is addressed in Chapter 6. Various applications are given, including dealing with combinations of estimating functions where there are nested strata of variation and providing methods of filtering and smoothing in time series estimation. The well-known Kalman filter is a special case.

Chapter 7 deals with projection methods that are useful in situations where a standard application of quasi-likelihood is precluded. Quasi-likelihood approaches are provided for constrained parameter estimation, for estimation in the presence of nuisance parameters, and for generalizing the E-M algorithm for estimation where there are missing data.

In Chapter 8 the focus is on deriving the score function, or more generally quasi-score estimating function, without use of the likelihood, which may be difficult to deal with, or fail to exist, under minor perturbations of standard conditions. Simple quasi-likelihood derivations of the score functions are provided for estimating the parameters in the covariance matrix, where the distribution is multivariate normal (REML estimation), in diffusion type models, and in hidden Markov random fields. In each case these remain valid as quasi-score estimating functions under significantly broadened assumptions over those of a likelihood based approach.

Chapter 9 deals briefly with issues of hypothesis testing. Generalizations of the classical efficient scores statistic and Wald test statistic are treated. These are shown to usually be asymptotically χ^2 distributed under the null hypothesis and to have asymptotically, noncentral χ^2 distributions, with maximum noncentrality parameter, under the alternative hypothesis, when the quasi-score estimating function is used.

Chapter 10 provides a brief discussion of infinite dimensional parameter (function) estimation. A sketch is given of the method of sieves, in which the dimension of the parameter is increased as the sample size increases. An informal treatment of estimation in linear semimartingale models, such as occur for counting processes and estimation of the cumulative hazard function, is also provided.

A diverse collection of applications is given in Chapter 11. Estimation is discussed for the mean of a stationary process, a heteroscedastic regression, the infection rate of an epidemic, and a population size via a multiple recapture experiment. Also treated are estimation via robustified estimating functions (possibly with components that are bounded functions of the data) and recursive estimation (for example, for on-line signal processing).

Chapter 12 treats the issues of consistency and asymptotic normality of estimators. Throughout the book it is usually expected that these will ordinarily hold under appropriate regularity conditions. The focus here is on martingale based methods, and general forms of martingale strong law and central limit theorems are provided for use in particular cases. The view is taken that it is mostly preferable directly to check cases individually rather than to rely on general theory with its multiplicity of regularity conditions.

Finally, in Chapter 13 a number of complementary issues involved in the use of quasi-likelihood methods are discussed. The chapter begins with a collection of methods for generating useful families of estimating functions. Integral transform families and the use of the infinitesimal generator of a Markov process are treated. Then, the numerical solution of estimating equations is considered, and methods are examined for dealing with multiple roots when a scalar objective function may not be available. The final section is concerned with resampling methods for the provision of confidence intervals, in particular the jackknife and bootstrap.

1.6 The Message of the Book

For estimation of parameters, in stochastic systems of any kind, it has become increasingly clear that it is possible to replace likelihood based techniques by quasi-likelihood alternatives, in which only assumptions about means and variances are made, in order to obtain estimators. There is often little, if any, loss in efficiency, and all the advantages of weighted least squares methods are also incorporated. Additional assumptions are, of course, required to ensure consistency of estimators and to provide confidence intervals.

If it is available, the likelihood approach does provide a basis for benchmarking of estimating functions but not more than that. It is conjectured that everything that can be done via likelihoods has a corresponding quasi-likelihood generalization.

1.7 Exercise

1. Suppose $\{X_i,\ i = 1, 2, \ldots\}$ is a sequence of independent rv's, X_i having a Bernoulli distribution with $P(X_i = 1) = p_i = \frac{1}{2} + \theta\, a_i,\ \ P(X_i = 0) = 1 - p_i$, and $0 < a_i \downarrow 0$ as $i \to \infty$. Show that there is a consistent estimator of θ if and only if $\sum_{i=1}^{\infty} a_i^2 = \infty$. (Adaped from Dion and Ferland (1995).)

Chapter 2

The General Framework

2.1 Introduction

Let $\{\boldsymbol{X}_t,\ t \le T\}$ be a sample of discrete or continuous data that is randomly generated and takes values in r-dimensional Euclidean space. The distribution of $\boldsymbol{X}_t$ depends on a "parameter" $\boldsymbol{\theta}$ taking values in an open subset $\boldsymbol{\Theta}$ of p-dimensional Euclidean space and the object of the exercise is the estimation of $\boldsymbol{\theta}$.

We assume that the possible probability measures for $\boldsymbol{X}_t$ are $\{P_\theta\}$ a union (possibly uncountable) of families of parametric models, each family being indexed by $\boldsymbol{\theta}$ and that each $(\Omega, \mathcal{F}, P_\theta)$ is a complete probability space.

We shall focus attention on the class $\mathcal{G}$ of zero mean, square integrable estimating functions $\boldsymbol{G}_T = \boldsymbol{G}_T(\{\boldsymbol{X}_t,\ t \le T\},\ \boldsymbol{\theta})$, which are vectors of dimension p for which $E\boldsymbol{G}_T(\boldsymbol{\theta}) = \boldsymbol{0}$ for each P_θ and for which the p-dimensional matrices $E\dot{\boldsymbol{G}}_T = (E\,\partial G_{T,i}(\boldsymbol{\theta})/\partial\theta_j)$ and $E\boldsymbol{G}_T\boldsymbol{G}_T'$ are nonsingular, the prime denoting transpose. The expectations are always with respect to P_θ. Note that $\dot{\boldsymbol{G}}$ is the transpose of the usual derivative of $\boldsymbol{G}$ with respect to $\boldsymbol{\theta}$.

In many cases P_θ is absolutely continuous with respect to some σ-finite measure λ_T giving a density $p_T(\boldsymbol{\theta})$. Then we write $\boldsymbol{U}_T(\boldsymbol{\theta}) = p_T^{-1}(\boldsymbol{\theta})\dot{p}_T(\boldsymbol{\theta})$ for the score function, which we suppose to be almost surely differentiable with respect to the components of $\boldsymbol{\theta}$. In addition we will also suppose that differentiation and integration can be interchanged in $E(\boldsymbol{G}_T\boldsymbol{U}_T')$ and $E(\boldsymbol{U}_T\boldsymbol{G}_T')$ for $\boldsymbol{G}_T \in \mathcal{G}$.

The score function $\boldsymbol{U}_T$ provides, modulo minor regularity conditions, a minimal sufficient partitioning of the sample space and hence should be used for estimation if it is available. However, it is often unknown, or in semiparametric cases, does not exist. The framework here allows a focus on models in which the error distribution has only its first and second moment properties specified, at least initially.

2.2 Fixed Sample Criteria

In practice we always work with specified subsets of $\mathcal{G}$. Take $\mathcal{H} \subseteq \mathcal{G}$ as such a set. As motivated in the previous chapter, optimality within $\mathcal{H}$ is achieved by maximizing the covariance matrix of the standardized estimating functions $G_T^{(s)} = -(E\dot{\boldsymbol{G}}_T)'(E\boldsymbol{G}_T\boldsymbol{G}_T')^{-1}\boldsymbol{G}_T$, $\boldsymbol{G}_T \in \mathcal{H}$. Alternatively, if $\boldsymbol{U}_T$ exists, an optimal estimating function within $\mathcal{H}$ is one with minimum dispersion distance from $\boldsymbol{U}_T$. These ideas are formalized in the following definition and equivalence, which we shall call criteria for O_F-optimality (fixed sample optimality). Later we shall introduce similar criteria for optimality to hold for all (sufficiently large) sample sizes. Estimating functions that are optimal in either sense will be referred to as *quasi-score estimating functions* and the estimators that come from equating these to zero and solving as *quasi-likelihood estimators.*

O_F-optimality involves choice of the estimating function $\boldsymbol{G}_T$ to maximize, in the partial order of nonnegative definite (nnd) matrices (sometimes known as the Loewner ordering), the information criterion

$$\mathcal{E}(\boldsymbol{G}_T) = E(\boldsymbol{G}_T^{(s)}\boldsymbol{G}_T^{(s)\prime}) = (E\dot{\boldsymbol{G}}_T)'(E\boldsymbol{G}_T\boldsymbol{G}_T')^{-1}(E\dot{\boldsymbol{G}}_T),$$

which is a natural generalization of Fisher information. Indeed, if the score function $\boldsymbol{U}_T$ exists,

$$\mathcal{E}(\boldsymbol{U}_T) = (E\dot{\boldsymbol{U}}_T)'(E\boldsymbol{U}_T\boldsymbol{U}_T')^{-1}(E\dot{\boldsymbol{U}}_T) = E\boldsymbol{U}_T\boldsymbol{U}_T'$$

is the Fisher information.

Definition 2.1 $\boldsymbol{G}_T^* \in \mathcal{H}$ is an O_F-optimal estimating function within $\mathcal{H}$ if

$$\mathcal{E}(\boldsymbol{G}_T^*) - \mathcal{E}(\boldsymbol{G}_T) \tag{2.1}$$

is nonnegative definite for all $\boldsymbol{G}_T \in \mathcal{H}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and P_θ.

The term **Loewner optimality** is used for this concept in the theory of optimal experimental designs (e.g., Pukelsheim (1993, Chapter 4)).

In the case where the score function exists there is the following equivalent form to Definition 2.1 phrased in terms of minimizing dispersion distance.

Definition 2.2 $\boldsymbol{G}_T^* \in \mathcal{H}$ is an O_F-optimal estimating function within $\mathcal{H}$ if

$$E\left(\boldsymbol{U}_T^{(s)} - \boldsymbol{G}_T^{(s)}\right)\left(\boldsymbol{U}_T^{(s)} - \boldsymbol{G}_T^{(s)}\right)' - E\left(\boldsymbol{U}_T^{(s)} - \boldsymbol{G}_T^{*(s)}\right)\left(\boldsymbol{U}_T^{(s)} - \boldsymbol{G}_T^{*(s)}\right)' \tag{2.2}$$

is nonnegative definite for all $\boldsymbol{G}_T \in \mathcal{H}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and P_θ.

Proof of Equivalence We drop the subscript T for convenience. Note that

$$E\left(\boldsymbol{G}^{(s)}\boldsymbol{U}^{(s)\prime}\right) = -(E\dot{\boldsymbol{G}})'(E\boldsymbol{G}\boldsymbol{G}')^{-1}E\boldsymbol{G}\boldsymbol{U}' = E\left(\boldsymbol{G}^{(s)}\boldsymbol{G}^{(s)\prime}\right)$$

since

$$E\boldsymbol{G}\boldsymbol{U}' = -E\dot{\boldsymbol{G}} \qquad \forall\, \boldsymbol{G} \in \mathcal{H}$$

$$\left(\begin{array}{lcl} E\boldsymbol{G}\boldsymbol{U}' & = & \displaystyle\int \boldsymbol{G}\left(\frac{\partial \log L}{\partial \boldsymbol{\theta}}\right)' L \\[2em] & = & \displaystyle\int \boldsymbol{G}\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)' = -\int \frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\theta}}\, L \end{array}\right)$$

and similarly

$$E\left(\boldsymbol{U}^{(s)}\,\boldsymbol{G}^{(s)\prime}\right) = E\left(\boldsymbol{G}^{(s)}\,\boldsymbol{G}^{(s)\prime}\right).$$

These results lead immediately to the equality of the expressions (2.1) and (2.2) and hence the equivalence of Definition 2.1 and Definition 2.2.

A further useful interpretation of quasi-likelihood can be given in a Hilbert space setting. Let $\mathcal{H}$ be a closed subspace of $L^2 = L^2(\Omega, \mathcal{F}, P_0)$ of (equivalence classes) of random vectors with finite second moment. Then, for $\boldsymbol{X}, \boldsymbol{Y} \in L^2$, taking inner product $(\boldsymbol{X}, \boldsymbol{Y}) = E(\boldsymbol{X}'\boldsymbol{Y})$ and norm $\|\boldsymbol{X}\| = (\boldsymbol{X}, \boldsymbol{X})^{1/2}$ the space L^2 is a Hilbert space. We say that $\boldsymbol{X}$ is orthogonal to $\boldsymbol{Y}$, written $\boldsymbol{X} \perp \boldsymbol{Y}$, if $(\boldsymbol{X}, \boldsymbol{Y}) = 0$ and that subsets L_1^2 and L_2^2 of L^2 are orthogonal, which holds if $\boldsymbol{X} \perp \boldsymbol{Y}$ for every $\boldsymbol{X} \in L_1^2$, $\boldsymbol{Y} \in L_2^2$ (written $L_1^2 \perp L_2^2$).

For $\boldsymbol{X} \in L^2$, let $\pi(\boldsymbol{X} \,|\, \mathcal{H})$ denote the element of $\mathcal{H}$ such that

$$\|\boldsymbol{X} - \pi(\boldsymbol{X} \,|\, \mathcal{H})\|^2 = \inf_{Y \in \mathcal{H}} \|\boldsymbol{X} - \boldsymbol{Y}\|^2,$$

that is, $\pi(\boldsymbol{X} \,|\, \mathcal{H})$ is the orthogonal projection of $\boldsymbol{X}$ onto $\mathcal{H}$.

Now suppose that the score function $\boldsymbol{U}_T \in \mathcal{G}$. Then, dropping the subscript T and using Definition 2.2, the standardized quasi-score estimating function $\boldsymbol{H}^{(s)} \in \mathcal{H}$ is given by

$$\inf_{H^{(s)} \in \mathcal{H}} E\left(\boldsymbol{U} - \boldsymbol{H}^{(s)}\right)\left(\boldsymbol{U} - \boldsymbol{H}^{(s)}\right)',$$

and since

$$\operatorname{tr} E(\boldsymbol{U} - \boldsymbol{H}^{(s)})(\boldsymbol{U} - \boldsymbol{H}^{(s)})' = \|\boldsymbol{U} - \boldsymbol{H}^{(s)}\|^2,$$

tr denoting trace, *the quasi-score is* $\pi(\boldsymbol{U} \,|\, \mathcal{H})$, *the orthogonal projection of the score function onto the chosen space* $\mathcal{H}$ *of estimating functions.* For further discusion of the Hilbert space approach see Small and McLeish (1994) and Merkouris (1992).

Next, the vector correlation that measures the association between $\boldsymbol{G}_T = (G_{T,1}, \ldots, G_{T,p})'$ and $\boldsymbol{U}_T = (U_{T,1}, \ldots, U_{T,p})'$, defined, for example, by Hotelling (1936), is

$$\rho^2 = \frac{(\det(E\boldsymbol{G}_T\boldsymbol{U}_T'))^2}{\det(E\boldsymbol{G}_T\boldsymbol{G}_T')\det(E\boldsymbol{U}_T\boldsymbol{U}_T')},$$

where det denotes determinant. However, under the regularity conditions that have been imposed, $E\dot{\boldsymbol{G}}_T = -E(\boldsymbol{G}_T\boldsymbol{U}_T')$, so a maximum correlation requirement is to maximize

$$(\det(E\dot{\boldsymbol{G}}_T))^2/\det(E\boldsymbol{G}_T\boldsymbol{G}_T'),$$

which can be achieved by maximizing $\mathcal{E}(\boldsymbol{G}_T)$ in the partial order of nonnegative definite matrices. This corresponds to the criterion of Definition 2.1.

Neither Definition 2.1 nor Definition 2.2 is of direct practical value for applications. There is, however, an essentially equivalent form (Heyde (1988a)), that is very easy to use in practice.

Theorem 2.1 $\boldsymbol{G}_T^* \in \mathcal{H}$ is an O_F-optimal estimating function within $\mathcal{H}$ if

$$E\left(\boldsymbol{G}_T^{*(s)} \boldsymbol{G}_T^{(s)'}\right) = E\left(\boldsymbol{G}_T^{(s)} \boldsymbol{G}_T^{*(s)'}\right) = E\left(\boldsymbol{G}_T^{(s)} \boldsymbol{G}_T^{(s)'}\right) \tag{2.3}$$

or equivalently

$$\left(E\dot{\boldsymbol{G}}_T\right)^{-1} E\boldsymbol{G}_T \boldsymbol{G}_T^{*}{}'$$

is a constant matrix for all $\boldsymbol{G}_T \in \mathcal{H}$. Conversely, if $\mathcal{H}$ is convex and $\boldsymbol{G}_T^* \in \mathcal{H}$ is an O_F-optimal estimating function, then (2.3) holds.

Proof. Again we drop the subscript T for convenience. When (2.3) holds,

$$E\left(\boldsymbol{G}^{*(s)} - \boldsymbol{G}^{(s)}\right)\left(\boldsymbol{G}^{*(s)} - \boldsymbol{G}^{(s)}\right)' = E\left(\boldsymbol{G}^{*(s)} \boldsymbol{G}^{*(s)'}\right) - E\left(\boldsymbol{G}^{(s)} \boldsymbol{G}^{(s)'}\right)$$

is nonnegative definite, $\forall\, \boldsymbol{G} \in \mathcal{H}$, since the left-hand side is a covariance function. This gives optimality via Definition 2.1.

Now suppose that $\mathcal{H}$ is convex and $\boldsymbol{G}^*$ is an O_F-optimal estimating function. Then, if $\boldsymbol{H} = \alpha\, \boldsymbol{G} + \boldsymbol{G}^*$, we have that

$$E\left(\boldsymbol{G}^{*(s)} \boldsymbol{G}^{*(s)'}\right) - E\left(\boldsymbol{H}^{(s)} \boldsymbol{H}^{(s)'}\right)$$

is nonnegative definite, and after inverting and some algebra this gives that

$$\begin{aligned} \alpha^2 & \left[E\boldsymbol{G}\boldsymbol{G}' - \left(E\dot{\boldsymbol{G}}\right)\left(E\dot{\boldsymbol{G}}^*\right)^{-1} E\boldsymbol{G}^*\boldsymbol{G}^{*'}\left(\left(E\dot{\boldsymbol{G}}^*\right)^{-1}\right)'\left(E\dot{\boldsymbol{G}}\right)'\right] \\ & - \alpha\left[-E\boldsymbol{G}\boldsymbol{G}^{*'} + \left(E\dot{\boldsymbol{G}}\right)\left(E\dot{\boldsymbol{G}}^*\right)^{-1}\left(E\boldsymbol{G}^*\boldsymbol{G}^{*'}\right)\right] \\ & - \alpha\left[-E\boldsymbol{G}^*\boldsymbol{G}' + \left(E\boldsymbol{G}^*\boldsymbol{G}^{*'}\right)\left(\left(E\dot{\boldsymbol{G}}^*\right)^{-1}\right)'\left(E\dot{\boldsymbol{G}}\right)'\right] \end{aligned}$$

is nonnegative definite. This is of the form $\alpha^2\boldsymbol{A} - \alpha\boldsymbol{B}$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric and $\boldsymbol{A}$ is nonnegative definite by Definition 2.1.

Let $\boldsymbol{u}$ be an arbitrary nonzero vector of dimension p. We have $\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u} \geq 0$ and

$$\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u} \geq \alpha^{-1}\boldsymbol{u}'\boldsymbol{B}\boldsymbol{u}$$

for all α, which forces $\boldsymbol{u}'\boldsymbol{B}\boldsymbol{u} = 0$ and hence $\boldsymbol{B} = \boldsymbol{0}$.

Now $\boldsymbol{B} = \boldsymbol{0}$ can be rewritten as

$$\left(E\boldsymbol{G}\boldsymbol{G}'\right)\left(\left(E\dot{\boldsymbol{G}}\right)'\right)^{-1}\boldsymbol{C} + \boldsymbol{C}'\left(E\dot{\boldsymbol{G}}\right)^{-1}\left(E\boldsymbol{G}\boldsymbol{G}'\right) = \boldsymbol{0},$$

where

$$\boldsymbol{C} = \left(E\boldsymbol{G}^{(s)}\,\boldsymbol{G}^{(s)'} - E\boldsymbol{G}^{(s)}\,\boldsymbol{G}^{*(s)'}\right)\left(E\dot{\boldsymbol{G}}^*\right)^{-1} E\boldsymbol{G}^*\,\boldsymbol{G}^{*'}$$

and, as this holds for all $\boldsymbol{G} \in \mathcal{H}$, it is possible to replace $\boldsymbol{G}$ by $\boldsymbol{DG}$, where $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ is an arbitrary constant matrix. Then, in obvious notation

$$\lambda_i \left[(E\boldsymbol{G}\boldsymbol{G}') \left((E\dot{\boldsymbol{G}})' \right)^{-1} \boldsymbol{C} \right]_j + \left[\boldsymbol{C}'(E\dot{\boldsymbol{G}})^{-1}(E\boldsymbol{G}\boldsymbol{G}') \right]_i \lambda_j = 0$$

for each i, j, which forces $\boldsymbol{C} = \boldsymbol{0}$ and hence (2.3) holds. This completes the proof.

In general, Theorem 2.1 provides a straightforward way to check whether an O_F-optimal estimating function exists for a particular family $\mathcal{H}$. It should be noted that existence is by no means guaranteed.

Theorem 2.1 is especially easy to use when the elements $G \in \mathcal{H}$ have orthogonal differences and indeed this is often the case in applications. Suppose, for example, that

$$\mathcal{H} = \left\{ H : \; H = \sum_{t=1}^{T} \boldsymbol{a}_t(\boldsymbol{\theta}) \, \boldsymbol{h}_t(\boldsymbol{\theta}) \right\},$$

with $\boldsymbol{a}_t(\boldsymbol{\theta})$ constants to be chosen, $\boldsymbol{h}_t$'s fixed and random with zero means and $E\boldsymbol{h}_s(\boldsymbol{\theta})\boldsymbol{h}_t'(\boldsymbol{\theta}) = \boldsymbol{0}$, $s \neq t$. Then

$$E\boldsymbol{H}\boldsymbol{H}^{*'} = \sum_{t=1}^{T} \boldsymbol{a}_t E\boldsymbol{h}_t\boldsymbol{h}_t'\boldsymbol{a}_t^{*'}$$

$$E\dot{\boldsymbol{H}} = \sum_{t=1}^{T} \boldsymbol{a}_t E\dot{\boldsymbol{h}}_t$$

and $(E\dot{\boldsymbol{H}})^{-1}E\boldsymbol{H}\boldsymbol{H}^{*'}$ is constant for all $\boldsymbol{H} \in \mathcal{H}$ if

$$\boldsymbol{a}_t^* = \left(E\dot{\boldsymbol{h}}_t \right)' (E\boldsymbol{h}_t\boldsymbol{h}_t')^{-1}.$$

An O_F-optimal estimating function is thus

$$\sum_{t=1}^{T} \left(E\dot{\boldsymbol{h}}_t(\boldsymbol{\theta}) \right)' \left(E\boldsymbol{h}_t(\boldsymbol{\theta})\boldsymbol{h}_t'(\boldsymbol{\theta}) \right)^{-1} \boldsymbol{h}_t(\boldsymbol{\theta}).$$

As an illustration consider the estimation of the mean of the offspring distribution in a Galton-Watson process $\{Z_t\}$, $\theta = E(Z_1 | Z_0 = 1)$. Here the data are $\{Z_0, \ldots, Z_T\}$.

Let $\mathcal{F}_n = \sigma(Z_0, \ldots, Z_n)$. We seek a basic martingale (MG) from the $\{Z_i\}$. This is simple since

$$Z_i - E\left(Z_i \,|\, \mathcal{F}_{i-1}\right) = Z_i - \theta \, Z_{i-1}$$

are MG differences (and hence orthogonal). Let

$$\mathcal{H} = \left\{ h : h_T = \sum_{t=1}^{T} a_t(\theta)(Z_t - \theta\, Z_{t-1}), \quad a_t(\theta) \text{ is } \mathcal{F}_{t-1} \text{ measurable} \right\}.$$

We find that the O_F-optimal choice for $a_t(\theta)$ is

$$a_t^*(\theta) = -1/\sigma^2,$$

where $\sigma^2 = \text{var}(Z_1|Z_0 = 1)$. The O_F-optimal estimator of θ is

$$(Z_1 + \ldots + Z_T)/(Z_0 + \ldots + Z_{T-1}).$$

We would call this a quasi-likelihood estimator from the family $\mathcal{H}$. It is actually the MLE for the power series family of offspring distributions

$$P(Z_1 = j|Z_0 = 1) = A(j)\frac{(a(\theta))^j}{F(\theta)}, \quad j = 0, 1, 2, \ldots$$

where

$$F(\theta) = \sum_{j=0}^{\infty} A(j)(a(\theta))^j.$$

These form the discrete exponential family for this context.

To obtain the MLE result for the power series family note that

$$\begin{aligned} & P(Z_0 = z_0, \ldots, Z_T = z_T) \\ &= P(Z_0 = z_0) \prod_{k=1}^{T} P\left(Z_k = z_k \middle| Z_{k-1} = z_{k-1}\right) \\ &= P(Z_0 = z_0) \prod_{k=1}^{T} \left[\sum_{j_1 + \ldots + j_{z_{k-1}} = z_k} A(j_1) \ldots A(j_{z_{k-1}}) \right] \frac{(a(\theta))^{z_k}}{(F(\theta))^{z_{k-1}}} \\ &= P(Z_0 = z_0) \frac{(a(\theta))^{z_1 + \ldots + z_T}}{(F(\theta))^{z_0 + \ldots + z_{T-1}}} \times \text{term not involving } \theta \end{aligned}$$

and hence if $L = L(Z_0, \ldots, Z_T; \theta)$ is the likelihood,

$$\frac{d \log L}{d\,\theta} = (Z_1 + \ldots + Z_T)\, \frac{\dot{a}(\theta)}{a(\theta)} - (Z_0 + \ldots + Z_{T-1})\, \frac{\dot{F}(\theta)}{F(\theta)}.$$

However

$$\theta = \sum_{j=1}^{\infty} j\, A(j)\, \frac{(a(\theta))^j}{F(\theta)}, \qquad 1 = \sum_{j=0}^{\infty} A(j)\, \frac{(a(\theta))^j}{F(\theta)}$$

and differentiating with respect to θ in the latter result,

$$0 = \sum_{j=1}^{\infty} j\, A(j) \dot{a}(\theta)\, \frac{(a(\theta))^{j-1}}{F(\theta)} - \frac{\dot{F}(\theta)}{F^2(\theta)} \sum_{j=0}^{\infty} A(j)(a(\theta))^j$$

so that

$$\theta\, \frac{\dot{a}(\theta)}{a(\theta)} = \frac{\dot{F}(\theta)}{F(\theta)}$$

and the score function is

$$U_T(\theta) = \frac{\dot{a}(\theta)}{a(\theta)}\, [(Z_1 + \ldots + Z_T) - \theta(Z_0 + \ldots + Z_{T-1})]\,.$$

The same estimator can also be obtained quite generally as a nonparametric MLE (Feigin (1977)).

This example illustrates one important *general strategy for finding optimal estimating functions.* This strategy is to compute the score function for some plausible underlying distribution (such as a convenient member of the appropriate exponential family) $\partial \log L_0 / \partial \theta$, say, then use the differences in this martingale to form $\boldsymbol{h}_t$'s. Finally, choose the optimal estimating function within the class $\mathcal{H} = \{\sum_{t=1}^{T} \boldsymbol{a}_t \boldsymbol{h}_t\}$ by suitably specifying the weights $\boldsymbol{a}_t$.

The previous discussion and examples have concentrated on the case of discrete time. However, the theory operates in entirely similar fashion in the case of continuous time.

For example, consider the diffusion process

$$dX_t = \theta\, X_t\, dt + dW_t,$$

where W_t is standard Brownian motion. We wish to estimate θ on the basis of the data $\{X_t,\ 0 \le t \le T\}$.

Here the natural martingale to use is W_t, and we seek an O_F-optimal estimating function from the set $\mathcal{H} = \{\int_0^T b_s\, dW_s, \quad b_s \text{ predictable}\}$. Note that, for convenience, and to emphasize the methodology, we shall often write estimating functions in a form that emphasizes the noise component of the model and suppresses the dependence on the observations. Here $\int_0^T b_s\, dW_s$ is to be interpreted as $\int_0^T b_s (dX_s - \theta\, X_s\, ds)$.

We have, writing $H_T = \int_0^T b_s\, dW_s, \quad H_T^* = \int_0^T b_s^*\, dW_s,$

$$H_T = \int_0^T b_s\, dW_s = \int_0^T b_s\, dX_s - \theta \int_0^T b_s\, X_s\, ds,$$

so that

$$E\dot{H}_T = -E \int_0^T b_s\, X_s\, ds, \qquad E H_T H_T^* = E \int_0^T b_s\, b_s^*\, ds$$

and $(E\dot{H}_T)^{-1}\, EH_T\, H_T^*$ is constant for all $H \in \mathcal{H}$ if $b_s^* = X_s$. Then,

$$H_T^* = \int_0^T X_s\, dX_s - \theta \int_0^T X_s^2\, ds$$

and the QLE is

$$\hat{\theta}_T = \int_0^T X_s\, dX_s \bigg/ \int_0^T X_s^2\, ds.$$

This is also the MLE.

As another example, we consider the multivariate counting process $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,p})'$, each $X_{t,i}$ being of the form

$$X_{t,i} = \theta_i \int_0^t J_i(s)\, ds + M_{t,i}$$

with multiplicative intensity $\Lambda_i(t) = \theta_i\, J_i(t)$, $J_i(t) > 0$ a.s. being predictable and $M_{t,i}$ a square integrable martingale. This is a special case of the framework considered by Aalen (1978), see also Andersen et.al. (1982) and it covers a variety of contexts for processes such as those of birth and death type. The case $p = 1$ has been discussed by Thavaneswaran and Thompson (1986).

The data are $\{\boldsymbol{X}_t,\ 0 \le t \le T\}$ and we write

$$\boldsymbol{X}_T = \left(\int_0^T \boldsymbol{J}(s)\, ds\right) \boldsymbol{\theta} + \boldsymbol{M}_T,$$

where

$$\boldsymbol{J}(s) = \operatorname{diag}(J_1(s), \ldots, J_p(s)), \quad \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)', \quad \boldsymbol{M}_T = (M_{T,1}, \ldots, M_{T,p})'$$

and we note that for counting processes $M_{T,i}$ and $M_{T,j}$ are orthogonal for $i \neq j$. Then, we seek an O_F-optimal estimating function from the set

$$\mathcal{H} = \left\{ \int_0^T \boldsymbol{b}_s\, d\boldsymbol{M}_s, \quad \boldsymbol{b}_s \text{ predictable} \right\}.$$

Let $\boldsymbol{H}_T = \int_0^T \boldsymbol{b}_s\, d\boldsymbol{M}_s$, $\boldsymbol{H}_T^* = \int_0^T \boldsymbol{b}_s^*\, d\boldsymbol{M}_s$. We have

$$E\dot{\boldsymbol{H}}_T = -E \int_0^T \boldsymbol{b}_s\, \boldsymbol{J}(s)\, ds,$$

$$E\boldsymbol{H}_T \boldsymbol{H}_T^{*'} = E \int_0^T \boldsymbol{b}_s\, \boldsymbol{J}(s)\, \boldsymbol{\theta}\, \boldsymbol{b}_s^{*'}\, ds,$$

and $(E\dot{\boldsymbol{H}}_T)^{-1} E\boldsymbol{H}_T \boldsymbol{H}_T^{*'}$ is constant for all $\boldsymbol{H} \in \mathcal{H}$ if $\boldsymbol{b}_s^* \equiv \boldsymbol{I}$. Thus $\boldsymbol{H}_T^* = \boldsymbol{M}_T$ is an O_F-optimal estimating function and the corresponding QLE is

$$\hat{\boldsymbol{\theta}}_T = \left(\int_0^T \boldsymbol{J}_s\, ds\right)^{-1} \boldsymbol{X}_T.$$

That this $\hat{\boldsymbol{\theta}}$ is also the MLE under rather general conditions follows from § 3.3 of Aalen (1978). The simplest particular case is where each $X_{t,i}$ is a Poisson process with parameter θ_i.

General comment: The art, as distinct from the science, in using quasi-likelihood methods is in a good choice of the family $\mathcal{H}$ of estimating functions with which to work. The ability to choose $\mathcal{H}$ is a considerable strength as the family can be tailor made to the requirements of the context, and regularity conditions can be built in. However, it is also a source of weakness, since it is by no means always clear what competing families of estimating functions might exist with better properties and efficiencies. Of course, the quasi-likelihood framework does provide a convenient basis for comparison of families via the information criterion.

For examples of estimators for autoregressive processes with positive or bounded innovations that are much better than the naive QLE chosen from the natural family of estimating functions for standard autoregressions see Davis and McCormick (1989).

2.3 Scalar Equivalences and Associated Results

Comparison of information matrices in the partial order of nonnegative definite matrices may be difficult in practice, especially if the information matrices are based on quite different families of estimating functions. In the case where an O_F-optimal estimating function exists, however, we may replace the matrix comparison by simpler scalar ones. The following result is essentially due to Chandrasekar and Kale (1984).

Theorem 2.2 Suppose that $\mathcal{H}$ is a space of estimating functions for which an O_F-optimal estimating function exists. The condition that $\boldsymbol{G}_T^*$ is O_F-optimal in $\mathcal{H}$, i.e., that $\mathcal{E}(\boldsymbol{G}_T^*) - \mathcal{E}(\boldsymbol{G}_T)$ is nonnegative definite for all $\boldsymbol{G}_T \in \mathcal{H}$, is equivalent to either of the two alternative conditions: for all $\boldsymbol{G}_T \in \mathcal{H}$,
(i) (trace (T) criterion) $\quad \operatorname{tr} \mathcal{E}(\boldsymbol{G}_T^*) \geq \operatorname{tr} \mathcal{E}(\boldsymbol{G}_T)$;
(ii) (determinant (D) criterion) $\quad \det(\mathcal{E}(\boldsymbol{G}_T^*)) \geq \det(\mathcal{E}(\boldsymbol{G}_T))$.

Remark There are further alternative equivalent conditions in addition to (i) and (ii), in particular
(iii) (smallest eigenvalue (E) criterion)

$$\lambda_{min}\left(\mathcal{E}\left(\boldsymbol{G}_T^*\right)\right) \geq \lambda_{min}\left(\mathcal{E}\left(\boldsymbol{G}_T\right)\right)$$

and
(iv) (average variance (A) criterion)

$$\left(p^{-1} \operatorname{tr}\left(\mathcal{E}\left(\boldsymbol{G}_T^*\right)\right)^{-1}\right)^{-1} \geq \left(p^{-1} \operatorname{tr}\left(\mathcal{E}\left(\boldsymbol{G}_T\right)\right)^{-1}\right)^{-1}.$$

Conditions (i) – (iv) have been widely used in the theory of optimal experimental design where they respectively correspond to T, D, E and A-optimality. See Pukelsheim (1993), e.g., Chapter 9, and references therein. For experimental designs it often happens that a Loewner optimal design (i.e., O_F-optimal estimating function) does not exist, but an A, D, E or T optimal design (i.e., estimating function optimal in the sense of the A, D, E or T criterion) can be found. See Pukelsheim (1993, p. 104) for a discussion of nonexistence of Loewner optimal designs.

Proof of Theorem 2.2 We shall herein drop the subscript T for convenience.
(i) The condition that $\mathcal{E}(\boldsymbol{G}^*) - \mathcal{E}(\boldsymbol{G})$ is nnd immediately gives

$$\operatorname{tr}(\mathcal{E}(\boldsymbol{G}^*) - \mathcal{E}(\boldsymbol{G})) = \operatorname{tr}\mathcal{E}(\boldsymbol{G}^*) - \operatorname{tr}\mathcal{E}(\boldsymbol{G}) \geq 0.$$

Conversely, suppose $\boldsymbol{H}$ satisfies $\operatorname{tr}\mathcal{E}(\boldsymbol{H}) \geq \operatorname{tr}\mathcal{E}(\boldsymbol{G})$ for all $\boldsymbol{G} \in \mathcal{H}$. If there is an O_F-optimal $\boldsymbol{G}^*$, then $\operatorname{tr}\mathcal{E}(\boldsymbol{H}) \geq \operatorname{tr}\mathcal{E}(\boldsymbol{G}^*)$. But from the definition of O_F-optimality we also have $\operatorname{tr}\mathcal{E}(\boldsymbol{G}^*) \geq \operatorname{tr}\mathcal{E}(\boldsymbol{H})$ and hence $\operatorname{tr}\mathcal{E}(\boldsymbol{G}^*) = \operatorname{tr}\mathcal{E}(\boldsymbol{H})$. Thus, we have that $\boldsymbol{A} = \mathcal{E}(\boldsymbol{G}^*) - \mathcal{E}(\boldsymbol{H})$ is nnd and $\operatorname{tr}\boldsymbol{A} = 0$. But $\boldsymbol{A}$ being symmetric and nnd implies that all its eigenvalues are positive, while $\operatorname{tr}\boldsymbol{A} = \boldsymbol{0}$ implies that the sum of all the eigenvalues of $\boldsymbol{A}$ is zero. This forces all the eigenvalues of $\boldsymbol{A}$ to be zero, which can only happen if $\boldsymbol{A} \equiv \boldsymbol{0}$, since the sum of squares of the elements of $\boldsymbol{A}$ is the sum of squares of its eigenvalues. Thus, $\mathcal{E}(\boldsymbol{G}^*) = \mathcal{E}(\boldsymbol{H})$ and we have an O_F-optimal estimating function.
(ii) Here we apply the Simultaneous Reduction Lemma (e.g., Rao (1973, p. 41)) which states that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices and $\boldsymbol{B}$ is positive definite (pd), then there is a nonsingular matrix $\boldsymbol{R}$ such that $\boldsymbol{A} = (\boldsymbol{R}^{-1})'\boldsymbol{\Lambda}\,\boldsymbol{R}^{-1}$ and $\boldsymbol{B} = (\boldsymbol{R}^{-1})'\,\boldsymbol{R}^{-1}$, where $\boldsymbol{\Lambda}$ is diagonal.

In the nontrivial case we first suppose that $\mathcal{E}(\boldsymbol{G}^*)$ is pd. Then using the Simultaneous Reduction Lemma we may suppose that there exists a nonsingular matrix $\boldsymbol{R}$ such that for fixed $\boldsymbol{G}$

$$\mathcal{E}(\boldsymbol{G}^*) = (\boldsymbol{R}^{-1})'\boldsymbol{R}^{-1}, \qquad \mathcal{E}(\boldsymbol{G}) = (\boldsymbol{R}^{-1})'\boldsymbol{\Lambda}_G\boldsymbol{R}^{-1},$$

where $\boldsymbol{\Lambda}_{\boldsymbol{G}}$ is diagonal. Then the condition that

$$\mathcal{E}(\boldsymbol{G}^*) - \mathcal{E}(\boldsymbol{G}) = (\boldsymbol{R}^{-1})'(\boldsymbol{I} - \boldsymbol{\Lambda}_G)\boldsymbol{R}^{-1}$$

is nnd forces

$$\det(\mathcal{E}(\boldsymbol{G}^*) - \mathcal{E}(\boldsymbol{G})) = (\det(\boldsymbol{R}^{-1}))^2 \det(\boldsymbol{I} - \boldsymbol{\Lambda}_G) \geq 0.$$

This means that $\det\boldsymbol{\Lambda}_G \leq 1$ and hence

$$\det(\mathcal{E}(\boldsymbol{G}^*)) = \det(\boldsymbol{R}^{-1})^2 \geq \det(\mathcal{E}(\boldsymbol{G})) = \det(\boldsymbol{R}^{-1})^2 \det(\boldsymbol{\Lambda}_G).$$

Conversely, suppose that $\boldsymbol{H}$ satisfies $\det(\mathcal{E}(\boldsymbol{H})) \geq \det(\mathcal{E}(\boldsymbol{G}))$ for all $\boldsymbol{G} \in \mathcal{H}$. As with the proof of (i) we readily find that $\det(\mathcal{E}(\boldsymbol{H})) = \det(\mathcal{E}(\boldsymbol{G}^*))$ when $\boldsymbol{G}^*$

is O_F-optimal. An application of the Simultaneous Reduction Lemma to the pair $\mathcal{E}(\boldsymbol{G}^*)$, $\mathcal{E}(\boldsymbol{H})$, the former taken as pd, leads immediately to $\mathcal{E}(\boldsymbol{G}^*) = \mathcal{E}(\boldsymbol{H})$ and an O_F-optimal solution.

Remark It must be emphasized that the existence of an O_F-optimal estimating function within $\mathcal{H}$ is a crucial assumption in the theorem. For example, if $\boldsymbol{G}^*$ satisfies the trace criterion (i), it is not ensured that $\boldsymbol{G}^*$ is an O_F-optimal estimating function within $\mathcal{H}$; there may not be one.

2.4 Wedderburn's Quasi-Likelihood

2.4.1 The Framework

Historically there have been two distinct approaches to parameter inference developed from both classical least squares and maximum likelihood methods. One is the optimal estimation approach introduced by Godambe (1960), and others, from the viewpoint of estimating functions. The other, introduced by Wedderburn (1974) as a basis for analyzing generalized linear regressions, was termed quasi-likelihood from the outset. Both approaches have seen considerable development in their own right. For those based on the Wedderburn approach see, for example, Liang and Zeger (1986), Morton (1987), and Mc Cullagh and Nelder (1989).

In this book our emphasis is on the optimal estimating functions approach and, as we shall show in this section, the Wedderburn approach can be regarded as a particular case of the optimal estimating function approach where we restrict the space of estimating functions to a special class.

Wedderburn observed that, from a computational point of view, the only assumptions on a generalized linear model necessary to fit the model were a specification of the mean (in terms of the regression parameters) and the relationship between the mean and the variance, not necessarily a fully specified likelihood. Therefore, he replaced the assumptions on the probability distribution by defining a function based solely on the mean-variance relationship, which had algebraic and frequency properties similarly to those of log-likelihoods. For example, for a regression model

$$\boldsymbol{Y} = \boldsymbol{\mu}(\boldsymbol{\theta}) + \boldsymbol{e}$$

with $E\boldsymbol{e} = \boldsymbol{0}$, we suppose that a function q can be defined by the differential equation

$$\frac{\partial q}{\partial \boldsymbol{\theta}} = \boldsymbol{Q}(\boldsymbol{\theta}) = \dot{\boldsymbol{\mu}}' \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})),$$

for matrices $\dot{\boldsymbol{\mu}} = (\partial \mu_i / \partial \theta_j)$ and $\boldsymbol{V} = E\boldsymbol{e}\boldsymbol{e}'$. Then

$$E\{\boldsymbol{Q}(\boldsymbol{\theta})\} = 0.$$

$$E\left\{\frac{\partial}{\partial \boldsymbol{\theta}} (\boldsymbol{Q}(\boldsymbol{\theta}))\right\} = -\dot{\boldsymbol{\mu}}' \boldsymbol{V}^{-1} \dot{\boldsymbol{\mu}}.$$

$$\operatorname{cov}\{\boldsymbol{Q}(\boldsymbol{\theta})\} = \dot{\boldsymbol{\mu}}'\boldsymbol{V}^{-1}\dot{\boldsymbol{\mu}}.$$

Thus $\boldsymbol{Q}(\boldsymbol{\theta})$ behaves like the derivative of a log-likelihood (a score function) and is termed a quasi-score or quasi-score estimating function from the viewpoint of estimating functions, while q itself is called a quasi-(log)likelihood. A common approach to get the quasi-score estimating function has been to first write down the general weighted sum of squares of residuals,

$$(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta}))'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})),$$

and then differentiate it with respect to $\boldsymbol{\theta}$ assuming $\boldsymbol{V}$ is independent of $\boldsymbol{\theta}$. We now put this approach into an estimating function setting.

Consider a model

$$\boldsymbol{Y} = \boldsymbol{\mu}(\boldsymbol{\theta}) + \boldsymbol{e}, \tag{2.4}$$

where $\boldsymbol{Y}$ is an $n \times 1$ data vector, $E\boldsymbol{e} = 0$ and $\boldsymbol{\mu}(\boldsymbol{\theta})$, which now may be random but for which $E(\boldsymbol{e}\,\boldsymbol{e}' \,|\, \dot{\boldsymbol{\mu}}) = \boldsymbol{V}$, involves an unknown parameter $\boldsymbol{\theta}$ of dimension p.

We consider the estimating function space

$$\mathcal{H} = \{\boldsymbol{A}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta}))\},$$

for $p \times p$ matrices $\boldsymbol{A}$ not depending on $\boldsymbol{\theta}$ which are $\dot{\boldsymbol{\mu}}$-measurable and satisfy the conditions that $E\boldsymbol{A}\,\dot{\boldsymbol{\mu}}$ and $E(\boldsymbol{A}\,\boldsymbol{e}\,\boldsymbol{e}'\,\boldsymbol{A}')$ are nonsingular. Then we have the following theorem.

Theorem 2.3 The estimating function

$$\boldsymbol{G}^* = \dot{\boldsymbol{\mu}}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \tag{2.5}$$

is a quasi-score estimating function within $\mathcal{H}$.

This result follows immediately from Theorem 2.1. If

$$\boldsymbol{G} = \boldsymbol{A}(\boldsymbol{Y} - \boldsymbol{\mu}),$$

we have

$$\begin{aligned} E\boldsymbol{G}\,\boldsymbol{G}^{*'} &= E(\boldsymbol{A}\,\boldsymbol{e}\,\boldsymbol{e}'\,\boldsymbol{V}^{-1}\dot{\boldsymbol{\mu}}) \\ &= E(\boldsymbol{A}E(\boldsymbol{e}\,\boldsymbol{e}' \,|\, \dot{\boldsymbol{\mu}})\boldsymbol{V}^{-1}\,\dot{\boldsymbol{\mu}}) \\ &= E(\boldsymbol{A}\,\dot{\boldsymbol{\mu}}) = -E\dot{\boldsymbol{G}} \end{aligned}$$

as required.

The estimating function (2.5) has been widely used in practice, in particular through the family of generalized linear models (e.g, McCullagh and Nelder (1989, Chapter 10)). A particular issue has been to deal with dispersion ϕ, which is modeled by using $\phi\,V(\boldsymbol{\theta})$ in place of the V in (2.5) (e.g., Nelder and

Lee (1992)). Another method of introducing extra variation into the model has been investigated by Morton (1989).

The Wedderburn estimating function (2.5) is also appropriate well beyond the usual setting of the generalized linear model. For example, estimation for all standard ARMA time series models are covered in Theorem 2.3. In practice, we obtain the familiar Yule-Walker equations for the quasi-likelihood estimation of the parameters of an autoregressive process under assumptions on only the first and second moments of the underlying distribution.

2.4.2 Limitations

It is not generally the case, however, that the Wedderburn estimating function (2.5) remains as a quasi-score estimating function if the class $\mathcal{H}$ of estimating functions is enlarged. Superior estimating functions may be found as we shall illustrate below.

It should be noted that, when $\boldsymbol{\mu}(\boldsymbol{\theta})$ is nonrandom, $\mathcal{H}$ confines attention to nonrandom weighting matrices $\boldsymbol{A}$. Allowing for random weights may improve precision in estimation.

As a simple example, suppose that $\boldsymbol{Y} = (x_1, \dots, x_n)'$, $\boldsymbol{e} = (e_1, \dots, e_n)'$ and the model (2.4) is of the form

$$x_i = \theta + e_i, \qquad i = 1, 2, \dots, n$$

with

$$E\left(e_i \,\middle|\, \mathcal{F}_{i-1}\right) = 0, \qquad E\left(e_i^2 \,\middle|\, \mathcal{F}_{i-1}\right) = \sigma^2\, x_{i-1}^2,$$

$\mathcal{F}_i$ being the σ-field generated by $x_1, \dots, x_i$. Then, it is easily checked using Theorem 2.1 that the quasi-score estimating function from the estimating function space

$$\mathcal{H} = \left\{ \sum_{i=1}^n a_i (x_i - \theta), \qquad a_i \text{ is } \mathcal{F}_{i-1} \text{ measurable} \right\}$$

is

$$G_1^* = \sigma^{-2} \sum_{i=1}^n x_{i-1}^{-2} (x_i - \theta)$$

in contrast to the Wedderburn estimating function (2.5), i.e.,

$$G^* = \sigma^{-2} \sum_{i=1}^n \left(E x_{i-1}^2\right)^{-1} (x_i - \theta),$$

which is the quasi-score estimating function from $\mathcal{H}$ in which all α_i are assumed constant. If θ_1^* and θ^* are, respectively, the solutions of $G_1^*(\theta_1^*) = 0$ and $G^*(\theta^*) = 0$, then

$$(E\dot{G}_1^*)^2 (E(G_1^*)^2)^{-1} \;=\; \sigma^{-2} \sum_{i=1}^n E(x_{i-1}^{-2})$$

$$
\begin{aligned}
&\geq \sigma^{-2} \sum_{i=1}^{n} (E x_{i-1}^2)^{-1} \\
&= (E\dot{G}^*)^2 (E(G^*)^2)^{-1}
\end{aligned}
$$

since $(EZ)(EZ^{-1}) \geq 1$ via the Cauchy-Schwarz inequality. Thus G_1^* is superior to G^*.

Now it can also happen that linear forms of the kind that are used in $\mathcal{H}$ are substantially inferior to nonlinear functions of the data. The motivation for $\mathcal{H}$ comes from exponential family considerations, and distributions that are far from this type may of course arise. These will fit within different families of estimating functions.

Suppose, for example, that $\boldsymbol{Y} = (x_1, \ldots, x_n)'$, $\boldsymbol{e} = (e_1, \ldots, e_n)'$ where

$$x_i = \theta + e_i, \qquad i = 1, \ldots, n$$

with the e_i being i.i.d. with density function

$$f(x) = 2\alpha^{1/4}\, e^{-\alpha x^4} / \Gamma(1/4), \qquad -\infty < x < \infty.$$

Then, we find that the true score function is

$$U(\theta) = 4\,\alpha \sum_{i=1}^{n} (x_i - \theta)^3.$$

This is much superior to the Wedderburn estimating function (2.5) for estimation of θ, i.e.,

$$G^* = \sigma^{-2} \sum_{i=1}^{n} (x_i - \theta)$$

where $\sigma^2 = \mathrm{var}(x_1) = E e_1^2$. Indeed, after some calculation we obtain

$$(E\dot{U})^{-2}\, E(U)^2 = \frac{E e_1^6}{(9(E e_1^2)^2) n} = \frac{\alpha^{-1/2}\, \Gamma(1/4)}{12\, n\, \Gamma(3/4)},$$

which is approximately 0.729477 of

$$(E\dot{G}^*)^{-2}\, E(G^*)^2 = \frac{1}{n}\, E\, e_1^2 = \frac{\alpha^{-1/2}\, \Gamma(3/4)}{n\, \Gamma(1/4)}.$$

This means that the length of an asymptotic confidence interval for θ derived from G^* will be $\sqrt{1/0.729477} \approx 1.1708$ times the corresponding one derived from U. The true score estimating function here could, for example, be regarded as an estimating function from the family

$$\mathcal{H}_2 = \left\{ \sum_{i=1}^{n} (a_i (x_i - \theta) + b_i (x_i - \theta)^3), \quad a_i\text{'s}, b_i\text{'s constants} \right\},$$

third moments being assumed zero, in contrast to G^* derived from

$$\mathcal{H} = \left\{ \sum_{i=1}^{n} a_i(x_i - \theta), \quad a_i\text{'s constant} \right\}.$$

From the above discussion, we see the flexibility of being able to choose an appropriate space of estimating functions in the optimal estimating functions approach.

2.4.3 Generalized Estimating Equations

Closely associated with Wedderburn's quasi-likelihood is the moment based generalized estimating equation (GEE) method developed by Liang and Zeger (1986), and Prentice (1988). The GEE approach was formulated to deal with problems of longitudiual data analysis where one typically has a series of repeated measurements of a response variable, together with a set of covariates, on each unit or individual observed chronologically over time. The response variables will usually be positively correlated. Furthermore, in the commonly occurring situation of data with discrete responses there is no comprehensive likelihood based approach analogous to that which comes from multivariate Gaussian assumptions. Consequently, there has been considerable interest in an approach that does not require full specification of the joint distribution of the repeated responses . For a recent survey of the area see Fitzmaurice, Laird and Rotnitzky (1993).

We shall follow Desmond (1996) in describing the formulation. This deals with a longitudinal data set consisting of responses $Y_{it},\ t = 1, 2, \ldots, n_i,\ i = 1, 2, \ldots, k$, say, where i indexes the individuals and t the repeated observations per individual. Observations on different individuals would be expected to be independent, while those on the same individual are correlated over time. Then, the vector of observations $\boldsymbol{Y} = (Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{k1}, \ldots, Y_{kn_k})'$ will have a covariance matrix $\boldsymbol{V}$ with block-diagonal structure

$$\boldsymbol{V} = \text{diag}\ (\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_k).$$

Suppose also that $\boldsymbol{V}_i = \boldsymbol{V}_i(\boldsymbol{\mu}_i, \lambda_i),\ i = 1, 2, \ldots, k$ where $\boldsymbol{\mu}_i = (\mu_{i_1}, \ldots, \mu_{in_i})$ is the vector of means for the ith individual and λ_i is a parameter including variance and correlation components. Finally, the means μ_{it} depend on covariates and a $p \times 1$ regression parameter $\boldsymbol{\theta}$, i.e.,

$$\mu_{it} = \mu_{it}(\boldsymbol{\theta}), \quad t = 1, 2, \ldots, n_i,\ i = 1, 2, \ldots, k.$$

For example, in the case of binary response variables and $n_i = T$ for each i, we may suppose that

$$P\left(Y_{it} = 1 \,\middle|\, \boldsymbol{x}_{it}, \boldsymbol{\theta}\right) = \mu_{it}, \quad \log\left[\mu_{it}/(1 - \mu_{it})\right] = \boldsymbol{x}'_{it}\,\boldsymbol{\theta},$$

where $\boldsymbol{x}_{it}$ represents a covariate vector associated with individual i and time t. This is the logit link function.

The basic model is then

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon},$$

say, where $E\boldsymbol{\epsilon} = \boldsymbol{0}$, $E\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}' = \boldsymbol{V}$ and, assuming the λ_i, $i = 1, 2, \dots, k$ known, the quasi-score estimating function from the family

$$\mathcal{H} = \{\boldsymbol{A}\,(\boldsymbol{Y} - \boldsymbol{\mu})\}$$

for $\sum_{i=1}^k n_i \times \sum_{i=1}^k n_i$ matrices $\boldsymbol{A}$ satisfying appropriate conditions is

$$\boldsymbol{Q}(\boldsymbol{\theta}) = \dot{\boldsymbol{\mu}}'\,\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$$

as in Theorem 2.3. Equivalently, upon making use of the block-diagonal structure of $\boldsymbol{V}$, this may be written as

$$\boldsymbol{Q}(\boldsymbol{\theta}) = \sum_{i=1}^{k} \dot{\boldsymbol{\mu}}_i'\,\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i),$$

where $\dot{\boldsymbol{\mu}}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\theta}$, $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{in_i})'$, $i = 1, 2, \dots, k$. The GEE is based on the estimating equation $\boldsymbol{Q}(\boldsymbol{\theta}) = \boldsymbol{0}$.

Now the particular feature of the GEE methodology is the use of "working" or approximate covariance matrices in place of the generally unknown $\boldsymbol{V}_i$. The idea is that the estimator thereby obtained will ordinarily be consistent regardless of the true correlation between the responses. Of course any replacement of the true $\boldsymbol{V}_i$ by other covariance matrices renders the GEE suboptimal. It is no longer based on a quasi-score estimating function although it may be asymptotically equivalent to one. See Chapter 5 for a discussion of asymptotic quasi-likelihood and, in particular, Chapter 5.5, Exercise 3.

Various common specifications of possible time dependence and the associated methods of estimation that have been proposed are detailed in Fitzmaurice, Laird and Rotnitzky (1993).

2.5 Asymptotic Criteria

Here we deal with the widely applicable situation in which families of estimating functions that are martingales are being considered.

Since the score function, if it exists, is usually a martingale, it is quite natural to approximate it using families of martingale estimating functions. Furthermore, the availability of comprehensive strong law and central limit results for martingales allows for straightforward discussion of issues of consistency, asymptotic normality, and efficiency in this setting.

For $n \times 1$ vector valued martingales $\boldsymbol{M}_T$ and $\boldsymbol{N}_T$, the $n \times n$ process $\langle \boldsymbol{M}, \boldsymbol{N}' \rangle_T$ is the *mutual quadratic characteristic*, a predictable increasing process such that $\boldsymbol{M}_T\boldsymbol{N}_T' - \langle \boldsymbol{M}, \boldsymbol{N}' \rangle_T$ is an $n \times n$ martingale. We shall write $\langle \boldsymbol{M} \rangle_T$ for $\langle \boldsymbol{M}, \boldsymbol{M}' \rangle_T$, the quadratic characteristic of $\boldsymbol{M}_T$. A convenient sketch of these concepts is given by Shiryaev (1981); see also Rogers and Williams (1987, IV. 26 and VI. 34).

Let $\mathcal{M}_1$ denote the subset of $\mathcal{G}$ that are square integrable martingales. For $\{\boldsymbol{G}_T\} \in \mathcal{M}_1$ there is, under quite broad conditions, a multivariate central limit result

$$\langle \boldsymbol{G} \rangle_T^{-\frac{1}{2}} \boldsymbol{G}_T \to MVN(\boldsymbol{0}, \boldsymbol{I}_p) \tag{2.6}$$

in distribution, as $T \to \infty$; see Chapter 12 for details.

For the general theory there are significant advantages for the random normalization using $\langle \boldsymbol{G} \rangle_T$ rather than constant normalization using the covariance $E\boldsymbol{G}_T\, \boldsymbol{G}_T'$. There are many cases in which normings by constants are unable to produce asymptotic normality, but instead lead to asymptotic mixed normality. These are cases in which operational time involves an intrinsic rescaling of ordinary time. Estimation of the mean of the offspring distribution in a Galton-Watson branching process described in Section 2.1 illustrates the point. The (martingale) quasi-score estimating function is $Q_T = \sum_{i=1}^T (Z_i - \theta Z_{i-1})$, the quadratic characteristic is $\langle Q \rangle_T = \sigma^2 \sum_{i=1}^T Z_{i-1}$ and, on the set $\{W = \lim_{n\to\infty} \theta^{-n} Z_n > 0\}$,

$$\langle Q \rangle_T^{-\frac{1}{2}} Q_T \xrightarrow{\mathrm{d}} N(0,1), \quad (EQ_T^2)^{-\frac{1}{2}} Q_T \xrightarrow{\mathrm{d}} W^{\frac{1}{2}} N(0,1),$$

the product in this last limit being a mixture of independent $W^{\frac{1}{2}}$ and $N(0,1)$ random variables. Later, in Chapter 4.5.1, it is shown that the normal limit form of central limit theorem has advantages, for provision of asymptotic confidence intervals, over the mixed normal form.

Let $\mathcal{M}_2 \subseteq \mathcal{M}_1$ be the subclass for which (2.6) obtains. Next, with $\boldsymbol{G}_T \in \mathcal{M}_2$, let $\boldsymbol{\theta}^*$ be a solution of $\boldsymbol{G}_T(\boldsymbol{\theta}) = \boldsymbol{0}$ and use Taylor's expansion to obtain

$$\boldsymbol{0} = \boldsymbol{G}_T(\boldsymbol{\theta}^*) = \boldsymbol{G}_T(\boldsymbol{\theta}) + \dot{\boldsymbol{G}}_T(\boldsymbol{\theta}^\dagger)(\boldsymbol{\theta}^* - \boldsymbol{\theta}), \tag{2.7}$$

where $\|\boldsymbol{\theta} - \boldsymbol{\theta}^\dagger\| \le \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|$, the norm denoting sum of squares of elements. Then, if $\dot{\boldsymbol{G}}_T(\boldsymbol{\theta})$ is nonsingular for $\boldsymbol{\theta}$ in a suitable neighborhood and

$$(\dot{\boldsymbol{G}}_T(\boldsymbol{\theta}^\dagger))^{-1} \dot{\boldsymbol{G}}_T(\boldsymbol{\theta}) \to \boldsymbol{I}_p$$

in probability as $T \to \infty$, expressions (2.6) and (2.7) lead to

$$\langle \boldsymbol{G}(\boldsymbol{\theta}) \rangle_T^{-\frac{1}{2}} \dot{\boldsymbol{G}}_T(\boldsymbol{\theta})(\boldsymbol{\theta}^* - \boldsymbol{\theta}) \to MVN(\boldsymbol{0}, \boldsymbol{I}_p)$$

in distribution.

Now, for $\{\mathcal{F}_T\}$ the filtration corresponding to $\{\boldsymbol{G}_T\}$ define the predictable process

$$\bar{\boldsymbol{G}}_t(\boldsymbol{\theta}) = \int_0^t E\left(d\dot{\boldsymbol{G}}_s(\boldsymbol{\theta}) \,\middle|\, \mathcal{F}_{s-} \right),$$

$\mathcal{F}_{s-}$ being the σ-field generated by $\bigcup_{r<s} \mathcal{F}_r$, and assume that $\dot{\boldsymbol{G}}_T(\boldsymbol{\theta})$ admits a Doob-Meyer type decomposition

$$\dot{\boldsymbol{G}}_T(\boldsymbol{\theta}) = \boldsymbol{M}_{G,T}(\boldsymbol{\theta}) + \bar{\boldsymbol{G}}_T(\boldsymbol{\theta}),$$

$\boldsymbol{M}_{G,T}(\boldsymbol{\theta})$ being a martingale. Then, under modest conditions, for example, if $\|\dot{\boldsymbol{G}}_T(\boldsymbol{\theta})\| \to \infty$ almost surely as $T \to \infty$,

$$\|\boldsymbol{M}_{G,T}(\boldsymbol{\theta})\| = o_p(\|\bar{\boldsymbol{G}}_T(\boldsymbol{\theta})\|)$$

as $T \to \infty$, o_p denoting small order in probability.

Thus, considerations which can be formalized under appropriate regularity conditions indicate that, for $\boldsymbol{G}_T$ belonging to some $\mathcal{M}_3 \subseteq \mathcal{M}_2$,

$$\langle \boldsymbol{G}(\boldsymbol{\theta})\rangle_T^{-\frac{1}{2}}\, \bar{\boldsymbol{G}}(\boldsymbol{\theta})(\boldsymbol{\theta}^* - \boldsymbol{\theta}) \to MVN(\boldsymbol{0}, \boldsymbol{I}_p)$$

in distribution, and hence

$$(\boldsymbol{\theta}^* - \boldsymbol{\theta})'\bar{\boldsymbol{G}}_T'(\boldsymbol{\theta})\langle \boldsymbol{G}(\boldsymbol{\theta})\rangle_T^{-1}\, \bar{\boldsymbol{G}}_T(\boldsymbol{\theta})(\boldsymbol{\theta}^* - \boldsymbol{\theta}) \to \chi_p^2 \tag{2.8}$$

in distribution. Best (asymptotic) estimation within a class $\mathcal{M} \subseteq \mathcal{M}_3$ of estimating functions is then achieved by choosing $\boldsymbol{G}_T^* \in \mathcal{M}$ so that Definition 2.3 below is satisfied. In this case we shall say that $\boldsymbol{G}_T^*$ is O_A-*optimal within* $\mathcal{M}$, O_A meaning optimal in the asymptotic sense.

Definition 2.3 $\boldsymbol{G}_T^* \in \mathcal{M}$ is an O_A-optimal estimating function within $\mathcal{M}$ if

$$\bar{\boldsymbol{G}}_T^{*'}(\boldsymbol{\theta})\langle \boldsymbol{G}^*(\boldsymbol{\theta})\rangle_T^{-1}\, \bar{\boldsymbol{G}}_T^*(\boldsymbol{\theta}) - \bar{\boldsymbol{G}}_T'(\boldsymbol{\theta})\langle \boldsymbol{G}(\boldsymbol{\theta})\rangle_T^{-1}\bar{\boldsymbol{G}}_T(\boldsymbol{\theta})$$

is almost surely nonnegative definite for all $\boldsymbol{G}_T \in \mathcal{M}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, P_θ and $T > 0$.

It is evident that maximizing the *martingale information*

$$\boldsymbol{I}_G(\boldsymbol{\theta}) = \bar{\boldsymbol{G}}_T'(\boldsymbol{\theta})\langle \boldsymbol{G}(\boldsymbol{\theta})\rangle_T^{-1}\, \bar{\boldsymbol{G}}_T(\boldsymbol{\theta})$$

leads to (asymptotic) confidence regions centered on $\boldsymbol{\theta}^*$ of minimum size (see for example Rao (1973, § 4b. 2)). Note that $\boldsymbol{I}_G(\boldsymbol{\theta})$ can be replaced by $\boldsymbol{I}_G(\boldsymbol{\theta}^*)$ in (2.8).

The relation between O_F and O_A optimality, both restricted to the same family of estimating functions, is very close and it should be noted that

$$E\bar{\boldsymbol{G}}_T(\boldsymbol{\theta}) = E\dot{\boldsymbol{G}}_T(\boldsymbol{\theta}), \qquad E\langle \boldsymbol{G}(\boldsymbol{\theta})\rangle_T = E\boldsymbol{G}_T(\boldsymbol{\theta})\, \boldsymbol{G}_T'(\boldsymbol{\theta}).$$

Below we shall consider the widely applicable class of martingale estimating function

$$\mathcal{A} = \left\{ \boldsymbol{G}_T(\boldsymbol{\theta}) = \int_0^T \boldsymbol{\alpha}_s(\boldsymbol{\theta})\, d\boldsymbol{M}_s(\boldsymbol{\theta}), \quad \boldsymbol{\alpha}_s \text{ predictable} \right\} \tag{2.9}$$

and it will be noted that O_F and O_A optimality coincide. Applications of this class of estimating functions will be discussed in some detail in Section 2.6.

As with Definition 2.1 the criterion of Definition 2.3 is hard to apply directly, but as an analogue of Theorem 2.1 we have the following result, which is easy to use in practice.

Theorem 2.4 Suppose that $\mathcal{M} \subseteq \mathcal{M}_1$. Then, $\boldsymbol{G}_T^*(\boldsymbol{\theta}) \in \mathcal{M}$ is an O_A-optimal estimating function within $\mathcal{M}$ if

$$\left(\bar{\boldsymbol{G}}_T(\boldsymbol{\theta})\right)^{-1} \langle \boldsymbol{G}(\boldsymbol{\theta}), \boldsymbol{G}^{*'}(\boldsymbol{\theta})\rangle_T = \left(\bar{\boldsymbol{G}}_T^*(\boldsymbol{\theta})\right)^{-1} \langle \boldsymbol{G}^*(\boldsymbol{\theta})\rangle_T \tag{2.10}$$

for all $\boldsymbol{G}_T \in \mathcal{M}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, P_θ and $T > 0$. Conversely, if $\mathcal{M}$ is convex and $\boldsymbol{G}_T^* \in \mathcal{M}$ is an O_A-optimal estimating function, then (2.10) holds.

Proof. This follows much the same lines as that of Theorem 2.1 and we shall just sketch the necessary modifications.

Write $\boldsymbol{G} = (G_1, \dots, G_p)'$ and $\boldsymbol{G}^* = (G_1^*, \dots, G_p^*)'$, where the subscript T has been deleted for convenience.

To obtain the first part of the theorem we write the $2p \times 2p$ quadratic characteristic matrix of the set of variables

$$Z' = \left(G_1, \dots, G_p, G_1^*, \dots, G_p^*\right)$$

in partitioned matrix form as

$$\boldsymbol{C} = \begin{bmatrix} \langle \boldsymbol{G}\rangle & \langle \boldsymbol{G}, \boldsymbol{G}^{*'}\rangle \\ \langle \boldsymbol{G}, \boldsymbol{G}^*\rangle' & \langle \boldsymbol{G}^*\rangle \end{bmatrix}.$$

Now $\boldsymbol{C}$ is a.s. nonnegative definite since for $\boldsymbol{u}$, an arbitrary $2p \times 1$ vector,

$$\boldsymbol{u}'\boldsymbol{C}\boldsymbol{u} = \boldsymbol{u}'\langle \boldsymbol{Z}, \boldsymbol{Z}\rangle \boldsymbol{u} = \langle \boldsymbol{u}'\boldsymbol{Z}, \boldsymbol{Z}'\boldsymbol{u}\rangle = \langle \boldsymbol{u}'\boldsymbol{Z}\rangle \geq 0$$

and then the method of Rao (1973, p. 327) gives the a.s. nonnegative definitness of

$$\langle \boldsymbol{G}\rangle - \langle \boldsymbol{G}, \boldsymbol{G}^{*'}\rangle \left(\langle \boldsymbol{G}^*\rangle\right)^{-1} \langle \boldsymbol{G}, \boldsymbol{G}^{*'}\rangle'. \tag{2.11}$$

But, condition (2.10) gives

$$\langle \boldsymbol{G}, \boldsymbol{G}^{*'}\rangle = (\bar{\boldsymbol{G}})(\bar{\boldsymbol{G}}^*)^{-1}\langle \boldsymbol{G}^*\rangle$$

and using this in (2.11) gives O_A-optimality via Definition 2.3.

The converse part of the proof carries through as with Theorem 2.1 using the fact that $\boldsymbol{H} = \alpha \boldsymbol{G} + \boldsymbol{G}^* \in \mathcal{M}$ for arbitrary scalar α.

With the aid of Theorem 2.4 we are now able to show that O_A-optimality implies O_F-optimality in an important set of cases. The reverse implication does not ordinary hold.

Theorem 2.5 Suppose that G_T^* is O_A-optimal within the convex class of martingale estimating functions $\mathcal{M}$. If $(\bar{G}_T^*)^{-1}\langle G^*\rangle_T$ is nonrandom for $T > 0$, then G_T^* is also O_F-optimal within $\mathcal{M}$.

Proof. For each $T > 0$,

$$\langle \boldsymbol{G}^* \rangle_T = \bar{\boldsymbol{G}}_T^* \boldsymbol{\eta}_T, \tag{2.12}$$

say, where $\boldsymbol{\eta}_T$ is a nonrandom $p \times p$ matrix. Then, from Theorem 2.4 we have

$$\langle \boldsymbol{G}, \boldsymbol{G}^{*'} \rangle_T = \bar{\boldsymbol{G}}_T \boldsymbol{\eta}_T, \tag{2.13}$$

and taking expectations in (2.12) and (2.13) leads to

$$E\boldsymbol{G}_T\boldsymbol{G}_T^{*'} = (E\dot{\boldsymbol{G}}_T)\boldsymbol{\eta}_T = (E\dot{\boldsymbol{G}}_T)(E\dot{\boldsymbol{G}}_T^*)^{-1}E\boldsymbol{G}_T^*\boldsymbol{G}_T^{*'}.$$

The required result then follows from Theorem 2.1.

For estimating functions of the form

$$\boldsymbol{G}_T(\boldsymbol{\theta}) = \int_0^T \boldsymbol{\alpha}_s(\boldsymbol{\theta})\, d\boldsymbol{M}_s(\boldsymbol{\theta}),$$

$\boldsymbol{\alpha}_s$ predictable, of the class $\mathcal{A}$ (see (2.9) above) it is easily checked that

$$\boldsymbol{G}_T^*(\boldsymbol{\theta}) = \int_0^T (d\bar{\boldsymbol{M}}_s)'(d\langle \boldsymbol{M}\rangle_s)^-\, d\boldsymbol{M}_s$$

is O_A-optimal within $\mathcal{A}$. Here $\boldsymbol{A}^-$ denotes generalized inverse of a matrix $\boldsymbol{A}$, which satisfies $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$, $\boldsymbol{A}^-\boldsymbol{A}\boldsymbol{A}^- = \boldsymbol{A}^-$. It is often convenient to use $\boldsymbol{A}^+$, the Moore-Penrose generalized inverse of a matrix $\boldsymbol{A}$, namely the unique matrix $\boldsymbol{A}^+$ possessing the properties $\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}$, $\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$, $\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^+$. Note that

$$\bar{\boldsymbol{G}}_T^* = \int_0^T (d\bar{\boldsymbol{M}}_s)'(\langle d\boldsymbol{M}\rangle_s)^-\, d\bar{\boldsymbol{M}}_s = \langle \boldsymbol{G}^* \rangle_T$$

and Theorem 2.5 applies to give O_F-optimality also.

2.6 A Semimartingale Model for Applications

Under ordinary circumstances the process of interest can, perhaps after suitable transformation, be modeled in terms of a *signal* plus *noise* relationship,

$$\textit{process} = \textit{signal} + \textit{noise}.$$

The *signal* incorporates the *predictable trend* part of the model and the *noise* is the *stochastic disturbance* left behind after the signal part of the model is fitted. Parameters of interest are involved in the signal term but may also be present in the noise.

More specifically, there is usually a special semimartingale representation. This framework is one in which there is a filtration $\{\mathcal{F}_t\}$ and the process of interest $\{\boldsymbol{X}_t\}$ is (uniquely) representable in the form

$$\boldsymbol{X}_t = \boldsymbol{X}_0 + \boldsymbol{A}_t(\theta) + \boldsymbol{M}_t(\theta),$$

where $\boldsymbol{A}_t(\theta)$ is a predictable finite variation process which is of locally integrable variation, $\boldsymbol{M}_t(\theta)$ is a local martingale and $\boldsymbol{A}_0 = \boldsymbol{M}_0 = \mathbf{0}$ (e.g., Rogers and Williams (1987, Chapter VI, 40)). The local martingale has a natural role to play in inference as it represents the residual stochastic noise after fitting of the signal which is encapsulated in the finite variation process.

The semimartingale restriction itself is not severe. All discrete time processes and most respectable continuous time processes are semimartingales. Indeed, most statistical models fit very naturally into a semimartingale framework.

Example 1. $\{X_t\}$ is an autoregression of order k. This is a model of the form

$$X_t = \mu + \alpha_1 X_{t-1} + \ldots + \alpha_k X_{t-k} + \epsilon_t$$

consisting of a trend term $\mu + \alpha_1 X_{t-1} + \ldots + \alpha_k X_{t-k}$ and a martingale difference disturbance ϵ_t.

Example 2. $\{X_t\}$ is a Poisson process with intensity θ and

$$X_t = \theta\, t + (X_t - \theta\, t),$$

which consists of a predictable trend $\theta\, t$ and a martingale disturbance $X_t - \theta\, t$.

Example 3. $\{X_t\}$ is a 1-type Galton-Watson branching process; $E(X_1 \mid X_0 = 1) = \theta$. Here

$$\begin{aligned} X_t &= X_{1,t-1}^{(1)} + \ldots + X_{1,t-1}^{(X_{t-1})} \\ &= \theta X_{t-1} + [(X_{1,t-1}^{(1)} - \theta) + \ldots + (X_{1,t-1}^{(X_{t-1})} - \theta)] \\ &= \theta X_{t-1} + \eta_t, \end{aligned}$$

say, where $X_{1,t-1}^{(i)}$ is the number of offspring of individual i in generation t, $i = 1, 2, \ldots, X_{t-1}$. Here θX_{t-1} represents the predictable trend and η_t a martingale difference disturbance.

Example 4. In mathematical finance the price $\{S_t\}$ of a risky asset is commonly modeled by

$$S_t = S_0 \exp\left[\left(\mu - \frac{1}{2}\sigma^2\right) t + \sigma\, W_t\right],$$

where W_t is standard Brownian motion (e.g. the Black-Scholes formulation; see e.g. Duffie (1992)). Then we can use the semimartingale representation

$$X_t = \log S_t = \log S_0 + \left(\mu - \frac{1}{2}\sigma^2\right) t + \sigma\, W_t. \tag{2.14}$$

It is worth noting, however, that the semimartingale property is lost if the Brownian motion noise is replaced by fractional Brownian motion (e.g. Cutland, Kopp and Willinger (1995)).

For various applications the modeling is done most naturally through a stochastic differential equation (sde) representation, which we may think of informally as

$$d\,(process) = d\,(signal) + d\,(noise).$$

In the case of the model (2.14) the corresponding sde is, using the Ito formula,

$$dS_t = S_t\,(\mu\,dt + \sigma\,dW_t)$$

and it should be noted that the parameter σ is no longer explicitly present in the signal part of the representation.

Of course the basic choice of the family of estimating functions from which the quasi-score estimating function is to be chosen always poses a fundamental question. There is, fortunately, a useful particular solution, which we shall call the *Hutton-Nelson solution*. This involves confining attention to the local $M\,G$ estimating functions belonging to the class

$$\mathcal{M} = \left\{ \boldsymbol{G}_t \in \mathcal{K}\, :\, \boldsymbol{G}_t(\boldsymbol{\theta}) = \int_0^t \boldsymbol{\alpha}_s(\boldsymbol{\theta})\, d\boldsymbol{M}_s(\boldsymbol{\theta}), \quad \boldsymbol{\alpha}_s \text{ predictable} \right\},$$

where the local $M\,G$ $\boldsymbol{M}_t$ comes from the semimartingale representation of the process under consideration, or in the discrete case

$$\mathcal{M} = \left\{ \boldsymbol{G}_t \in \mathcal{K}\, :\, \boldsymbol{G}_t(\boldsymbol{\theta}) = \sum_{s=1}^{t} \boldsymbol{\alpha}_s(\boldsymbol{\theta})\, \boldsymbol{m}_s(\boldsymbol{\theta}), \quad \boldsymbol{\alpha}_s\ \mathcal{F}_{s-1} \text{ measurable} \right\},$$

where $\boldsymbol{M}_t = \sum_{s=1}^{t} \boldsymbol{m}_s$. As usual, elements of $\mathcal{M}$ are interpreted, via the semimartingale representation, as functions of the data and parameter $\boldsymbol{\theta}$ of interest.

Within this family $\mathcal{M}$ the quasi-score estimating function is easy to write down, for example, using Theorem 2.1. It is

$$\boldsymbol{G}_t^*(\boldsymbol{\theta}) = \sum_{s=1}^{t} \left(E\left(\dot{\boldsymbol{m}}_s \,\middle|\, \mathcal{F}_{s-1} \right) \right)' \left(E\left(\boldsymbol{m}_s\, \boldsymbol{m}_s' \,\middle|\, \mathcal{F}_{s-1} \right) \right)^{-} \boldsymbol{m}_s$$

(discrete time)

$$= \int_0^t (d\bar{\boldsymbol{M}}_s)' (d\langle \boldsymbol{M} \rangle_s)^{-}\, d\boldsymbol{M}_s$$

(continuous time)

where

$$d\bar{\boldsymbol{M}}_s = E\left(d\dot{\boldsymbol{M}}_s \,\middle|\, \mathcal{F}_{s-} \right),$$

$\langle \boldsymbol{M} \rangle_s$ being the quadratic characteristic and $-$ referring to generalized inverse.

The Hutton-Nelson solution is very widely usable, so widely usable that it may engender a false sense of security. Certainly it cannot be applied thoughtlessly and we shall illustrate the need for care in a number of ways.

First, however, we give a straightforward application to the estimation of parameters in a neurophysiological model. It is known that under certain circumstances the membrane potential $V(t)$ across a neuron is well described by a stochastic differential equation

$$dV(t) = (-\varrho\, V(t) + \lambda)\, dt + dM(t)$$

(e.g. Kallianpur (1983)), where $M(t)$ is a martingale with discontinuous sample paths and a (centered) generalized Poisson distribution. Here $\langle M \rangle_t = \sigma^2 t$ for some $\sigma > 0$. The Hutton-Nelson quasi-score estimating function for $\boldsymbol{\theta} = (\varrho, \lambda)'$ on the basis of a single realization $\{V(s),\, 0 \le s \le T\}$ gives

$$\boldsymbol{G}_T^* = \int_0^T (-V(t)\; 1)' \, \{dV(t) - (-\varrho\, V(t) + \lambda)\, dt\,\}.$$

The estimators $\hat{\varrho}$ and $\hat{\lambda}$ are then obtained from the estimating equations

$$\begin{aligned} \int_0^T V(t)\, dV(t) &= \int_0^T \left(-\hat{\varrho}\, V(t) + \hat{\lambda}\right) V(t)\, dt, \\ V(T) - V(0) &= \int_0^T \left(-\hat{\varrho}\, V(t) + \hat{\lambda}\right) dt, \end{aligned}$$

and it should be noted that these do not involve detailed properties of the stochastic disturbance $M(t)$, only a knowledge of $\langle M \rangle_t$. In particular, they remain the same if $M(t)$ is replaced (as holds in a certain limiting sense; see Kallianpur (1983)) by $\sigma^2 W(t)$, where $W(t)$ is standard Brownian motion. In this latter case $\hat{\varrho}$ and $\hat{\lambda}$ are actually the respective maximum likelihood estimators.

There are, of course, some circumstances in which a parameter in the noise component of the semimartingale model needs to be estimated, such as the scale parameter σ associated with the noise in the above membrane potential model. This can ordinarily be handled by transforming the original semimartingale model into a new one in which the parameter of interest is no longer in the noise component. For the membrane potential model, for example, we can consider a new semimartingale $\{[M]_t - \langle M \rangle_t\}$, $\{[M]_t\}$ being the quadratic variation process. Then, since

$$\int_0^T (dV(t))^2 = \int_0^T (dM(t))^2 = [M]_T$$

and

$$\langle M \rangle_T = \int_0^T d\langle M \rangle_t = \sigma^2 T,$$

it is clear that σ^2 can be estimated by $T^{-1} \int_0^T (dV(t))^2$.

Now for stochastic processes it is important to focus on the interplay between the *data* that is observed and the *sources of variation* in the model for which estimators are sought.

Stochastic process estimation raises a rich diversity of confounding and nonidentifiability problems that are often less transparent than those which occur in a nonevolutionary statistical environment. Furthermore, there are more often limitations on the data that can be collected and consequent limitations on the parameters that can be estimated.

It is worth illustrating with some examples in which one must be careful to focus individually on the different sources of variation in the model. The first example (Sørensen (1990)) concerns the process

$$dX_t = \theta\, X_t\, dt + dW_t + dN_t, \quad t \geq 0,\ X_0 = x_0,$$

where W_t is a standard Brownian motion and N_t is a Poisson process with intensity λ. This process may be written in semimartingale form as

$$dX_t = \theta\, X_t\, dt + \lambda\, dt + dM_t,$$

where

$$M_t = W_t + N_t - \lambda\, t$$

and the Hutton-Nelson quasi-score estimating function based on M_t is

$$\boldsymbol{G}_T^* = -\frac{1}{1+\lambda} \int_0^T \begin{pmatrix} X_{s-} \\ 1 \end{pmatrix} dM_s.$$

Then, the estimating equations are

$$\begin{aligned} \int_0^T X_{s-}\, dX_s &= \hat{\theta} \int_0^T X_s^2\, ds + \hat{\lambda} \int_0^T X_s\, ds \\ X_T &= \hat{\theta} \int_0^T X_s\, ds + \hat{\lambda}\, T. \end{aligned}$$

These, however, are the maximum likelihood estimating equations for the model

$$dX_t = (\theta\, X_t + \lambda)\, dt + dW_t,$$

that is, where N_t has been replaced by its compensator $\lambda\, t$. The entire stochastic fluctuation is described by the Brownian motion and this would only be realistic if $\lambda \ll 1$.

However, if we rewrite the original model by separating the discrete and continuous parts and treating them individually, we get true MLE's. Put

$$dX_t^c = \theta\, X_t\, dt + dW_t \qquad X_t^c = X_t - N_t \tag{2.15}$$

$$dX_t - dX_t^c = dN_t = \lambda\, dt + dN_t - \lambda\, dt. \tag{2.16}$$

The Hutton-Nelson quasi-score estimating functions based on (2.15) and (2.16) individually are, respectively,

$$\begin{aligned} \int_0^T X_{s-}\, dX_s^c &= \tilde{\theta} \int_0^T X_s^2\, ds, \\ N_T = \tilde{\lambda}\, T. & \end{aligned}$$

The improvement over the earlier approach can easily be assessed. For example, $E(\tilde{\lambda}-\lambda)^2 \sim \lambda/T$ while $E(\hat{\lambda}-\lambda)^2 \sim (1+\lambda)/T$ as $T \to \infty$.

In the next example (Sørensen (1990)) $\{X_t\}$ behaves according to the compound Poisson process

$$X_t = \sum_{i=1}^{N_t} Y_i,$$

where $\{N_t\}$ is a Poisson process with intensity λ and the $\{Y_i\}$ are i.i.d. with mean θ (and independent of $\{N_t\}$). Here the semimartingale representation is

$$X_t = \lambda\,\theta\,t + \sum_{i=1}^{N_t} (Y_i - \theta) + \theta(N_t - \lambda\,t) = \lambda\,\theta\,t + M_t,$$

say, $\{M_t\}$ being a martingale. Using $\{M_t\}$, the estimating equation based on the Hutton-Nelson quasi-score is

$$X_T = (\lambda\,\theta)\,T$$

so that only the product $\lambda\,\theta$ is estimable.

However, if the data provide both $\{X_t,\ 0 \le t \le T\}$ and $\{N_t,\ 0 \le t \le T\}$, then we can separate out the two martingales $\left\{\sum_1^{N_t} (Y_i - \theta)\right\}$, $\{N_t - \lambda\,t\}$ and obtain the Hutton-Nelson quasi-score estimating functions based on each. The QL estimators are

$$\hat{\theta} = X_T/N_T, \qquad \hat{\lambda} = N_T/T$$

and these are MLE's when the Y_t's are exponentially distributed.

General conclusion: Identify relevant martingales that focus on the sources of variation and combine them.

Relevant martingales can be constructed in many ways. The simplest single approach is, in the discrete time setting, to take the increments of the observation process, substract their conditional expectations and sum. Ordinary likelihoods and their variants such as marginal, conditional or partial likelihoods can also be used to generate martingales (see, e.g., Barndorff-Nielsen and Cox (1994, Sections 8.6 and 3.3)). General methods for optimal combination of estimating functions are discussed in Chapter 6.

2.7 Some Problem Cases for the Methodology

It is by no means assured that a quasi-score estimating function is practicable or computable in terms of the available data even if it is based on the semimartingale representation as described in Section 2.6.

To provide one example we take the population process in a random environment

$$X_t = (\theta + \eta_t)\,X_{t-1} + \epsilon_t,$$

where $(\theta + \eta_t)$ is multiplicative noise coming from environmental stochasticity and ϵ_t is additive noise coming from demographic stochasticity. Let $\{\mathcal{F}_t\}$ denote the past history σ-fields. We shall take

$$E\left(\eta_t \,\middle|\, \mathcal{F}_{t-1}\right) = 0 \quad \text{a.s.,} \qquad E\left(\epsilon_t \,\middle|\, \mathcal{F}_{t-1}\right) = 0 \quad \text{a.s.,}$$

$$E\left(\epsilon_t^2 \,\middle|\, \mathcal{F}_{t-1}\right) = \sigma^2\, X_{t-1},$$

the last result being by analogy with the Galton-Watson branching process. The problem is to estimate θ on the basis of data $\{X_0, \dots, X_T\}$. Here the semimartingale representation is

$$X_t = \theta\, X_{t-1} + u_t,$$

where the $u_t = \eta_t\, X_{t-1} + \epsilon_t$ are $M\,G$ differences. The Hutton-Nelson quasi-score estimating function based on $\{u_t\}$ is

$$Q_T^*(\theta) = \sum_1^T \frac{X_{t-1}}{E\left(u_t^2 \,\middle|\, \mathcal{F}_{t-1}\right)}\,(X_t - \theta\, X_{t-1})$$

and

$$E\left(u_t^2 \,\middle|\, \mathcal{F}_{t-1}\right) = X_{t-1}^2\, E\left(\eta_t^2 \,\middle|\, \mathcal{F}_{t-1}\right) + 2X_{t-1}\, E\left(\eta_t\, \epsilon_t \,\middle|\, \mathcal{F}_{t-1}\right) + E\left(\epsilon_t^2 \,\middle|\, \mathcal{F}_{t-1}\right)$$

may not be explicitly computable in terms of the data; indeed a tractable form requires independent environments.

If we attempt to restrict the family of estimating functions to ones that are explicit functions of the data

$$\left\{ H \;:\; H_T = \sum_{t=1}^T a_t(X_{t-1}, \dots, X_0;\, \theta)\,(X_t - \theta\, X_{t-1}) \right\},$$

then there is not in general a solution to the optimality problem of minimizing var H_T^S.

Suppose now that the environments in the previous example are independent and $\eta^2 = E(\eta_t^2 \mid \mathcal{F}_{t-1})$. Then,

$$E\left(u_t^2 \,\middle|\, \mathcal{F}_{t-1}\right) = \eta^2\, X_{t-1}^2 + \sigma^2\, X_{t-1},$$

so that

$$Q_T^* = \sum_1^T \frac{X_{t-1}}{\eta^2 X_{t-1}^2 + \sigma^2 X_{t-1}}\,(X_t - \theta\, X_{t-1}).$$

The QLE is

$$\hat{\theta}_T = \sum_1^T \frac{X_t\, X_{t-1}}{\eta^2 X_{t-1}^2 + \sigma^2 X_{t-1}} \bigg/ \sum_1^T \frac{X_{t-1}^2}{\eta^2 X_{t-1}^2 + \sigma^2 X_{t-1}},$$

which contains the nuisance parameters σ^2, η^2. However, on the set $\{X_t \to \infty\}$, $\hat{\theta}_T$ will have the same asymptotic behavior as

$$T^{-1} \sum_{t=1}^{T} \frac{X_t}{X_{t-1}}.$$

This kind of consideration suggests the need to formulate a precise concept of asymptotic quasi-likelihood and this is done in Chapter 5.

Another awkward example concerns the estimation of the growth parameter in a logistic map system whose observation is subject to additive error (e.g., Berliner (1991), Lele (1994)). Here we have observations $\{Y_t,\, t = 1, 2, \ldots, T\}$ where

$$Y_t = X_t + e_t \tag{2.17}$$

with the $\{X_t\}$ process given by the deterministic logistic map

$$X_{t+1} = \theta\, X_t(1 - X_t), \quad t = 0, 1, 2, \ldots,$$

while the e's are i.i.d. with zero mean and variance σ^2 (assumed known).

Based on the representation (2.17) one might consider the family of estimating functions

$$\mathcal{H} = \left\{ \sum_{t=1}^{T} c_t(Y_t - X_t), \quad c_t\text{'s constants} \right\}, \tag{2.18}$$

although some adjustments are clearly necessary to avoid nuisance parameters, the X's not being observable. Nevertheless, proceeding formally, the quasi-score estimating function from $\mathcal{H}$ is

$$Q_T(\theta) = \sum_{t=1}^{T} \frac{dX_t}{d\theta}\, (Y_t - \theta\, X_{t-1}(1 - X_{t-1})).$$

The weight functions $dX_t/d\theta$ are, however, not practicable to work with. Note that the system $\{X_t\}$ is uniquely determined by X_0 and θ and that X_t may be written (unhelpfully) as a polynomial of degree $2^t - 1$ in θ and 2^t in X_0.

It may be concluded that, despite its apparent simplicity, $\mathcal{H}$ is a poor choice as a family of estimating functions and that one based more directly on the available data may be more useful. With this in mind, and noting that

$$Y_t(1 - Y_t) + \sigma^2 - X_t(1 - X_t) = e_t(1 - 2X_t) + (\sigma^2 - e_t^2),$$

so that $Y_t(1 - Y_t) + \sigma^2$ is an unbiased estimator of $X_t(1 - X_t)$, we can consider the family

$$\mathcal{K} = \left\{ \sum_{t=1}^{T} c_t(Y_t - \theta(Y_{t-1}(1 - Y_{t-1}) + \sigma^2)), \quad c_t\text{'s constants} \right\} \tag{2.19}$$

of estimating functions. The element of this family for which $c_t = 1$ for all t has been shown by Lele (1994) to provide an estimator that is strongly consistent for θ and asymptotically normally distributed. Quasi-likelihood methods do offer the prospect of some improvement in efficiency at the price of considerable increase in complexity (see Exercise 4, Section 2.8).

2.8 Complements and Exercises

Exercise 1. Show that an equivalent form to Criterion 2.2 when the score function $\boldsymbol{U}_T$ exists is

$$E\left(\boldsymbol{U}_T - \boldsymbol{G}_T^{*(s)}\right) \boldsymbol{G}_T^{(s)'} = \mathbf{0} = E\boldsymbol{G}_T^{(s)} \left(\boldsymbol{U}_T - \boldsymbol{G}_T^{*(s)}\right)'.$$

(Note that $\boldsymbol{U}_T = \boldsymbol{U}_T^{(s)}$.) (Godambe and Heyde (1987)).

Exercise 2. Extend Theorem 2.2 to include the smallest eigenvalue criterion and average variance criterion in the remark immediately following the statement of the theorem as additional equivalent conditions (e.g., Pukelsheim (1993)).

Exercise 3. (Geometric waiting times) If time is measured in discrete periods, a model that is often used for the time X to failure of an item is

$$P(X = k) = \theta^{k-1}(1-\theta), \quad k = 1, 2, \ldots.$$

Suppose that we have a set of i.i.d. observations $X_1, \ldots, X_n$ from X.

(i) Show that this is an exponential family model and find the MLE $\hat{\theta}$ of θ.

(ii) Show that $EX = (1-\theta)^{-1}$ and that the QLE for θ from the family $\mathcal{H} = \{\sum_i^n c_i(X_i - EX_i),\ c_i \text{ constants}\}$ coincides with the MLE.

Now suppose that we have censoring of the data. Suppose that we only record the time of failure if failure occurs on or before time r and otherwise we just note that the item has survived at least time $r+1$. Thus we observe $Y_1, \ldots, Y_n$, which have the distribution

$$P_\theta(Y_i = k) = \theta^{k-1}(1-\theta), \quad k = 1, 2, \ldots, r$$

$$P_\theta(Y_i = r+1) = 1 - P_\theta(X \le r+1) = \theta^r.$$

Let M be the number of indices i such that $Y_i = r+1$.

(iii) Show that the MLE of θ based on $Y_1, \ldots, Y_n$ is

$$\hat{\theta}(Y_1, \ldots, Y_n) = \frac{\sum_{i=1}^n Y_i - n}{\sum_{i=1}^n Y_i - M}.$$

(iv) Show that the QLE for θ based on the family

$$\mathcal{H} = \left\{ \sum_1^n c_i(Y_i - EY_i), \quad c_i \text{ constants} \right\}$$

does not coincide with the MLE $\hat{\theta}(Y_1, \dots, Y_n)$.

(v) Now subdivide the data into those indices i such that $Y_i = r+1$ and those indices i for which $Y_i \le r$ and consider these sets separately. To make this clear, define

$$Z_i = Y_i \, I(Y_i = r+1), \quad W_i = Y_i \, I(1 \le Y_i \le r),$$

$i = 1, 2, \dots, n$, I denoting the indicator function. Next consider the families of estimating functions

$$\mathcal{H}_1 = \left\{ \sum_1^n c_i(Z_i - EZ_1), \quad c_i \text{ constants} \right\},$$

$$\mathcal{H}_2 = \left\{ \sum_1^n d_i(W_i - EW_1), \quad d_i \text{ constants} \right\},$$

and show that these lead to QSEF's

$$(r+1)(M - n\,\theta^r) \tag{2.20}$$

and

$$\sum_1^n Y_i - (r+1)\,M - n\left(\frac{1-\theta^r}{1-\theta} - r\,\theta^r\right), \tag{2.21}$$

respectively. Note that the first of these suggests the use of M to estimate $n\theta^r$ and substituting this into

$$\sum_1^n Y_i - (r+1)\,M - n\left(\frac{1-\hat{\theta}^r}{1-\hat{\theta}} - r\,\hat{\theta}^r\right) = 0$$

with $M = n\hat{\theta}^r$ leads to the $\hat{\theta}(Y_1, \dots, Y_n)$ defined above in (iii).

(vi) Show that the estimating functions (2.20) and (2.21) can also be added to produce $\hat{\theta}(Y_1, \dots, Y_n)$. That is, there are multipliers α^* and β^* such that

$$\alpha^*(r+1)(M - n\bar{\theta}^r) + \beta^*\left(\sum_1^n Y_i - (r+1)M - n\left(\frac{1-\bar{\theta}^r}{1-\bar{\theta}} - r\bar{\theta}^r\right)\right) = 0$$

has solution $\bar{\theta} = \hat{\theta}(Y_1, \dots, Y_n)$. Find α^*/β^*.

(vii) Show that the solution in (vi) is the QS estimating function for the family

$$\mathcal{H} = \left\{ \alpha \sum_1^n (Z_i - EZ_1) + \beta \sum_1^n (W_i - EW_1), \quad \alpha, \beta \text{ constants} \right\}.$$

Exercise 4. For the logistic map observed subject to noise investigate the quasi-score estimating function from the family (2.19) together with practicable variants. If

$$\epsilon_t = Y_t - \theta \left[Y_{t-1} \left(1 - Y_{t-1}\right) + \sigma^2 \right],$$

note that ϵ_s and ϵ_t are not independent unless $|t - s| > 1$. Obtain strong consistency and asymptotic normality results for estimators if possible and compare asymptotic variances with that obtained for the estimator based on the case $c_t = 1$, all t.

Infinite moments

The quasi-likelihood theory developed herein is based on estimating functions with zero means and finite variances. When the data are derived from distributions that do not possess finite second moments, linear functions of the data cannot be used as a basis for constructing estimating functions and transformation of the data is required. Exercise 5 provides a simple illustation of the possibilities. More information is provided in Chapter 13.

Exercise 5. Let $X_i,\ i = 1, 2, \ldots, n$ be i.i.d.r.v. having a Cauchy distribution with density $b/(\pi(b^2 + x^2)),\ -\infty < x < \infty$. Show that $n^{-1} \sum_{i=1}^n \cos X_i$ can be obtained as a QLE for e^{-b}. Extending this reasoning, show that, for any real t, $n^{-1} \sum_{i=1}^n \cos t X_i$ is a QLE for e^{-tb} and investigate the question of a suitable choice for t.

Non-regular cases

In assessing the appropriateness of likelihood-based methods for parameter inference in a particular context, it is necessary to check the following basic set of requirements:

(1) Is the likelihood differentiable with respect to the parameter (i.e., does the score function exist)?

(2) Does Fisher information exist?

(3) Does the Fisher information of the sample increase unboundedly as the sample size increases?

(4) Can the parameter of interest take a value on the boundary of the parameter set?

(5) Is the likelihood zero outside a domain that depends on the unknown parameter?

Failure of any of these regularity conditions is typically a warning signal that non-standard behavior may be expected. Similar considerations apply also to quasi-likelihood methods in respect to (2)-(4) and the information $\mathcal{E}$. The following exercises illustrate consequences of warning signals (4), (5).

Exercise 6. Let $X_1, \ldots, X_n$ be i.i.d. uniform $U(0, \theta)$ random variables.

(i) Show that the maximum likelihood estimator $X_{(n)} = \max_{k \leq n} X_k$ does not satisfy the likelihood equation (obtained from equating the score function to zero).

(ii) Show that $n(\theta - X_{(n)})$ converges in distribution to an exponential law.

(iii) Let $S_n = [n + 1/n]\, X_{(n)}$. Show that δ_n is preferable asymptotically to $X_{(n)}$ by verifying that

$$E\left[n(X_{(n)} - \theta)^2\right] \to 2\theta^2, \quad E\left[n(\delta_n - \theta)^2\right] \to \theta^2$$

as $n \to \infty$.

Exercise 7. Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. such that $Y_j = |X_j|$, where X_j has a $N(\mu, 1)$ distribution. The situation arises, for example, if X_j is the difference between a matched pair of random variables whose control and treatment labels are lost. The sign of the parameter μ is unidentifiable so we may work with $\theta = |\mu|$, with parameter space $\Theta = [0, \infty)$. Show that the score function is

$$\sum_{j=1}^{n} [Y_j + \tanh(\theta Y_j) - \theta]$$

and observe that for $\theta > 0$ the problem is a regular one and

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{\text{d}} N(0, i(\theta))$$

with

$$i(\theta) = 1 - \frac{2}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\theta^2} \int_0^\infty y^2 e^{-\frac{1}{2}y^2} \operatorname{sech}(\theta y)\, dy.$$

For $\theta = 0$ all this breaks down and the score function is identically zero. However, show that a first-order expansion of the score function gives

$$\hat{\theta}_n^2 = \frac{3\left(\sum_1^n Y_j^2 - n\right)}{\sum_1^n Y_j^4}$$

provided $\sum_1^n Y_j^2 > n$. Show that $\hat{\theta}^2 \sqrt{n}$ has asymptotically a $N(0, 2)$ distribution truncated at zero (Cox and Hinkley (1974, pp. 303-304)).

Chapter 3

An Alternative Approach: E-Sufficiency

3.1 Introduction

In classical statistics, the concept of "sufficiency" and its dual concept of "ancillarity" play a key role and sufficient statistics provide the essential approach to parameter inference in cases where they are available. To adapt these basic concepts to the context of estimating functions, Small and McLeish developed the expectation based concepts of E-sufficiency, E-ancillarity, local E-sufficiency and local E-ancillarity. The basic idea is that an ancillary statistic is one whose distribution is insentive to changes in the parameter and that, in an estimating function context, changes to the parameter are first evident through changes to the expectation of the estimating function. For a space Ψ of estimating functions the E-ancillary subset, $\mathcal{A}$ (say), are defined on the basis of expectation insensitivity to parameter changes and the E-sufficient estimating functions belong to the orthogonal complement of $\mathcal{A}$ with respect to Ψ. It is argued that an optimum estimating function from Ψ should be chosen from the E-sufficient or locally E-sufficient subset of Ψ, if this exists. Detailed expositions are given in McLeish and Small (1988), Small and McLeish (1994, Chapter 4).

It turns out that our framework of optimal or quasi-score estimating functions in this book corresponds most closely to the notation of locally E-sufficient estimating functions and this chapter is concerned with their relationship. Our study shows that whether a quasi-score estimating function is locally E-sufficient depends strongly on the estimating function space chosen, and also that, under certain regularity and other conditions, the quasi-score estimating function will be locally E-sufficient. In many cases the two approaches both lead to the same estimating equation. The treatment here follows Lin (1994a).

3.2 Definitions and Notation

Let $\mathcal{X}$ be a sample space and $\mathcal{P}$ be a class of probability measures P on $\mathcal{X}$. For each P, let v_P denote the p-dimensional vector space of vector-valued functions $\{\boldsymbol{f}\}$ which are defined on the sample space $\mathcal{X}$ and satisfy $E_P \|\boldsymbol{f}\|^2 < \infty$, where $\|\boldsymbol{f}\|^2 = \boldsymbol{f}'(x)\,\boldsymbol{f}(x)$.

Let $\boldsymbol{\theta}$ be a real-valued function on the class of probability measures $\mathcal{P}$ and $\Theta = \{\boldsymbol{\theta}(P),\ P \in \mathcal{P}\} \subseteq \mathbb{R}^p$ be the parameter space.

Now we consider a space Ψ in which every estimating function $\boldsymbol{\psi}(\boldsymbol{\theta}, x) = \boldsymbol{\psi}(\boldsymbol{\theta})$ is a mapping from the parameter space and the sample space into $\mathbb{R}^p$.

Each element of Ψ is unbiased and square integrable, i.e.,

$$E_P[\boldsymbol{\psi}(\boldsymbol{\theta}(P))] = \mathbf{0},$$

and

$$\|\boldsymbol{\psi}\|^2_{\theta(P)} \equiv E_P[\boldsymbol{\psi}'(\boldsymbol{\theta}(P))\,\boldsymbol{\psi}(\boldsymbol{\theta}(P))] < \infty, \tag{3.1}$$

for all P and $\boldsymbol{\theta} = \boldsymbol{\theta}(P)$. Assume that the space has constant covariance structure, i.e., the inner product $E_P[\boldsymbol{\psi}'(\boldsymbol{\theta}(P))\,\boldsymbol{\psi}(\boldsymbol{\theta}(P))]$ depends on P only through $\boldsymbol{\theta}(P)$, for all $\boldsymbol{\psi}$, $\boldsymbol{\phi} \in \Psi$. As usual, we also require that Ψ be a Hilbert space of estimating functions and weak square closed, i.e., for any sequence of functions $\boldsymbol{\psi}_n \in \Psi$, if $\lim_{n\to\infty,\, m\to\infty} \|\boldsymbol{\psi}_n - \boldsymbol{\psi}_m\|_\theta = 0$, then there is a function $\boldsymbol{\psi} \in \Psi$ such that $\lim_{n\to\infty} \|\boldsymbol{\psi}_n - \boldsymbol{\psi}\|_\theta = 0$. The choice of $\boldsymbol{\psi}$ will, in general, depend on $\boldsymbol{\theta}$; see McLeish and Small (1988), Small and McLeish (1994).

Definition 3.1 An unbiased estimating function $\boldsymbol{\psi}(\boldsymbol{\theta})$ is locally E-ancillary if it is a weak square limit of functions $\{\boldsymbol{\psi}\} \in \Psi$ that satisfy

$$E_P[\boldsymbol{\psi}(\boldsymbol{\theta})] = o(|\boldsymbol{\theta}(P) - \boldsymbol{\theta}|), \qquad \text{for} \quad \boldsymbol{\theta}(P) \to \boldsymbol{\theta}. \tag{3.2}$$

We note that if the order of the differentiation and integration can be interchanged, then equation (3.2) can be rewritten as

$$\left(E_\theta\left(\frac{\partial \psi_i}{\partial \theta_j}\right)\right)_{p\times p} = \mathbf{0}.$$

Therefore, under suitable regularity conditions, Definition 3.1 can be replaced by Definition 3.2 below.

Definition 3.2 An unbiased estimating function $\boldsymbol{\psi}(\boldsymbol{\theta})$ is locally E-ancillary if it is a weak square limit of functions $\{\boldsymbol{\psi}\} \in \Psi$ that satisfy

$$\left(E_\theta\left(\frac{\partial \psi_i}{\partial \theta_j}\right)\right)_{p\times p} = \mathbf{0}.$$

The locally E-ancillary estimating functions form a subset of the estimating function space, which will be denoted by $\mathcal{A}_{\text{loc}}$.

Following the definition of local E-ancillarity, when $\boldsymbol{\psi} \in \mathcal{A}_{\text{loc}}$, i.e., $(\psi_1, \ldots, \psi_p)' \in \mathcal{A}_{\text{loc}}$, then all of the estimating functions $\boldsymbol{a}_{i,j} = (0, \ldots, a_i, \ldots, 0)$, where $a_i = \psi_j$ and $i, j = 1, \ldots, p$, belong to $\mathcal{A}_{\text{loc}}$.

Definition 3.3 A subset $\mathcal{S}_{\text{loc}}$ of the estimating function space is complete locally E-sufficient when $\psi \in \mathcal{S}_{\text{loc}}$ if and only if

$$E_\theta(\boldsymbol{\psi}'(\boldsymbol{\theta})\,\boldsymbol{\phi}(\boldsymbol{\theta})) = 0 \qquad \text{for all} \quad P,\, \boldsymbol{\theta} = \boldsymbol{\theta}(P) \quad \text{and} \quad \boldsymbol{\phi} \in \mathcal{A}_{\text{loc}}.$$

Because of the special property of $\mathcal{A}_{\text{loc}}$, when $\psi \in \mathcal{S}_{\text{loc}}$, for each element of $\mathcal{A}_{\text{loc}}$, say $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$, we have $E(\boldsymbol{\psi}'\boldsymbol{a}_{ij}) = (0, \ldots, a_i, \ldots, 0)'$, $a_i = \phi_j$, and $i, j = 1, \ldots, p$, i.e,

$$E(\psi_i \phi_j) = 0 \qquad (i, j = 1, \ldots, p).$$

So $E(\boldsymbol{\psi}\, \boldsymbol{\phi}') = \mathbf{0}$, which in turn implies that $E(\boldsymbol{\psi}'\boldsymbol{\phi}) = 0$. Therefore, the definition of $\mathcal{S}_{\text{loc}}$ has the following equivalent formulation.

Definition 3.3′ A subset $\mathcal{S}_{\text{loc}}$ of the estimating function space is complete locally E-sufficient when $\boldsymbol{\psi} \in \mathcal{S}_{\text{loc}}$ if and only if

$$E_\theta(\boldsymbol{\psi}(\boldsymbol{\theta})\, \boldsymbol{\phi}(\boldsymbol{\theta})') = \mathbf{0} \qquad \text{for all} \quad P, \boldsymbol{\theta} = \boldsymbol{\theta}(P) \quad \text{and} \quad \boldsymbol{\phi} \in \mathcal{A}_{\text{loc}}.$$

It follows from the definition of $\mathcal{S}_{\text{loc}}$ that $\mathcal{S}_{\text{loc}} = \mathcal{A}_{\text{loc}}^{\perp}$, and by the assumptions on Ψ, we have

$$\Psi = \mathcal{S}_{\text{loc}} \oplus \mathcal{A}_{\text{loc}}.$$

However the decomposition is associated with $\boldsymbol{\theta} \in \Theta$. A more detailed discussion of the space Ψ can be found in Small and McLeish (1991).

In general there is no guarantee of the existence of a complete locally E-sufficient subspace. However, if it does exist it is a closed linear subspace of Ψ. Completeness is dispensed with in the following definition.

Definition 3.4 A subset L of Ψ is locally E-sufficient if for $\boldsymbol{\phi} \in \Psi$,

$$E_\theta(\boldsymbol{\phi}'(\boldsymbol{\theta})\, \boldsymbol{\psi}(\boldsymbol{\theta})) = 0 \qquad \text{for all} \quad \boldsymbol{\psi} \in L \quad \text{and} \quad \boldsymbol{\theta} \in \Theta$$

implies that $\boldsymbol{\phi}$ is locally E-ancillary.

As a simple illustration we consider the square integrable scalar process $\{X_t\}$, with X_t adapted to the increasing sequence of a fields $\mathcal{F}_t$, and such that

$$E_\theta \left(X_t \,\middle|\, \mathcal{F}_{t-1} \right) = \mu_t \theta, \qquad \text{var}_{\,\theta} \left(X_t \,\middle|\, \mathcal{F}_{t-1} \right) = \Sigma_t(\theta)$$

for θ an unknown parameter and $\mathcal{F}_{t-1}$-measurable μ_t, $\Sigma_t(\theta)$ whose forms are specified. Let Ψ be the space of square integrable estimating functions of the form

$$\left\{ \psi(\theta) = \sum_{i=1}^{n} A_i(\theta)(X_i - \mu_i \theta), \quad A_i\text{'s are } \mathcal{F}_{i-1} \text{ measurable} \right\}. \tag{3.3}$$

Then, if the function

$$\psi^*(\theta) = \sum_{i=1}^{n} \mu_i \Sigma_i^{-1}(X_i - \mu_i \theta)$$

is in the space Ψ, it generates the locally E-sufficient subspace.

To check this, we note that of $\psi(\theta)$ is of the form (3.3), then

$$\begin{aligned} E_\theta(\psi^*(\theta)\,\psi(\theta)) &= E_\theta \sum_{i=1}^{n} \mu_i\,\Sigma_i^{-1}\,\Sigma_i A_i(\theta) \\ &= E_\theta \sum_{i=1}^{n} \mu_i\,A_i(\theta) \\ &= \frac{d}{d\xi} E_\xi \sum_{i=1}^{n} (\xi-\theta)\mu_i\,A_i(\theta)\,\Big|_{\xi=\theta} \\ &= \frac{d}{d\xi} E_\xi\,\psi(\theta)\,\Big|_{\xi=\theta} \end{aligned}$$

and if this is zero then ψ is locally E-ancillary.

3.3 Results

In the following, assume that Ψ satisfies the usual regularity conditions, so that Definition 3.2 can be used to define the concept of local E-ancillarity.

Let $\mathcal{G}_1$ be a subset of Ψ. Each element $\boldsymbol{G}$ of $\mathcal{G}_1$ satisfies the conditions that $E\dot{\boldsymbol{G}}(\boldsymbol{\theta}) = (E(\partial G_i/\partial\theta_j))_{p\times p}$ and $E\boldsymbol{G}(\boldsymbol{\theta})\,\boldsymbol{G}'(\boldsymbol{\theta})$ are nonsingular. The score function is as usual denoted by $\boldsymbol{U}(\boldsymbol{\theta})$.

The fact that $\Psi = \mathcal{S}_{\text{loc}} \oplus \mathcal{A}_{\text{loc}}$ leads us to seek conditions under which a quasi-score estimating function is in $\mathcal{S}_{\text{loc}}$ or $\mathcal{A}_{\text{loc}}$. In the following discussion, a quasi-score estimating function in $\mathcal{G}_1$ is always denoted by $\boldsymbol{G}^*$ and, for an estimating function $\boldsymbol{G}$, if $\boldsymbol{G} = \boldsymbol{0}$ a.s., we write $\boldsymbol{G} = \boldsymbol{0}$.

Theorem 3.1 If $\mathcal{G}_1$ is a convex set and the quasi-score estimating function $\boldsymbol{G}^* \in \mathcal{G}_1 \cap \mathcal{A}_{\text{loc}}$ (resp. $\boldsymbol{G}^* \in \mathcal{G}_1 \cap \mathcal{S}_{\text{loc}}$), then $\mathcal{G}_1 \cap \mathcal{S}_{\text{loc}} = \emptyset$ (resp. $\mathcal{G}_1 \cap \mathcal{A}_{\text{loc}} = \emptyset$).

Proof. We give only the first part of the proof. The proof of the second part is analogous.

Applying Theorem 2.1, if $\boldsymbol{G}^* \in \mathcal{G}_1 \cap \mathcal{A}_{\text{loc}}$, but $\mathcal{G}_1 \cap \mathcal{S}_{\text{loc}} \neq \emptyset$, then there is a $\boldsymbol{G}^* \in \mathcal{G}_1 \cap \mathcal{S}_{\text{loc}}$ such that

$$\left(E\dot{\boldsymbol{G}}\right)^{-1}\left(E\boldsymbol{G}\,\boldsymbol{G}^{*'}\right) = \left(E\dot{\boldsymbol{G}}^*\right)^{-1}\left(E\boldsymbol{G}^*\boldsymbol{G}^{*'}\right).$$

Following the definition of $\mathcal{S}_{\text{loc}}$, we have $E\boldsymbol{G}\,\boldsymbol{G}^{*'} = \boldsymbol{0}$. So the left hand side of the above equation is equal to zero. This implies

$$E\boldsymbol{G}^*\boldsymbol{G}^{*'} = \boldsymbol{0},$$

and $\boldsymbol{G}^*$ has to be a zero vector. This is contrary to the definition of a quasi-score estimating function. Therefore $\mathcal{G}_1 \cap \mathcal{S}_{\text{loc}} = \emptyset$.

From Theorem 3.1 we note that, if there are no other restrictions on $\mathcal{G}_1$, it is possible that $\boldsymbol{G}^*$ is outside $\mathcal{S}_{\text{loc}}$. However in general, under regularity conditions, $\mathcal{G}_1 \cap \mathcal{A}_{\text{loc}} = \emptyset$. This implies that $\boldsymbol{G}^*$ can be only in the complement of $\mathcal{A}_{\text{loc}}$. But this conclusion has not answered the questions of whether $\boldsymbol{G}^* \in \mathcal{G}_{\text{loc}}$ and what the necessary conditions are for $\boldsymbol{G}^*$ to belong to $\mathcal{S}_{\text{loc}}$. The conditions that force $\boldsymbol{G}^* \in \mathcal{S}_{\text{loc}}$ are of interest and are provided in the following theorems.

Theorem 3.2 Suppose that $\mathcal{G}_1$ is a convex set and $\boldsymbol{G}^* \in \mathcal{G}_1 - \mathcal{G}_1 \cap \mathcal{A}_{\text{loc}}$. Let $\boldsymbol{G}^*_{ps}$ be the projection of $\boldsymbol{G}^*$ on $\mathcal{S}_{\text{loc}}$. If $\boldsymbol{G}^*_{ps} \in \mathcal{G}_1$ and $\boldsymbol{G}^*_{ps}$ satisfies the condition that $\boldsymbol{G}^*_{ps} + \boldsymbol{G} \in \mathcal{G}_1$ for any $\boldsymbol{G}$ with $E\dot{\boldsymbol{G}} = 0$, then $\boldsymbol{G}^* \in \mathcal{S}_{\text{loc}}$.

Proof. Assume that $\boldsymbol{G}^* = \boldsymbol{G}^*_{ps} + \boldsymbol{G}^*_{pa}$, where $\boldsymbol{G}^*_{pa} \in \mathcal{A}_{\text{loc}}$ and $\boldsymbol{G}^*_{ps} \in \mathcal{S}_{\text{loc}}$. Since $\boldsymbol{G}^*_{pa} \in \mathcal{A}_{\text{loc}}$, there is a sequence $\{\boldsymbol{G}_{pa,n}\}$ in $\mathcal{A}_{\text{loc}}$ with $E\dot{\boldsymbol{G}}_{pa,n} = \boldsymbol{0}$ $(n = 1, 2, \ldots)$, such that $E((\boldsymbol{G}^*_{pa} - \boldsymbol{G}_{pa,n})'(\boldsymbol{G}^*_{pa} - \boldsymbol{G}_{pa,n})) \to 0$, as $n \to \infty$.

Let $\boldsymbol{G}_n = \boldsymbol{G}^*_{ps} + \boldsymbol{G}_{pa,n}$. By the assumption of the theorem, we have $\boldsymbol{G}_n \in \mathcal{G}_1$. So by Theorem 2.1,

$$\left(E\dot{\boldsymbol{G}}_n\right)^{-1}\left(E\boldsymbol{G}_n\boldsymbol{G}^{*'}\right) = \left(E\dot{\boldsymbol{G}}^*\right)^{-1}\left(E\boldsymbol{G}^*\boldsymbol{G}^{*'}\right),$$

i.e.,

$$\begin{aligned}\left(E\dot{\boldsymbol{G}}^*_{ps}\right)^{-1}&\left(E\boldsymbol{G}^*_{ps}\boldsymbol{G}^{*'}_{ps} + E\boldsymbol{G}_{pa,n}\boldsymbol{G}^{*'}_{pa}\right)\\ &= \left(E\dot{\boldsymbol{G}}^*_{ps} + E\dot{\boldsymbol{G}}^*_{pa}\right)^{-1}\left(E\boldsymbol{G}^*_{ps}\boldsymbol{G}^{*'}_{ps} + E\boldsymbol{G}^*_{pa}\boldsymbol{G}^{*'}_{pa}\right). \quad (3.4)\end{aligned}$$

Since

$$\begin{aligned}\left\|E\left(\boldsymbol{G}_{pa,n}\boldsymbol{G}^{*'}_{pa} - \boldsymbol{G}^*_{pa}\boldsymbol{G}^{*'}_{pa}\right)\right\| &= \left\|E\left(\left(\boldsymbol{G}_{pa,n} - \boldsymbol{G}^*_{pa}\right)\boldsymbol{G}^{*'}_{pa}\right)\right\|\\ &\le \left\|E\left(\boldsymbol{G}_{pa,n} - \boldsymbol{G}^*_{pa}\right)\right\|\,\left\|E\left(\boldsymbol{G}^*_{pa}\right)\right\|\\ &\qquad\qquad \to 0 \quad \text{as } n \to \infty,\end{aligned}$$

letting $n \to \infty$ in (3.4) gives

$$\begin{aligned}\left(E\dot{\boldsymbol{G}}^*_{ps}\right)^{-1}&\left(E\boldsymbol{G}^*_{ps}\boldsymbol{G}^{*'}_{ps} + E\boldsymbol{G}^*_{pa}\boldsymbol{G}^{*'}_{pa}\right)\\ &= \left(E\dot{\boldsymbol{G}}^*_{ps} + E\dot{\boldsymbol{G}}^*_{pa}\right)^{-1}\left(E\boldsymbol{G}^*_{ps}\boldsymbol{G}^{*'}_{ps} + E\boldsymbol{G}^*_{pa}\boldsymbol{G}^{*'}_{pa}\right),\end{aligned}$$

which implies that $E\dot{\boldsymbol{G}}^*_{pa} = \boldsymbol{0}$. Applying Theorem 2.1 again, we have

$$\begin{aligned}\left(E\dot{\boldsymbol{G}}^*_{ps}\right)^{-1}\left(E\boldsymbol{G}^*_{ps}\boldsymbol{G}^{*'}\right) &= \left(E\dot{\boldsymbol{G}}^*\right)^{-1}\left(E\boldsymbol{G}^*_{ps}\boldsymbol{G}^{*'}_{ps} + E\boldsymbol{G}^*_{pa}\boldsymbol{G}^{*'}_{pa}\right)\\ &= \left(E\dot{\boldsymbol{G}}^*_{ps}\right)^{-1}\left(E\boldsymbol{G}^*_{ps}\boldsymbol{G}^{*'}_{ps} + E\boldsymbol{G}^*_{pa}\boldsymbol{G}^{*'}_{pa}\right),\end{aligned}$$

i.e.,

$$\left(E\dot{\boldsymbol{G}}^*_{ps}\right)^{-1}\left(E\boldsymbol{G}^*_{ps}\boldsymbol{G}^{*'}_{ps}\right) = \left(E\dot{\boldsymbol{G}}^*_{ps}\right)^{-1}\left(E\boldsymbol{G}^*_{ps}\boldsymbol{G}^{*'}_{ps} + E\boldsymbol{G}^*_{pa}\boldsymbol{G}^{*'}_{pa}\right).$$

Therefore $E\boldsymbol{G}^*_{pa}\boldsymbol{G}^{*'}_{pa} = \boldsymbol{0}$ and hence $\boldsymbol{G}^*_{pa} = \boldsymbol{0}$. Thus,

$$\boldsymbol{G}^* \in \mathcal{G}_1 \cap \mathcal{S}_{\text{loc}}.$$

Theorem 3.3 Suppose that $\mathcal{G}_1$ is a convex set. Let $\boldsymbol{G}^* = \boldsymbol{G}^*_{ps} + \boldsymbol{G}^*_{pa}$, where $\boldsymbol{G}^*_{ps} \in \mathcal{S}_{\text{loc}} \cap \mathcal{G}_1$ and $\boldsymbol{G}^*_{pa} \in \mathcal{A}_{\text{loc}}$. If $E_\theta\dot{\boldsymbol{G}}^* = E_\theta\dot{\boldsymbol{G}}^*_{ps}$ for all $\boldsymbol{\theta}$, then $\boldsymbol{G}^* \in \mathcal{S}_{\text{loc}}$.

The proof of Theorem 3.3 is analogous to that of Theorem 3.2 and is omitted.

Before another important property for the quasi-score estimating function is mentioned, we introduce some new notation and definitions.

Let $\tilde{\mathcal{G}}_1 = \{\boldsymbol{G},\ E\dot{\boldsymbol{G}}$ is nonsingular and $\boldsymbol{G} \in \Psi\}$, so that $\tilde{\mathcal{G}}_1 \supseteq \mathcal{G}_1$.

Definition 3.5 An estimating function $\boldsymbol{G}^*_T$ is a sub-quasi-score estimating function in $\tilde{\mathcal{G}}_1$, if

$$\left(E\dot{\boldsymbol{G}}_T\right)^{-1}\left(E\boldsymbol{G}_T\boldsymbol{G}'_T\right)\left(E\dot{\boldsymbol{G}}'_T\right)^{-1} - \left(E\dot{\boldsymbol{G}}^*_T\right)^{-1}\left(E\boldsymbol{G}^*_T\boldsymbol{G}^{*'}_T\right)\left(E\dot{\boldsymbol{G}}^{*'}_T\right)^{-1}$$

is nonnegative-definite for all $\boldsymbol{G}_T \in \tilde{\mathcal{G}}_1$ and all $\boldsymbol{\theta} \in \Theta$.

Remark. The relationship between a sub-quasi-score estimating function and a quasi-score estimating function is as follows. If a sub-quasi-score estimating function $\boldsymbol{G}^*_T$ is an element of $\mathcal{G}_1$, then $\boldsymbol{G}^*_T$ is a quasi-score estimating function in $\mathcal{G}_1$. However, if $\boldsymbol{G}^*_T$ is a quasi-score estimating function in $\mathcal{G}_1$, $\boldsymbol{G}^*_T$ need not to be a sub-quasi-score estimating function in $\tilde{\mathcal{G}}_1$ because the space $\tilde{\mathcal{G}}_1$ is larger than $\mathcal{G}_1$

Obviously a sub-quasi-score estimating function is not a quasi-score estimating function in general. However, if we restrict our consideration to the model

$$X_t = \int_0^t f_s\, d\lambda_s + m_s,$$

and consider the estimating function space $\Psi_T = \{\int_0^T \alpha_s(\theta)\, dm_s,\ \alpha_s(\theta) \in \mathcal{F}_{s-}\}$ with $\mathcal{G}_1 \subseteq \tilde{\mathcal{G}}_1 \subseteq \Psi_T$, where $\{m_s\}$ is a martingale with respect to some σ-fields $\{\mathcal{F}_t\}$ and $\{\lambda_s\}$ is an increasing function, then following the proof in Godambe and Heyde (1987, p. 238), we can show that the quasi-score estimating function $\boldsymbol{G}^*$ is a sub-quasi-score estimating function in $\tilde{\mathcal{G}}_1$.

Since the conditions required for the sub-quasi-score estimating function are weaker than those for the quasi-score estimating function, it should be easy to find the conditions under which the sub-quasi-score estimating function is an E-sufficient estimating function.

We now restrict Ψ by requiring that the subset $\mathcal{A}_{\text{loc}}$ of Ψ satisfy the following condition.

Condition 3.1 $\boldsymbol{G} \in \mathcal{A}_{\text{loc}}$ if and only if $E_\theta \dot{\boldsymbol{G}} = \boldsymbol{0}$ for all $\boldsymbol{\theta}$.

Theorem 3.4 Let $\boldsymbol{G}^*$ be a sub-quasi-score estimating function on $\tilde{\mathcal{G}}_1 \subseteq \Psi$. Then, under Condition 3.1, $\boldsymbol{G}^* \in \mathcal{S}_{\text{loc}}$.

Proof. Express $\boldsymbol{G}^*$ as $\boldsymbol{G}^* = \boldsymbol{G}^*_{ps} + \boldsymbol{G}^*_{pa}$, where $\boldsymbol{G}^*_{ps} \in \mathcal{S}_{\text{loc}}$ and $\boldsymbol{G}^*_{pa} \in \mathcal{A}_{\text{loc}}$. Since $\boldsymbol{G}^*_{pa} \in \mathcal{A}_{\text{loc}}$ and $\boldsymbol{G}^* \in \tilde{\mathcal{G}}_1$, it follows that $E_\theta \dot{\boldsymbol{G}}^*_{ps} = E_\theta \dot{\boldsymbol{G}}^*$ is nonsingular for all $\boldsymbol{\theta} \in \Theta$. Following Definition 3.5 we have that

$$\left(E\dot{\boldsymbol{G}}^*_{ps}\right)^{-1} \left(E\boldsymbol{G}^*_{ps}\boldsymbol{G}^{*'}_{ps}\right) \left(E\dot{\boldsymbol{G}}^{*'}_{ps}\right)^{-1} - \left(E\dot{\boldsymbol{G}}^*\right)^{-1} \left(E\boldsymbol{G}^*\boldsymbol{G}^{*'}\right) \left(E\dot{\boldsymbol{G}}^{*'}\right)$$

is nonnegative-definite. That is,

$$\left(E\dot{\boldsymbol{G}}^*_{ps}\right)^{-1} \left(E\boldsymbol{G}^*_{ps}\boldsymbol{G}^{*'}_{ps}\right) \left(E\dot{\boldsymbol{G}}^{*'}_{ps}\right)^{-1}$$
$$- \left(E\dot{\boldsymbol{G}}^*_{ps}\right)^{-1} \left(E\boldsymbol{G}^*_{ps}\boldsymbol{G}^{*'}_{ps} + E\boldsymbol{G}^*_{pa}\boldsymbol{G}^{*'}_{pa}\right) \left(E\dot{\boldsymbol{G}}^{*'}_{ps}\right)^{-1}$$

is nonnegative-definite. Consequently, $-E\boldsymbol{G}^*_{pa}\boldsymbol{G}^{*'}_{pa}$ is nonnegative-definite, which gives $E\boldsymbol{G}^*_{pa}\boldsymbol{G}^{*'}_{pa} = \boldsymbol{0}$ and $\boldsymbol{G}^* \in \mathcal{S}_{\text{loc}}$.

Theorem 3.4 shows that, if Condition 3.1 is satisfied and if a quasi-score estimating function is a sub-quasi-score estimating function, then it will belong to $\mathcal{S}_{\text{loc}}$. Therefore, another way to check whether $\boldsymbol{G}^* \in \mathcal{S}_{\text{loc}}$ is by verifying Condition 3.1 and showing that $\boldsymbol{G}^*$ is a sub-quasi-score estimating function.

As an example, consider a stochastic process $\{X_t\}$ that satisfies a stochastic diferential equation of the form

$$dX_t = a_t(\theta)\, d\langle M(\theta)\rangle_t + dM_t(\theta) \qquad (t \le T), \tag{3.5}$$

where $\{M_t\}$ is a square integrable martingale with predictable variation process $\langle M\rangle_t$ observable up to the value of the parameter θ. Also assume that the predictable process $\{a_t(\theta)\}$ is observable up to the parameter value and there exists a real-valued predictable process $f(t;\, \eta, \theta)$ such that

$$\int_0^s a_t(\eta)\, d\langle M(\eta)\rangle_t = \int_0^s f(t; \eta, \theta)\, d\langle M(\theta)\rangle_t,$$

for all $0 < s < T$ and all η sufficiently close to θ. Noting that $f(t; \theta, \theta) = a_t(\theta)$ (see McLeish and Small (1988, p. 103)) and restricting our consideration to a space of estimating functions, in which each element has the form $\psi(\theta) = \int_0^T H_s(\theta)\, dM_s(\theta)$, where H is a predictable process, we obtain the following proposition.

Proposition 3.1 Under the conditions of the model (3.5), Condition 3.1 is satisfied, i.e.,

$$\mathcal{A}_{\mathrm{loc}} = \{\phi : \ E_\theta \, \dot{\phi} = 0, \quad \text{all } \theta \in \Theta\}.$$

Proof. We note that, for any estimating function ψ in the estimating function space and parameters η and θ in the parameter space,

$$\begin{aligned} E_\eta \psi(\theta) &= E_\eta \left\{ \int_0^T H_t(\theta)[dM_t(\eta) + a_t(\eta)\, d\langle M(\eta)\rangle_t - a_t(\theta)\, d\langle M(\theta)\rangle_t] \right\} \\ &= E_\eta \left\{ \int_0^T H_t(\theta)[f(t;\eta,\theta) - f(t;\theta,\theta)]\, d\langle M(\theta)\rangle_t \right\}. \end{aligned}$$

Dividing by $\eta - \theta$ in the above equation and then letting $\eta \to \theta$, we obtain

$$E_\theta \dot{\psi}(\theta) = \frac{\partial}{\partial \eta} E_\eta \, \psi(\theta) \Big|_{\eta=\theta} = E_\theta \int_0^T H_t(\theta) \left[\frac{\partial}{\partial \theta} f(t;\theta,\theta) \right] d\langle M(\theta)\rangle_t .$$

For any $\psi_0(\theta) = \int_0^T H_{t,0}(\theta)\, dM_t(\theta)$ in $\mathcal{A}_{\mathrm{loc}}$, we can show that $E_\theta \, \dot{\psi}_0 = 0$. From the definition of $\mathcal{A}_{\mathrm{loc}}$, there is a sequence of estimating functions $\{\psi_n = \int_0^T H_{t,n}\, dM_t(\theta)\}$ with

$$E_\theta \dot{\psi}_n = E_\theta \int_0^T H_{t,n}(\theta) \left[\frac{\partial}{\partial \theta} f(t;\theta,\theta)\, d\langle M(\theta)\rangle_t \right] = 0,$$

for all θ, which weakly square converges to ψ_0. Therefore, applying the Kunita-Watanabe inequality (see e.g. Elliott (1982, Corollary 10.12, p. 102)), we obtain

$$\begin{aligned} &\left| E_\theta \dot{\psi}_0(\theta) \right| = \left| E_\theta \left(\dot{\psi}_0(\theta) - \dot{\psi}_n(\theta) \right) \right| \\ &= \left| E_\theta \int_0^T (H_{t,0}(\theta) - H_{t,n}(\theta)) \left[\frac{\partial}{\partial \theta} f(t;\theta,\theta) \right] d\langle M(\theta)\rangle_t \right| \\ &\leq \left(E_\theta \int_0^T (H_{t,0}(\theta) - H_{t,n}(\theta))^2 d\langle M(\theta)\rangle_t \right)^{\frac{1}{2}} \\ &\qquad \cdot \left(E_\theta \int_0^T \left(\frac{\partial}{\partial \theta} f(t;\theta,\theta) \right)^2 d\langle M(\theta)\rangle_t \right)^{\frac{1}{2}} \\ &= \|\psi_0 - \psi_n\|_2 \left(E_\theta \int_0^T \left(\frac{\partial}{\partial \theta} f(t;\theta,\theta) \right)^2 d\langle M(\theta)\rangle_t \right)^{\frac{1}{2}} \to 0 \quad \text{as } n \to \infty, \end{aligned}$$

which yields $E_\theta \dot{\psi}_0 = 0$ for all $\theta \in \Theta$. Therfore,

$$\mathcal{A}_{\text{loc}} = \{\psi : \; E_\theta \dot{\psi}(\theta) = 0, \quad \text{all}\, \theta, \psi \in \Psi\}.$$

Applying Proposition 3.1 and following the remark under Definition 3.5 for models satisfying (3.5), all quasi-score estimating functions from the space $\mathcal{G}_1 \in \Psi = \{\int_0^T H_s(\theta)\, dM_s(\theta); \quad H$ a predictable process$\}$ lie in $\mathcal{S}_{\text{loc}}$.

As another example consider the widely used model

$$dX_t = f_t(\theta)\, d\lambda + dM_t,$$

where $\{f_t(\theta)\}$ is a predictable process and M is a square integrable martingale with $d\langle M(\theta)\rangle_t = a_t(\theta)\, d\lambda$. Since this model can be rewritten in the form (3.5) and also satisfies the assumption required for (3.5), the quasi-score estimating function in the space

$$\Psi = \left\{ \int_0^T H_t(\theta)\, [dX_t - f_t(\theta)\, d\lambda], \quad \{H_t\} \;\text{ a predictable process} \right\}$$

is locally E-sufficient.

Condition 3.1 plays an important role in our discussion. However, checking this condition in general is still an issue. Of course Proposition 3.1 says that Condition 3.1 is always true for the model 3.5, which covers many problems and a similar analysis can be undertaken in other particular cases.

3.4 Complement and Exercise

The theory in this book has been formulated on the assumption that the estimating functions $G(\theta)$ under consideration are differentiable with respect to θ. It is possible, however, to extend this theory by replacing the $-E_\theta \dot{G}(\theta)$ used herein by

$$\nabla G(\theta) = \frac{\partial}{\partial \eta}\, E_\eta G(\theta) \Big|_{\eta=\theta}.$$

This feature has been incorporated in the theory of this chapter. The following exercise indicates the kind of result that is thereby included.

1. Estimation of the parameter θ in the density function $\frac{1}{2} e^{-|x-\theta|}$, $-\infty < x < \infty$, is required on the basis of a random sample $\{X_1, \ldots, X_T\}$. Consider the space Ψ of estimating functions that contains just the two elements

$$\psi_1 = \sum_{t=1}^{T} |X_t - \theta| - T, \qquad \psi_2 = \sum_{t=1}^{T} \text{sgn}\,(X_t - \theta).$$

Show that ψ_1 is locally E-ancillary and ψ_2 is locally E-sufficient for estimation of θ. Note that ψ_2 leads to the sample median as an estimator of θ.

Chapter 4

Asymptotic Confidence Zones of Minimum Size

4.1 Introduction

Thus far in the book we have concentrated on the optimal properties of estimating functions that are derived from maximum information, minimum distance, or, in the case of Chapter 3, sufficiency ideas, which all involve fixed finite samples. Now we shall proceed to asymptotic considerations about confidence zones that generalize classical results holding for maximum likelihood. However, some comments about maximum likelihood itself are appropriate as a prelude since, although it is one of the most widely used statistical techniques, it does suffer from various problems.

The principal justification of maximum likelihood is asymptotic. Under certain regularity conditions the maximum likelihood estimator is asymptotically unbiased, consistent, and asymptotically normally distributed with minimum size asymptotic confidence zones for the unknown parameter. Unfortunately the regularity conditions are rather stringent, and there is no shortage of striking examples of things that can go wrong, even in such seemingly innocent contexts as for exponential families. Le Cam (1990b) has given an entertaining account of problems for i.i.d. variables. However for all its faults, maximum likelihood methods do provide a benchmark against which others are judged and the exceptions just serve to underline the principle of reviewing each application individually for compliance.

Not surprisingly, whatever problems can occur for likelihood based methods can also occur for quasi-likelihood. However, the quasi-likelihood framework has one big advantage and that is the opportunity to choose the family of estimating functions within which to work. Appropriate regularity can be demanded of these estimating functions whereas there is ordinarily little control over the choice of model and consequent form of the likelihood.

Despite the foregoing comments we shall ordinarily take counterparts of the standard regularity conditions for maximum likelihood for granted in developing the corresponding quasi-likelihood theory. Exceptions typically need rather special treatment.

The reader is reminded that the standard regularity conditions for maximum likelihood involve a finite dimensional parameter space Θ, say, which is an open subset of the corresponding Euclidean space and a true parameter which is an interior point of Θ. A likelihood which is at least twice differentiable with respect to the parameter θ is postulated and differentiation under the integral sign is permitted. Various other technical conditions are typically imposed which are sufficient to ensure consistency and asymptotic normality of the estimators (e.g., Cox and Hinkley (1974, Chapter 9), Basawa and Prakasa Rao (1980, particularly Chapter 7)). The most common exceptions involve unknown endpoint problems (i.e., one endpoint of the range of the distribution in question is the unknown parameter), parameters on the boundary of the parameter space and unbounded likelihoods. For discussion of nonregular cases for the MLE see, e.g., Smith (1985), (1989), Cheng and Traylor (1995). Whenever something can be done for the MLE, it may be expected that corresponding quasi-likelihood analogues can be developed.

Suppose that we have experiments indexed by $t \in [0, \infty)$ and $\{\mathcal{F}_t\}$ denotes a standard filtration generated from these experiments. Both discrete and continuous time are covered. With little loss in generality we confine attention to the class $\mathcal{G}$ of zero mean, square integrable, $\mathcal{F}_t$-measurable, semimartingale estimating functions $\boldsymbol{G}_t(\boldsymbol{\theta})$ which are vectors of dimension p such that the p dimensional matrices

$$\dot{\boldsymbol{G}}_t(\boldsymbol{\theta}) = (\partial G_{t,i}(\boldsymbol{\theta})/\partial \theta_j),$$

$E\dot{\boldsymbol{G}}_t(\boldsymbol{\theta})$, $E\boldsymbol{G}_t(\boldsymbol{\theta})\,\boldsymbol{G}'_t(\boldsymbol{\theta})$, and $[\boldsymbol{G}(\boldsymbol{\theta})]_t$ are (a.s.) nonsingular for each $t > 0$, $[\boldsymbol{G}(\boldsymbol{\theta})]_t$ denoting the quadratic variation process.

Here

$$[\boldsymbol{G}(\boldsymbol{\theta})]_t = [\boldsymbol{G}^{cm}(\boldsymbol{\theta})]_t + \sum_{0<s\le t} (\Delta\,\boldsymbol{G}_s)(\Delta\,\boldsymbol{G}_s)',$$

where $\boldsymbol{G}^{cm}$ is the unique continuous martingale part of $\boldsymbol{G}$ and

$$\Delta\,\boldsymbol{G}_s = \boldsymbol{G}_s(\boldsymbol{\theta}) - \boldsymbol{G}_{s-}(\boldsymbol{\theta});$$

see, e.g., Rogers and Williams (1987, p. 391) for a discussion.

An effective choice of $\boldsymbol{G}_t$ is vital and in the case where the likelihood function exists, is known, and is tractable, the score function provides the benchmark and should ordinarily be used for $\boldsymbol{G}_t$. It is well known, for example, that maximum likelihood (ML) estimation is associated with minimum-size asymptotic confidence zones under suitable regularity conditions (e.g., Hall and Heyde (1980, Chapter 6)). We shall, however, show that for quasi-likelihood properties similar to that of the ML estimator hold within a more restricted setting.

4.2 The Formulation

Asymptotic confidence statements about the "parameter" $\boldsymbol{\theta}$ are based on consideration of the asymptotic properties of a suitably chosen estimating function $\boldsymbol{G}_t(\boldsymbol{\theta})$ as $t \to \infty$. For this purpose we shall assume, as it typically the case in regular problems of genuine physical relevance, that $\boldsymbol{G}_t(\boldsymbol{\theta})$ has a limit distribution under some approximate normalization.

Indeed, for $\boldsymbol{G}_t(\boldsymbol{\theta}) \in \mathcal{G}$, the result

$$(E\boldsymbol{G}_t\boldsymbol{G}'_t)^{-1/2}\,\boldsymbol{G}_t \xrightarrow{\text{d}} \boldsymbol{X} \tag{4.1}$$

for some proper law $\boldsymbol{X}$, which is not in general normal, but does not depend on the choice of $\boldsymbol{G}_t$, seems to be uniquitous.

As a basis for obtaining confidence zones for $\boldsymbol{\theta}$ we suppose that data are available for $0 \le t \le T$ and, letting $\boldsymbol{\theta}_T^*$ be a solution of $\boldsymbol{G}_T(\boldsymbol{\theta}) = \mathbf{0}$, we use Taylor expansion to obtain

$$\mathbf{0} = \boldsymbol{G}_T(\boldsymbol{\theta}_T^*) = \boldsymbol{G}_T(\boldsymbol{\theta}) + \dot{\boldsymbol{G}}_T(\boldsymbol{\theta}_{1,T})\,(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}), \tag{4.2}$$

where $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{1,T}\| \le \|\boldsymbol{\theta} - \boldsymbol{\theta}_T^*\|$, the norm denoting sum of squares of elements. Then, assuming that

$$(E\boldsymbol{G}_T(\boldsymbol{\theta})\,\boldsymbol{G}_T'(\boldsymbol{\theta}))^{-1/2}\,\dot{\boldsymbol{G}}_T(\boldsymbol{\theta}_{1,T})\left(E\dot{\boldsymbol{G}}_T(\boldsymbol{\theta})\right)^{-1}\left(E\boldsymbol{G}_T(\boldsymbol{\theta})\,\boldsymbol{G}_T'(\boldsymbol{\theta})\right)^{1/2} \xrightarrow{\text{p}} Y\boldsymbol{I}_p$$

for some random variable Y $(>0$ a.s.$)$ we have, as $T \to \infty$,

$$(E\boldsymbol{G}_T(\boldsymbol{\theta})\,\boldsymbol{G}_T'(\boldsymbol{\theta}))^{-1/2}\left(E\dot{\boldsymbol{G}}_T(\boldsymbol{\theta})\right)(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}) \xrightarrow{\text{d}} \boldsymbol{Z}, \tag{4.3}$$

say, not depending on the choice of $\boldsymbol{G}_T$.

The size of confidence zones for $\boldsymbol{\theta}$ is then governed by the scaling "information"

$$\begin{aligned}\mathcal{E}(\boldsymbol{G}_t(\boldsymbol{\theta})) &= \left(E\dot{\boldsymbol{G}}_t(\boldsymbol{\theta})\right)'\left(E\boldsymbol{G}_t(\boldsymbol{\theta})\,\boldsymbol{G}_t'(\boldsymbol{\theta})\right)^{-1} E\left(\dot{\boldsymbol{G}}_t(\boldsymbol{\theta})\right) \\ &= E\boldsymbol{G}_t^{(s)}(\boldsymbol{\theta})\,\boldsymbol{G}_t^{(s)'}(\boldsymbol{\theta})\end{aligned}$$

and we prefer estimating function $\boldsymbol{G}_{1,t}$ to $\boldsymbol{G}_{2,t}$ if $\mathcal{E}(\boldsymbol{G}_{1,t}) \ge \mathcal{E}(\boldsymbol{G}_{2,t})$ for each $t \ge t_0$ in the Loewner ordering (partial order of nonnegative definite matrices). This is precisely the requirement of O_F-optimality, as discussed in Chapter 2, but for all sufficiently large samples.

We shall henceforth suppose that a quasi-score or asymptotic quasi-score estimating function $\boldsymbol{Q}_t(\boldsymbol{\theta})$ has been chosen for which (4.3) holds and our considerations regarding confidence zones will be based on $\boldsymbol{Q}_t$.

Meta Theorem: *Under sensible regularity conditions a quasi-likelihood estimator within some $\mathcal{H}$ is strongly consistent and, with suitable norming, asymptotically normally distributed. It can be used to construct minimum size asymptotic confidence zones for estimators within $\mathcal{H}$.*

This is a result for which a satisfactory formal statement and proof are elusive to the extent that any readily formulated set of sufficient conditions has obvious exceptions for which the desired results continue to hold. Furthermore, it is usually preferable, and more economical, to check directly in a particular case to see whether consistency and asymptotic normality hold than to try to verify the sufficient conditions of a general theorem that might not quite apply. The results in this area mostly make use of martingale strong laws and central limit theorems and we shall provide powerful versions of these in

Chapter 12. They can be used in particular cases. For specific results along the lines of the meta theorem we refer the reader to Hutton and Nelson (1986), Theorem 3.1 (strong consistency) and Theorem 4.1 (asymptotic normality), Hutton, Ogunyemi and Nelson (1991), Theorem 12.1 (strong consistency) and Theorem 12.2 (asymptotic normality) and Greenwood and Wefelmeyer (1991), Proposition 10.1. See also Section 3 of Barndorff-Nielsen and Sørensen (1994).

Of course the issue of "sensible regularity conditions" is crucial. For some examples where asymptotic normality of the quasi-likelihood estimators does not hold, see Chan and Wei (1988). These involve unstable autoregressive processes.

Much of the complication of regularity conditions can be drawn off into a single stochastic equicontinuity condition (Pollard (1984, Chapter 7)) in the case where a stochastic maximization or minimization is involved, or similarly, into stochastic differentiability conditions (Hoffmann-Jørgensen (1994, Chapter 14)).

4.3 Confidence Zones: Theory

At the outset it is important to emphasize that confidence intervals based on (4.3) are mostly difficult to formulate unless $\boldsymbol{Z}$ is normally distributed and it is desirable, wherever possible, to renormalize to obtain asymptotic normality. Indeed, as we shall see in Section 4.4, there is a specific sense in which confidence intervals based on asymptotic normality are preferable, on average, to those based on an alternative limiting distribution, at least in the scalar case.

For $\boldsymbol{Q}_t(\boldsymbol{\theta}) \in \mathcal{G}$, the result

$$[\boldsymbol{Q}]_t^{-1/2}\, \boldsymbol{Q}_t \xrightarrow{\mathrm{d}} MVN(0, \boldsymbol{I}_p) \tag{4.4}$$

seems to encapsulate the most general form of asymptotic normality result. Norming by a random process such as $[\boldsymbol{Q}]_t$ is essential in what is termed the nonergodic case (for which $(E[\boldsymbol{Q}]_t)^{-1}\, [\boldsymbol{Q}]_t \overset{\mathrm{p}}{\not\longrightarrow}$ constant as $t \to \infty$). A simple example of this is furnished by the pure birth process N_t with intensity $\theta\, N_{t-}$, where we take for Q_t the score function and

$$\begin{aligned} Q_t &= \theta^{-1}\,(N_t - 1) - \int_0^t N_{s-}\, ds, \\ [Q]_t &= \theta^{-2}\,(N_t - 1), \end{aligned}$$

while

$$[Q]_t^{-1/2}\, Q_t \xrightarrow{\mathrm{d}} N(0,1), \quad (EQ_t^2)^{-1/2}\, Q_t \xrightarrow{\mathrm{d}} W^{1/2}\, N(0,1)$$

and

$$(E[Q]_t)^{-1}\, [Q]_t \xrightarrow{\mathrm{a.s.}} W,$$

where W has a gamma distribution with the form parameter N_0 and shape parameter N_0 and $W^{1/2}\, N(0,1)$ is distributed as the product of independent

$W^{1/2}$ and $N(0,1)$ variables. Here

$$\mathcal{E}(Q_t(\theta)) = E[Q]_t = EQ_t^2,$$

the MLE $\hat{\theta}_t$ satisfies

$$\hat{\theta}_t = (N_t - 1) \Big/ \int_0^t N_{s-}\, ds$$

and

$$\frac{1}{\theta}\, [Q]_t^{\frac{1}{2}}\, (\hat{\theta}_t - \theta) \quad \xrightarrow{\text{d}} \quad N(0,1),$$

$$\frac{1}{\theta}\, \left(\mathcal{E}(Q_t(\theta))^{\frac{1}{2}}\, (\hat{\theta}_t - \theta)\right) \quad \xrightarrow{\text{d}} \quad W^{-\frac{1}{2}}\, N(0,1).$$

On average, confidence intervals for θ from the former are shorter than those from the latter (Section 4.4).

To obtain confidence zones for θ from (4.4) we may use the Taylor expansion (4.2) and then under appropriate continuity conditions for $\dot{\boldsymbol{Q}}_t$,

$$\dot{\boldsymbol{Q}}_T(\boldsymbol{\theta}_{1,T})\, \left(\dot{\boldsymbol{Q}}_T(\boldsymbol{\theta})\right)^{-1} \xrightarrow{\text{p}} \boldsymbol{I}_p$$

and, when (4.4) holds

$$[\boldsymbol{Q}(\boldsymbol{\theta})]_T^{-1/2}\, \dot{\boldsymbol{Q}}_T(\boldsymbol{\theta})\, (\boldsymbol{\theta}_T^* - \boldsymbol{\theta}) \xrightarrow{\text{d}} MVN(0, \boldsymbol{I}_p) \tag{4.5}$$

as $T \to \infty$.

For the construction of confidence intervals we actually need this convergence to be uniform in compact intervals of θ; we shall use $\xrightarrow{\text{u.d.}}$ to denote such convergence and henceforth suppose that (4.5) holds in this mode. We shall also write

$$\bar{\mathcal{E}}(\boldsymbol{Q}_t(\boldsymbol{\theta})) = \left(-\dot{\boldsymbol{Q}}_t(\boldsymbol{\theta})\right)'\, [\boldsymbol{Q}(\boldsymbol{\theta})]_t^{-1}\, \left(-\dot{\boldsymbol{Q}}_t(\boldsymbol{\theta})\right). \tag{4.6}$$

If $\boldsymbol{C}$ is a column vector of dimension p, convergence in (4.5) is mixing in the sense of Rényi (see Hall and Heyde (1980, p. 64)) and $\bar{\mathcal{E}}(\boldsymbol{Q}_T(\boldsymbol{\theta}))$ behaves asymptotically like a constant matrix. Then, replacing $\bar{\mathcal{E}}(\boldsymbol{Q}_T(\boldsymbol{\theta}))$ by the estimated $\bar{\mathcal{E}}(\boldsymbol{Q}_T(\boldsymbol{\theta}_T^*))$, we obtain

$$P\left(\boldsymbol{C}'\boldsymbol{\theta} \in \boldsymbol{C}'\boldsymbol{\theta}_T^* \pm z_{\alpha/2}\, \boldsymbol{C}'\, \left(\bar{\mathcal{E}}(\boldsymbol{Q}_T(\boldsymbol{\theta}_T^*))\right)^{-1}\, \boldsymbol{C}\right) \approx 1 - \alpha$$

where $\Phi(z_\beta) = 1 - \beta$, Φ denoting the standard normal distribution function. In particular, this provides asymptotic confidence results for the individual elements of $\boldsymbol{\theta}$ and also any nuisance parameters can conveniently be deleted.

Other confidence statements of broader scope may also be derived. Noting that

$$(\boldsymbol{\theta}_T^* - \boldsymbol{\theta})'\bar{\mathcal{E}}\, (\boldsymbol{Q}_T(\boldsymbol{\theta}_T^*))\, (\boldsymbol{\theta}_T^* - \boldsymbol{\theta}) \xrightarrow{\text{u.d.}} \chi_p^2$$

and that

$$(\boldsymbol{\theta}_T^* - \boldsymbol{\theta})'\bar{\mathcal{E}}\,(\boldsymbol{Q}_T(\boldsymbol{\theta}_T^*))\,(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}) = \max_{\boldsymbol{C}} \frac{(\boldsymbol{C}'(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}))^2}{\boldsymbol{C}'(\bar{\mathcal{E}}\,(\boldsymbol{Q}_T(\boldsymbol{\theta}_T^*)))^{-1}\boldsymbol{C}},$$

we have that

$$P\left(\max_{\boldsymbol{C}} \frac{|\boldsymbol{C}'(\boldsymbol{\theta}_T^* - \boldsymbol{\theta})|}{(\boldsymbol{C}'(\bar{\mathcal{E}}\,(\boldsymbol{Q}_T(\boldsymbol{\theta}_T^*)))^{-1}\boldsymbol{C})^{1/2}} \le \chi_{p,\alpha}\right) \approx 1 - \alpha,$$

where $\chi^2_{p,\alpha}$ is the upper α point of the chi-squared distribution with p degrees of freedom, so that

$$P\left(\boldsymbol{C}'\,\boldsymbol{\theta}_0 \in \boldsymbol{C}'\,\boldsymbol{\theta}_T^* \pm \chi_{p,\alpha}\left(\boldsymbol{C}'(\bar{\mathcal{E}}\,(\boldsymbol{Q}_T(\boldsymbol{\theta}_T^*)))^{-1}\boldsymbol{C}\right)^{1/2} \quad \text{for all}\,\boldsymbol{C}\right) \approx 1 - \alpha.$$

Also, if we let $\boldsymbol{c}_\alpha$ be the set of all possible $\boldsymbol{\theta}$ satisfying the inequality

$$(\boldsymbol{\theta}_T^* - \boldsymbol{\theta})'\bar{\mathcal{E}}\,(\boldsymbol{Q}_T(\boldsymbol{\theta}_T^*))\,(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}) \le \chi^2_{p,\alpha},$$

the so-called ellipsoids of Wald in the classical case (e.g., Le Cam (1990a)), then simultaneous confidence intervals for general functions $g_i(\boldsymbol{\theta})$, $i = 1, 2, \ldots, p$ with asymptotic confidence possibly greater than $(1 - \alpha)$ are given by

$$\left\{\min_{\boldsymbol{\theta}\in c_\alpha} g_i(\boldsymbol{\theta}), \max_{\boldsymbol{\theta}\in c_\alpha} g_i(\boldsymbol{\theta})\right\}, \qquad i = 1, 2, \ldots, p.$$

This allows us to deal conveniently with confidence intervals for nonlinear functions of the components $\boldsymbol{\theta}$.

Other methods are available in particular cases and, it often happens, in cases where likelihood-based theory is available and tractable, that the ML ratio produces better-behaved confidence zones than the ML estimator (e.g. Cox and Hinkley (1974, p. 343)). In such a case, if $\boldsymbol{\theta} = (\boldsymbol{\psi}', \boldsymbol{\lambda}')'$ and $\Omega_0 = \{\boldsymbol{\psi}_0, \boldsymbol{\lambda} \in \Omega_\lambda\}$, then the set

$$\left\{\Omega_0 : \sup_{\Omega} L_t(\boldsymbol{\theta}) - \sup_{\Omega_0} L_T(\boldsymbol{\theta}) \le \frac{1}{2}\,\chi^2_{d,\alpha}\right\} \tag{4.7}$$

based on the likelihood function L, and with $d = \dim \boldsymbol{\psi}$, may give an asymptotic confidence region for $\boldsymbol{\psi}$ of size α. It should be noted that (4.7) is invariant under transformation of the parameter of interest.

It is usually the case that the quasi-score or asymptotic quasi-score estimating function $\boldsymbol{Q}_t(\boldsymbol{\theta})$ is a martingale and subject to suitable scaling,

$$-\,\dot{\boldsymbol{Q}}_t(\boldsymbol{\theta}) - [\boldsymbol{Q}(\boldsymbol{\theta})]_t, \tag{4.8}$$

$$-\,\dot{\boldsymbol{Q}}_t(\boldsymbol{\theta}) - \langle \boldsymbol{Q}(\boldsymbol{\theta})\rangle_t, \tag{4.9}$$

and

$$[\boldsymbol{Q}(\boldsymbol{\theta})]_t - \langle \boldsymbol{Q}(\boldsymbol{\theta})\rangle_t \tag{4.10}$$

are zero mean martingales, $\langle \boldsymbol{Q}(\boldsymbol{\theta})\rangle_t$ being the quadratic characteristic of $\boldsymbol{Q}_t$. The quantity $-\dot{\boldsymbol{Q}}_t(\boldsymbol{\theta})$ is a generalized version of what is known as the observed information, while

$$E\boldsymbol{Q}_t(\boldsymbol{\theta})\,\boldsymbol{Q}_t'(\boldsymbol{\theta}) = E[\boldsymbol{Q}(\boldsymbol{\theta})]_t = E\langle \boldsymbol{Q}(\boldsymbol{\theta})\rangle_t = -E\dot{\boldsymbol{Q}}_t(\boldsymbol{\theta})$$

is a generalized Fisher information. The martingale relationships (4.8) – (4.10) and that fact that each of the quantities $-\dot{\boldsymbol{Q}}_t(\boldsymbol{\theta})$, $[\boldsymbol{Q}(\boldsymbol{\theta})]_t$ and $\langle \boldsymbol{Q}(\boldsymbol{\theta})\rangle_t$ goes a.s. to infinity as t increases usually implies their asymptotic equivalence.

Which of the asymptotically equivalent forms should be used for asymptotic confidence zones based on (4.5) is unclear in general despite the fact that many special investigations have been conducted. Let $\mathcal{S}$ be the set of (possibly random) normalizing sequences $\{\boldsymbol{A}_t\}$ that are positive definite and such that

$$\boldsymbol{A}_t^{-1}(\boldsymbol{\theta})\,[\boldsymbol{Q}(\boldsymbol{\theta})]_t^{-1/2}\,\dot{\boldsymbol{Q}}_t(\boldsymbol{\theta}) \xrightarrow{\text{p}} \boldsymbol{I}_p, \tag{4.11}$$

as $t \to \infty$. Then, each element of $\mathcal{S}$ determines an asymptotic confidence zone for $\boldsymbol{\theta}$, $\bar{\mathcal{E}}(\boldsymbol{Q}_t(\boldsymbol{\theta}))$ being replaced by $\boldsymbol{A}_t'(\boldsymbol{\theta})\,\boldsymbol{A}_t(\boldsymbol{\theta})$. Of course (4.11) ensures that zones will be similar for large T with high probability.

In the ergodic case when likelihoods rather than quasi-likelihoods are being considered, there is evidence to support the use of the observed information $-\dot{\boldsymbol{Q}}_t(\boldsymbol{\theta})$ rather than the expected information $E\boldsymbol{Q}_t(\boldsymbol{\theta})\,\boldsymbol{Q}_t'(\boldsymbol{\theta})$ (e.g., Efron and Hinkley (1978)).

In the general case, Barndorff-Nielsen and Sørensen (1994) have used a number of examples to argue the advantages of (4.6), which, as we have seen, comes naturally from (4.4). On the other hand, in Chapter 2.5 we have used the form

$$\left(-\bar{\boldsymbol{Q}}_t(\boldsymbol{\theta})\right)'\,\langle \boldsymbol{Q}(\boldsymbol{\theta})\rangle_t^{-1}\left(-\bar{\boldsymbol{Q}}_t(\boldsymbol{\theta})\right),$$

where $\bar{\boldsymbol{Q}}_t(\boldsymbol{\theta})$ is a matrix of predictable processes such that $\dot{\boldsymbol{Q}}_t(\boldsymbol{\theta}) - \bar{\boldsymbol{Q}}_t(\boldsymbol{\theta})$ is a martingale. This has the advantage of a concrete interpretation as an empirical information and is a direct extension of the classical Fisher information and is closely related to $\mathcal{E}(\boldsymbol{Q}_t(\boldsymbol{\theta}))$.

The general situation is further complicated by the fact that $\bar{\mathcal{E}}(\boldsymbol{Q}_t(\boldsymbol{\theta}))$ ordinarily involves the unknown $\boldsymbol{\theta}$ and consequently has to be replaced by $\bar{\mathcal{E}}(\boldsymbol{Q}_t(\boldsymbol{\theta}_t^*))$ in confidence statements. The variety of possibilities is such that each case should be examined on an individual basis.

Finally, some special comments on the nonergodic case are appropriate. Suppose that

$$(E[\boldsymbol{Q}]_t)^{-1}[\boldsymbol{Q}]_t \xrightarrow{\text{p}} \boldsymbol{W}(>0 \text{ a.s.})$$

as $t \to \infty$. As always, a sequence of normalizing matrices $\{\boldsymbol{A}_t\}$ is sought so that

$$\boldsymbol{A}_T(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}) \xrightarrow{\text{d}} MVN(0, \boldsymbol{I}_p)$$

as $T \to \infty$. However, it has been argued that one should condition on the limit random variable $\boldsymbol{W}$ and then treat the unobserved value $\boldsymbol{w}$ of $\boldsymbol{W}$ as a nuisance parameter to be estimated. This approach has the attraction of reducing a nonergodic model to an ergodic one but at the price of introducing asymptotics which, although plausible, may be very difficult to formalize in practice (e.g., Basawa and Brockwell (1984)). A related approach based on conditioning on an asymptotic ancillary statistic has been discussed by Sweeting (1986). This also poses considerable difficulties in formalization but may be useful in some cases. The problem of precise choice of normalization, of course, usually remains.

In conclusion, we also need to comment on the point estimate itself. If confidence statements are based on (4.5) and $\bar{\boldsymbol{\theta}}_T$ is an estimator such that

$$[\boldsymbol{Q}(\boldsymbol{\theta})]_T^{-1/2}\,\dot{\boldsymbol{Q}}_T(\boldsymbol{\theta})(\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*) \xrightarrow{\text{p}} \mathbf{0}$$

as $T \to \infty$, then

$$[\boldsymbol{Q}(\boldsymbol{\theta})]_T^{-1/2}\,\dot{\boldsymbol{Q}}_T(\boldsymbol{\theta})(\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}) \xrightarrow{\text{d}} MVN(0, \boldsymbol{I}_p)$$

clearly offers an alternative to (4.5). There may sometimes be particular reasons to favor a certain choice of estimator, such as for the avoidance of a nuisance parameter.

4.4 Confidence Zones: Practice

In the foregoing theoretical discussion we have chosen to focus on the asymptotics of the estimator $\boldsymbol{\theta}^*$ rather than the quasi-score estimating function $\boldsymbol{Q}(\boldsymbol{\theta})$ from which it is derived. However, if $\boldsymbol{Q}$ is asymptotically normal or mixed normal itself, there is significant empirical evidence in favor of basing confidence statements on $\boldsymbol{Q}$ directly.

Suppose that

$$(E\boldsymbol{Q}_T(\boldsymbol{\theta})\,\boldsymbol{Q}_T'(\boldsymbol{\theta}))^{-\frac{1}{2}}\,\boldsymbol{Q}_T(\boldsymbol{\theta}) \xrightarrow{\text{d}} MVN(\mathbf{0}, \boldsymbol{I}_p)$$

$$\left(\text{resp.}\quad \langle \boldsymbol{Q}(\boldsymbol{\theta})\rangle_T^{-\frac{1}{2}}\,\boldsymbol{Q}_T(\boldsymbol{\theta}) \xrightarrow{\text{d}} MVN(\mathbf{0}, \boldsymbol{I}_p)\right).$$

Then, we have the confidence regions

$$\left\{\boldsymbol{\theta} :\ \boldsymbol{Q}_T'(\boldsymbol{\theta})\left(E\boldsymbol{Q}_T(\boldsymbol{\theta})\,\boldsymbol{Q}_T'(\boldsymbol{\theta})\right)^{-1}\boldsymbol{Q}_T(\boldsymbol{\theta}) \le \chi^2_{p,\alpha}\right\} \tag{4.12}$$

$$\left(\text{resp.}\quad \left\{\boldsymbol{\theta} :\ \boldsymbol{Q}_T'(\boldsymbol{\theta})\langle \boldsymbol{Q}(\boldsymbol{\theta})\rangle_T^{-1}\,\boldsymbol{Q}_T(\boldsymbol{\theta}) \le \chi^2_{p,\alpha}\right\}\right).$$

The form of such regions will depend heavily on the context but it can be argued that they will usually be preferable to regions constructed on the basis of the asymptotic distribution of the estimator $\boldsymbol{\theta}^*$. It may be expected that our primary information concerns the distributions of $\boldsymbol{Q}(\boldsymbol{\theta})$ directly and that the Taylor expansion used to obtain $\boldsymbol{\theta}^*$ from $\boldsymbol{Q}(\boldsymbol{\theta})$ introduces an unnecessary component of approximation.

A simple example concerns nonlinear regression where $\boldsymbol{X}_i$ are independent random vectors with the distribution $MVN(\boldsymbol{f}_i(\boldsymbol{\theta}), \sigma^2\boldsymbol{I})$, $i = 1, 2, \ldots, T$. Then, the Hutton-Nelson quasi-score estimating function based on the martingale differences $\{\boldsymbol{m}_i(\boldsymbol{\theta}) = \boldsymbol{X}_i - \boldsymbol{f}_i(\boldsymbol{\theta})\}$ is

$$\sigma^{-2}(\dot{\boldsymbol{f}}(\boldsymbol{\theta}))'\,\boldsymbol{m}(\boldsymbol{\theta})$$

where $\dot{\boldsymbol{f}}(\boldsymbol{\theta}) = \left(\dot{\boldsymbol{f}}_1'(\boldsymbol{\theta}) \,\vdots\, \dots \,\vdots\, \dot{\boldsymbol{f}}_T'(\theta)\right)'$ with $\dot{\boldsymbol{f}}_r(\theta) = (\partial f_r(\theta_i)/\partial\theta_j)$, and $\boldsymbol{m}(\boldsymbol{\theta}) = \left(\boldsymbol{m}_1'(\boldsymbol{\theta}) \,\vdots\, \dots \,\vdots\, \boldsymbol{m}_T'(\boldsymbol{\theta})\right)'$ which has an $MVN(0, \sigma^2(\dot{\boldsymbol{f}}(\theta))'\,\dot{\boldsymbol{f}}(\theta))$ distribution. Consequently, if σ^2 is known, the confidence zone corresponding to (4.12), namely,

$$\left\{\boldsymbol{\theta} :\, \sigma^{-2}\boldsymbol{m}'(\boldsymbol{\theta})\,\dot{\boldsymbol{f}}(\boldsymbol{\theta})((\dot{\boldsymbol{f}}(\boldsymbol{\theta}))'\dot{\boldsymbol{f}}(\boldsymbol{\theta}))^{-1}(\dot{\boldsymbol{f}}(\boldsymbol{\theta}))'\boldsymbol{m}(\boldsymbol{\theta}) \le \chi^2_{p,\alpha}\right\},$$

is exact. On the other hand, the distribution of the quasi-likelihood estimator $\boldsymbol{\theta}^*$ may be seriously skewed, leading to confidence regions based on $(\boldsymbol{\theta}_T^* - \boldsymbol{\theta})'\,\mathcal{E}(\boldsymbol{Q}_T(\boldsymbol{\theta}^*))(\boldsymbol{\theta}_T^* - \boldsymbol{\theta})$, say, which are unreliable and sometimes very poor. This is despite the satisfactory asymptotics. The separate issue of possibly needing to replace σ^2 by an estimated value does not alter the general preference for confidence zones based on $\boldsymbol{Q}(\boldsymbol{\theta})$.

The problem referred to above can be alleviated by bias reduction methods and these have been widely studied. The usual approach is to seek a first order asymptotic expression for the bias

$$E\boldsymbol{\theta}_T^* = \boldsymbol{\theta} + \boldsymbol{b}_T(\boldsymbol{\theta}) + \boldsymbol{r}_T(\boldsymbol{\theta}),$$

say, where $\|\boldsymbol{r}_T(\boldsymbol{\theta})\| = o(\|\boldsymbol{b}_T(\boldsymbol{\theta})\|)$ and $\|\boldsymbol{b}_T(\boldsymbol{\theta})\| \to 0$ as $T \to \infty$ for fixed $\boldsymbol{\theta}$ and then to replace $\boldsymbol{\theta}_T^*$ by the bias corrected estimator

$$\boldsymbol{\theta}_T^{(BC)} = \boldsymbol{\theta}_T^* - \boldsymbol{b}_T(\boldsymbol{\theta}_T^*).$$

In the regular classical setting of i.i.d. random variables, $\boldsymbol{b}_T(\boldsymbol{\theta})$ is of the form $\boldsymbol{b}(\boldsymbol{\theta})/T$. For a discussion of the basic methods of bias correction see Cox and Hinkley (1974, Sections 8.4 and 9.2).

A more general approach to this problem involves modification of the score (or quasi-score) rather than the estimator itself. This has been developed by Firth (1993) for the case of maximum likelihood estimation but the ideas are readily adapted to the quasi-likelihood context. We replace the quasi-score estimating function $\boldsymbol{Q}_T(\boldsymbol{\theta})$ by

$$\boldsymbol{Q}_T^{(BC)}(\boldsymbol{\theta}) = \boldsymbol{Q}_T(\boldsymbol{\theta}) + \boldsymbol{A}_T(\boldsymbol{\theta}),$$

where $\boldsymbol{A}_T(\boldsymbol{\theta})$ is chosen to be of the form $\dot{\boldsymbol{Q}}_T(\boldsymbol{\theta})\,\boldsymbol{b}_T(\boldsymbol{\theta})$. This form is used since Taylor expansion gives

$$\boldsymbol{0} = \boldsymbol{Q}_T(\boldsymbol{\theta}^*) \simeq \boldsymbol{Q}_T(\boldsymbol{\theta}) + \dot{\boldsymbol{Q}}_T(\boldsymbol{\theta})\,(\boldsymbol{\theta}_T^* - \boldsymbol{\theta})$$

so that we have

$$\boldsymbol{Q}^{(BC)}(\boldsymbol{\theta}) \simeq -\dot{\boldsymbol{Q}}_T(\boldsymbol{\theta})\,(\boldsymbol{\theta}^* - \boldsymbol{\theta} - \boldsymbol{b}_T(\boldsymbol{\theta})),$$

which has been bias corrected to first order. The theory underlying such corrections can usefully be interpreted in differential geometry terms related to correcting for curvature (e.g., Barndorff-Nielsen and Cox (1994, Chapter 6) and references therein).

For a recent further adaption to provide higher order corrections to estimating functions see Li (1996b). The idea is to remove first order bias while minimizing the mean squared error. This allows for situations in which, for example, skewness problems preclude the use of the Edgeworth expansion and the kurtosis is unknown.

For another approach to adjusting score and quasi-score estimating functions, based on martingale methods, see Mykland (1995).

4.5 On Best Asymptotic Confidence Intervals

4.5.1 Introduction and Results

Much asymptotic inference for stochastic processes is based on use of some obvious consistent estimator but involves a choice between competing normalizations, one being of constants and the other of random variables. In this section we study the common situation where either asymptotic normality or asymptotic mixed normality is achievable through suitable normalization. It should be remarked that asymptotic mixed nomality is also obtained under a variety of non-regular conditions (e.g., Kutoyants and Vostrikova (1995)) as well as regular ones. It is shown that there is a certain sense in which, on average, confidence intervals based on the asymptotic normality result are preferable to those based on asymptotic mixed normality. The result here is from Heyde (1992a).

Theorem 4.1 Let $\{\hat{\theta}_t\}$ be a sequence of estimators that is consistent for θ and $\{c_t\}$, $\{d_t\}$ be norming sequences, possibly random, such that

$$c_t(\hat{\theta}_t - \theta) \xrightarrow{\text{d}} W, \tag{4.13}$$

$$d_t(\hat{\theta}_t - \theta) \xrightarrow{\text{d}} \eta^{-1}\, W \tag{4.14}$$

as $t \to \infty$, where W is standard normal and $\eta > 0$ a.s. is random, independent of W and such that

$$c_t\, d_t^{-1} \xrightarrow{\text{d}} \eta$$

as $t \to \infty$. Suppose that $L_t^{(i)}(\delta)$, $i = 1, 2$, is the minimum length of an $100(1-\delta)\%$ confidence interval for θ based on an exact, approximate or assumed distribution of the pivot in (4.13) and (4.14), respectively, for which the specified convergence result holds. Then

$$\liminf_{t\to\infty} E\left(L_t^{(2)}(\delta)/L_t^{(1)}(\delta)\right) \geq 1.$$

Remark 1. The theorem gives a sense in which, on average, confidence intervals based on the pivot in (4.13) are to be preferred to those based on the pivot in (4.14) whether or not a random norming is required.

Remark 2. A principal application of the theorem is in the context of inference for the nonergodic models that are common in stochastic process estimation (e.g., Basawa and Scott (1983), Hall and Heyde (1980, Chapter 6)). In this context we have $\{c_t\}$ as random and $\{d_t\}$ as constants so that $L_t^{(1)}(\delta)$ is a random variable and $L_t^{(2)}(\delta)$ is a constant.

One of the most important cases is that of locally asymptotic mixed normal (LAMN) families. This is the situation in which the log-likelihood ratio has asymptotically a mixed normal distribution and results of type (4.14) hold with $\{d_t\}$ as constants and (4.13) with $\{c_t\}$ as random variables, while $Ec_t = d_t$ for each n. For LAMN families one has, under modest regularity conditions, a striking result known as Hajék's convolution theorem (e.g., Theorem 2, page 47, of Basawa and Scott (1983)) to the effect that if $\{T_t\}$ is any other sequence of consistent estimators of θ such that

$$d_t(T_t - \theta) \xrightarrow{\mathrm{d}} U \tag{4.15}$$

for some nondegenerate U, then

$$U \overset{\mathrm{d}}{=} V + \eta^{-1}W,$$

where V is independent of η and W.

It is readily shown that confidence intervals based on the pivot in (4.15) are wider than those for the corresponding result (4.14). That is, using (4.13) is better on average than (4.14), which is better than (4.15). If Φ is the standard normal distribution function, we have for any real a, β, α with $\beta > \alpha$,

$$\Phi((\beta - a)\, y) - \Phi((\alpha - a)\, y) \leq \Phi\left(\frac{1}{2}(\beta - \alpha)\, y\right) - \Phi\left(-\frac{1}{2}(\beta - \alpha)\, y\right) \tag{4.16}$$

and hence, integrating with respect to $dP(\eta \leq y)$,

$$P(\alpha < \eta^{-1}W + a < \beta) \leq P\left(-\frac{1}{2}(\beta - \alpha) < \eta^{-1}W < \frac{1}{2}(\beta - \alpha)\right), \tag{4.17}$$

so that

$$\begin{aligned} P(\alpha < \eta^{-1}W + V < \beta) &= \int_{-\infty}^{\infty} P(\alpha < \eta^{-1}W + v < \beta)\, dP(V \leq v) \\ &\leq P\left(-\frac{1}{2}(\beta - \alpha) < \eta^{-1}W < \frac{1}{2}(\beta - \alpha)\right). \end{aligned}$$

Remark 3. A particular case of the result of the theorem has been obtained by Glynn and Iglehart (1990) who studied the problem of finding minimum size asymptotic confidence intervals for steady state parameters of the simulation output process from a single simulation run. They contrasted the approach of consistently estimating the variance constant in the relevant central limit theorem with the standardized time series approach which avoids estimation of the variance in a manner reminiscent of the t-statistic and suggested that the former approach is preferable on average.

4.5.2 Proof of Theorem 4.1

Suppose that Φ_t and Ψ_t are exact, approximate or assumed distribution functions of the pivots $c_t(\hat{\theta}_t - \theta)$ and $d_t(\hat{\theta}_t - \theta)$, respectively. We are given the complete convergence results

$$\Phi_t \xrightarrow{\text{c}} \Phi, \qquad \Psi_t \xrightarrow{\text{c}} \Psi$$

as $t \to \infty$, where

$$\Phi(x) = P(W \leq x) \quad \text{and} \quad \Psi(x) = P(\eta^{-1} W \leq x) = \int_0^\infty \Phi(xy)\, dG(y)$$

with $G(x) = P(\eta \leq x)$.

Now let $a_t,\ b_t$ be any numbers for which

$$\Phi_t(b_t) - \Phi_t(a_t) \geq 1 - \delta.$$

This gives a confidence interval for θ of length

$$l_t^{(1)}(\delta) = (b_t - a_t)/c_t.$$

By passing to a subsequence, we may suppose that $a_t \to a$ and $b_t \to b$, where $-\infty \leq a \leq b \leq \infty$. Then, by complete convergence,

$$\Phi_t(a_t) - \Phi_t(b_t) \to \Phi(b) - \Phi(a),$$

and, in view of (4.16), $b - a \geq 2z_\delta$, where $2\Phi(z_\delta) = 2 - \delta$. Therefore, noting that $c_t l_t^{(1)}(\delta)$ is not random, we have

$$\liminf_{t\to\infty} c_t\, L_t^{(1)}(\delta) \geq 2z_\delta. \tag{4.18}$$

Next, let z_t be the smallest z for which

$$\Phi_t(z) - \Phi_t(-z) \geq 1 - \delta.$$

As above, we may suppose that $z_t \to z_\infty$ and, because Φ is continuous,

$$\Phi_t(z_t) - \Phi_t(-z_t) = 1 - \delta + o(1)$$

as $t \to \infty$. By complete convergence and using the result $\Phi(-x) = 1 - \Phi(x)$, $x > 0$, it follows that $2\Phi(z_\infty) = 2 - \delta$ and hence that $z_\infty = z_\delta$. Then, $L_t^{(1)}(\delta) \leq 2z_t/c_t$, so that

$$\liminf_{t\to\infty} c_t \, L_t^{(1)}(\delta) \leq 2z_\delta, \tag{4.19}$$

and (4.18) and (4.19) imply

$$\lim_{t\to\infty} c_t \, L_t^{(1)}(\delta) = 2z_\delta. \tag{4.20}$$

Similar reasoning to that which led to (4.20) also applies to show that

$$\lim_{t\to\infty} d_t \, L_t^{(2)}(\delta) = 2\zeta_\delta, \tag{4.21}$$

where $2\Psi(\zeta_\delta) = 2 - \delta$. We merely replace (4.16) and (4.17) to show that the symmetric confidence interval is best, while $\Psi(-x) = 1 - \Psi(x)$, $x > 0$, is evident from the definition and the corresponding result for Φ.

From (4.18) and (4.21), it then follows that, as $t \to \infty$,

$$d_t \, L_t^{(2)}(\delta)/c_t \, L_t^{(1)}(\delta) \to \zeta_\delta/z_\delta.$$

Since $c_t/d_t \xrightarrow{\text{d}} \eta$, Slutsky's Theorem gives

$$L_t^{(2)}(\delta)/L_t^{(1)}(\delta) \xrightarrow{\text{d}} \eta \, \zeta_\delta/z_\delta.$$

and then, using Billingsley ((1968), Theorem 5.3, page 32),

$$\liminf_{t\to\infty} E\left(L_t^{(2)}(\delta)/L_t^{(1)}(\delta)\right) \geq (E\eta) \, \zeta_\delta/z_\delta.$$

Thus, in order to complete the proof, it remains to show that

$$(E\eta) \, \zeta_\delta \geq z_\delta. \tag{4.22}$$

Now note that $\zeta_\delta = \zeta(\eta; \delta)$ solves the equation

$$\Psi(\zeta(\eta;\delta)) = P(\eta^{-1}W \leq \zeta(\eta;\delta)) = 1 - \delta/2.$$

Then, taking $b > 0$, we have

$$1 - \frac{\delta}{2} = P(W \leq \zeta(\eta;\delta)\eta) = P\left(W \leq \frac{1}{b}\,\zeta(\eta;\delta)\, b\,\eta\right) = P(W \leq \zeta(b\,\eta;\delta)\, b\,\eta)$$

so that continuity and strict monotonicity of Ψ imply that

$$\zeta(b\,\eta;\delta) = \frac{1}{b}\,\zeta(\eta;\delta)$$

and hence, if

$$\psi(\eta) = (E\eta)\,\zeta(\eta;\delta),$$

we have

$$\psi(b\,\eta) = \psi(\eta). \tag{4.23}$$

Now we see from (4.22) that it is required to show

$$\psi(\eta) \geq \Phi^{-1}(1 - \delta/2)$$

and using (4.23) we may, without loss of generality, scale η so that

$$\zeta(\eta; \delta) = 1. \tag{4.24}$$

But (4.24) implies

$$\Psi(1) = 1 - \delta/2,$$

or equivalently,

$$\int_0^\infty \Phi(y)\, G(y) = E\Phi(\eta) = 1 - \delta/2.$$

Thus we have to show that

$$\psi(\eta) = E\eta \geq \Phi^{-1}(1 - \delta/2)$$

subject to

$$E\Phi(\eta) = 1 - \delta/2,$$

and since Φ is monotone, this holds if

$$\Phi(E\eta) \geq E\Phi(\eta). \tag{4.25}$$

However, since $1 - \Phi(x)$ is convex for $x > 0$, as is easily checked by differentiation, we have from Jensen's inequality that

$$1 - \Phi(E\eta) \leq E(1 - \Phi(\eta))$$

and (4.25) follows. This completes the proof.

Final Remarks It is worth noting that (4.22), which relates the average confidence intervals when (4.13) and (4.14) hold exactly rather than as limits, is a paraphrase of the fact that, on average, extraneous randomization does not improve confidence intervals for a normal mean. That is, if X is distributed as $N(\theta, 1)$, the usual symmetric confidence interval for θ based on $X - \theta$ is not improved, on average, if one instead uses the pivot $(x - \theta)/\eta$, where η is independent of θ. In particular, when $\bar{X}_t$, s_t^2 are the mean and variance of a random sample of size n from the $N(\theta, \sigma^2)$ distribution where σ is *known*, using the t-intervals for θ based on $(\bar{X}_n - \theta)/s_n$ produces confidence intervals which are longer on average than z-intervals.

4.6 Exercises

1. Suppose that $X_1, \ldots, X_T$ are i.i.d. random variables with mean $1/\theta$ and variance σ^2. Show that a quasi-score estimating function for the estimation of θ is $Q_T(\theta) = T(\theta^{-1} - \bar{X})$, where $\bar{X} = T^{-1} \sum_{t=1}^{T} X_t$ and that a suitable choice for a bias corrected version is $Q_T^{(BC)}(\theta) = Q_T(\theta) - \sigma^2\, \theta$. In the case where the X_t have a Poisson distribution develop alternative forms of confidence intervals using (a) bias correction and (b) the quasi-score function directly. (Adapted from Firth (1993).)

2. Let $X_1, \ldots, X_T$ be a random sample from a distribution with mean θ and variance $C\theta^2$, the coefficient of variation C being known. Taking $C = 1$ for convenience, consider the estimating function

$$G_T(\theta) = \theta^{-2} \sum_{t=1}^{T} (X_t - \theta) + \theta^{-3} \sum_{t=1}^{T} \left[(X_t - \theta)^2 - \theta^2\right],$$

which is the score function when the X_t are normally distributed. Show that a suitable choice for a bias corrected version of G_T is

$$G_T^{(BC)}(\theta) = G_T(\theta) + \frac{2}{3}\,\theta^{-1}.$$

(Adapted from Firth (1993).)

Chapter 5

Asymptotic Quasi-Likelihood

5.1 Introduction

Discussion in earlier chapters on optimality of estimating functions and quasi-likelihood has been concerned with exact results, where a specific criterion holds for either fixed T or for each T as $T \to \infty$. Here we address the situation where the criteria for optimality are not satisfied exactly but hold in a certain asymptotic sense to be made precise below. These considerations give rise to an equivalence class of asymptotic quasi-likelihood estimator, which enjoy the same kind of properties as ordinary quasi-likelihood estimators, such as having asymptotic confidence zones of minimum size, within a specified family, for the "parameter" in question.

One particular difficulty with the exact theory is that a quasi-likelihood estimator may contain an unknown parameter or parameters. The asymptotic quasi-likelihood theory herein allows one to focus on issues such as whether there is loss of information when an estimator is replaced by a consistent estimator thereof or, under some circumstances, is asymptotically irrelevant.

As a simple illustration of the ideas take the case where $\{X_t\}$ is a subcritical Galton-Watson branching process with immigration and $\boldsymbol{\theta}' = (m, \lambda)$ is to be estimated on the basis of data $\{X_t,\ t = 0, 1, \dots, T\}$, where m and λ are, respectively, the means of the offspring and immigration distributions.

This model has been widely used in practice, for example, for particle counts in colloidal solutions, and an account of various applications is given in Heyde and Seneta (1972) and Winnicki (1988).

Noting that for the model in question the $(t+1)$th generation is obtained from the independent reproduction of each of the individuals in the tth generation, each with the basic offspring distribution, plus an independent immigration input with the immigration distribution, we obtain

$$E\left(X_t \,|\, \mathcal{F}_{t-1}\right) = m\, X_{t-1} + \lambda$$

so that there is a semimartingale representation

$$\sum_{t=1}^{T} X_t = m \sum_{t=1}^{T} X_{t-1} + T\,\lambda + \sum_{t=1}^{T} m_t,$$

where

$$m_t = X_t - E\left(X_t \,|\, \mathcal{F}_{t-1}\right)$$

are martingale differences.

Now suppose that X_0 has a distribution for which $E\,X_0^2 < \infty$ and that the variances of the offspring and immigration distributions are $\sigma^2\,(<\infty)$ and $\eta^2\,(<\infty)$, respectively. Then using Theorem 2.1, the quasi-score estimating function based on the martingale $\{\sum_{s=1}^t m_s\}$ is

$$\boldsymbol{Q}_T = \left(\begin{array}{c} \sum_{t=1}^T X_{t-1}\,(\sigma^2\,X_{t-1}+\eta^2)^{-1}\,(X_t - m\,X_{t-1} - \lambda) \\ \sum_{t=1}^T (\sigma^2\,X_{t-1}+\eta^2)^{-1}\,(X_t - m\,X_{t-1} - \lambda) \end{array} \right), \tag{5.1}$$

which in general involves the nuisance parameters σ^2, η^2. If no immigration is present, so that $\lambda = \eta^2 = 0$, the nuisance parameter σ^2 disappears in the estimating equation $\boldsymbol{Q}_t = \boldsymbol{0}$. Furthermore, if we have a parametric model based on m, λ alone, such as in the case where the offspring and immigration distributions are both Poisson, there is no nuisance parameter problem. Indeed, in the Poisson case it is straightforward to check that (5.1) is a constant multiple of the true score estimating function, which leads to the maximum likelihood estimators.

In general, nuisance parameters must be estimated or avoided. In this case estimation is possible using strongly consistent estimators of σ^2 and η^2 suggested by Yanev and Tchoukova-Dantcheva (1980). We can take these estimators as

$$\hat{\sigma}^2(\bar{m}_T, \bar{\lambda}_T) = \frac{T\sum_{t=1}^T X_{t-1}\,\hat{U}_t^2 - \sum_{t=1}^T X_{t-1}\sum_{t=1}^T \hat{U}_t^2}{T\sum_{t=1}^T X_{t-1}^2 - \left(\sum_{t=1}^T X_{t-1}\right)^2},$$

$$\hat{\eta}_T^2(\bar{m}_T, \bar{\lambda}_T) = \frac{\sum_{t=1}^T \hat{U}_t^2 \sum_{t=1}^T X_{t-1}^2 - \sum_{t=1}^T X_{t-1}\sum_{t=1}^T X_{t-1}\,\hat{U}_t^2}{T\sum_{t=1}^T X_{t-1}^2 - \left(\sum_{t=1}^T X_{t-1}\right)^2}$$

where $\hat{U}_t^2 = X_t - \bar{m}_T\,X_{t-1} - \bar{\lambda}_T$, $\bar{m}_T$ and $\bar{\lambda}_T$ being strongly consistent estimators of m and λ, respectively.

Wei and Winnicki (1989) studied an estimating function closely related to (5.1) in which the term $\sigma^2\,X_{t-1} + \eta^2$ is replaced by $X_{t-1} + 1$. Details of the corresponding limit theory are given in Theorem 3.C of Winnicki (1988). Furthermore, for $\hat{m}_T$, $\hat{\lambda}_T$ based on (5.1) with estimated σ^2 and η^2 as suggested above (using $\hat{m}_T, \hat{\lambda}_T$), a similar analysis to that undertaken by Wei and Winnicki shows that

$$T^{\frac{1}{2}} \left(\begin{array}{c} \hat{m}_T - m \\ \hat{\lambda}_T - \lambda \end{array} \right) \xrightarrow{d} N(0, \boldsymbol{W}^{-1})$$

as $T \to \infty$ where, if X has the stationary limit distribution of the $\{X_t\}$ process,

$$\boldsymbol{W} = \left(\begin{array}{cc} E\,X^2\,(\sigma^2\,X+\eta^2)^{-1} & E\,X\,(\sigma^2\,X+\eta^2)^{-1} \\ E\,X\,(\sigma^2\,X+\eta^2)^{-1} & E\,(\sigma^2\,X+\eta^2)^{-1} \end{array} \right),$$

the same result as one obtains if σ^2 and η^2 are known. The quasi-likelihood framework ensures that this estimation procedure is optimal from the point of view of asymptotic efficiency.

The substitution of estimated values for the nuisance parameters amounts to replacing the quasi-likelihood estimator by an asymptotic quasi-likelihood estimator that has the same asymptotic confidence zones for the unknown parameter.

Exact solutions of the optimal estimation problem leading to ordinary quasi-likelihood estimators require certain orthogonality properties, which are often only approximately satisfied in practice. For example, using Theorem 2.1 and omitting the explicit $\boldsymbol{\theta}$ for convenience, if $\boldsymbol{Q}_T \in \mathcal{H} \subseteq \mathcal{G}$ is O_F-optimal within $\mathcal{H}$, then for the standardized estimating functions $\boldsymbol{G}_T^{(s)}$, $\boldsymbol{Q}_T^{(s)} \in \mathcal{H}$ we need to have

$$E\left(\boldsymbol{G}_T^{(s)} - \boldsymbol{Q}_T^{(s)}\right)\left(\boldsymbol{G}_T^{(s)}\right)' = E\,\boldsymbol{G}_T^{(s)}\left(\boldsymbol{G}_T^{(s)} - \boldsymbol{Q}_T^{(s)}\right)' = \boldsymbol{0}. \tag{5.2}$$

Here we shall typically be dealing with circumstances under which we have sequences of estimating functions and the quantities in (5.2) tend to zero as $T \to \infty$.

5.2 The Formulation

We confine our attention to the space $\mathcal{G}$ of sequences of zero mean and finite variance estimating functions $\{\boldsymbol{G}_T(\boldsymbol{\theta}) = \boldsymbol{G}_T(\{\boldsymbol{X}_t, 1 \le t \le T\}; \boldsymbol{\theta})\}$, which are vectors of dimension p and are *a.s.* differentiable with respect to the component of $\boldsymbol{\theta}$ and such that $E\dot{\boldsymbol{G}}_T(\boldsymbol{\theta})$ and $E\boldsymbol{G}_T(\boldsymbol{\theta})\boldsymbol{G}_T'(\boldsymbol{\theta})$ are nonsingular for each T, the prime denoting transpose. We shall write $\boldsymbol{G}_T^{(n)} = (E\dot{\boldsymbol{G}}_T)^{-1}\boldsymbol{G}_T$ for the normalized estimating function and drop the $\boldsymbol{\theta}$ for convenience. It is convenient in this context to use a different standardization of estimating functions from that of Chapters 1 and 2 that we have referred to as $\boldsymbol{G}_T^{(s)}$.

The asymptotic optimality of a sequence of estimating functions will be defined as the maximization of $\mathcal{E}(\boldsymbol{G}_T)$ in the partial order of nonnegative definite matrices in a certain asymptotic sense. Before the definition is spelt out, however, we need to discuss certain concepts concerning matrices.

For a matrix $\boldsymbol{A} = (a_{i,j})$ we shall denote by $\|\boldsymbol{A}\|$ the Frobenius (Euclidean) norm given by

$$\|\boldsymbol{A}\| = \left(\sum_i \sum_j a_{i,j}^2\right)^{\frac{1}{2}}.$$

A sequence of symmetric matrices $\{\boldsymbol{A}_n\}$ is said to be *asymptotically nonnegative definite* if there exists a sequence of matrices $\{\boldsymbol{D}_n\}$ such that $\boldsymbol{A}_n - \boldsymbol{D}_n$ is nonnegative definite and $\|\boldsymbol{D}_n\| \to 0$ as $n \to \infty$.

In the sequel we need square roots of positive definite matrices. Let $\boldsymbol{A}^{\frac{1}{2}}$ ($\boldsymbol{A}^{\frac{T}{2}}$) be a left (resp. right) square root of the positive definite matrix $\boldsymbol{A}$, i.e., $\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{A}^{\frac{T}{2}} = \boldsymbol{A}$. In addition, let $\boldsymbol{A}^{-\frac{1}{2}} = (\boldsymbol{A}^{\frac{1}{2}})^{-1}$, $\boldsymbol{A}^{-\frac{T}{2}} = (\boldsymbol{A}^{\frac{T}{2}})^{-1}$. The left Cholesky square root matrix is defined as the unique lower triangular matrix satisfying the square root condition, and it can be calculated easily without solving any eigenvalue problems. Though we do not require any specific form of the square root matrix for our results to hold, Cholesky's square root matrix is preferred for ease of application.

Definition 5.1 Suppose that $\{\boldsymbol{Q}_T\} \in \mathcal{H} \subseteq \mathcal{G}$, and

$$\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{-\frac{1}{2}} E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime} \left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{-\frac{T}{2}} - \boldsymbol{I}$$

is asymptotically nonnegative definite for all $\{\boldsymbol{G}_T\} \in \mathcal{H}$, where $\boldsymbol{I}$ is the $p \times p$ identity matrix. Then we say that $\{\boldsymbol{Q}_T\}$ is an asymptotic quasi-score (AQS) sequence of estimating functions within $\mathcal{H}$. The solution $\boldsymbol{\theta}_T^*$ of $\boldsymbol{Q}_T(\boldsymbol{\theta}) = \boldsymbol{0}$ will be called an asymptotic quasi-likelihood estimator from $\mathcal{H}$.

Remark 5.1 The choice of the subset $\mathcal{H}$ of estimating functions is open and it can be tailored to suit the particular problem. For example, features such as being linear or bounded functions of the data could be incorporated.

Remark 5.2 If $\boldsymbol{G}_T$ is O_F-optimal within $\mathcal{H}$ for each T (see Heyde (1988a)), then $\{\boldsymbol{G}_T\}$ is an asymptotic quasi-score sequence of estimating functions within $\mathcal{H}$.

Remark 5.3 The asymptotic quasi-score formulation provides a natural framework for extension of Rao's concept of asymptotic first order efficiency (e.g., Rao (1973, pp. 348-349)). In the case of a scalar parameter and estimator Z_T, Rao's definition takes the form

$$\left(EU_T^2\right)^{\frac{1}{2}} \left[Z_T - \theta - \gamma(\theta)\frac{U_T}{EU_T^2}\right] \xrightarrow{P} 0 \tag{5.3}$$

as $T \to \infty$ for some $\gamma(\theta)$ not depending on T or the observations. Here U_T is the score function and the idea is that $Z_T - \theta$ has basically the same asymptotic properties as the score function. For example, as indicated by Rao, the Fisher information in Z_T is asymptotic to the Fisher information in U_T under minor additional conditions.

In this chapter we generalize this framework in a variety of ways. First, it is possible to confine attention to estimation within a subset of estimation functions, $\mathcal{H}$ say, in which case the benchmark role of an estimating function containing maximum information is a quasi-score estimating function, Q_T, say. In addition, we can replace the linear (or linearized) estimating function by a general normalized estimating function $G_T^{(n)} = (E\dot{G}_T)^{-1}G_T$. Now, since under usual regularity conditions for quasi-score estimating functions

$$E\dot{Q}_T = -EQ_T^2,$$

we have

$$Q_T^{(n)} = -(EQ_T^2)^{-1}Q_T.$$

Thus, we can think of

$$\left(EQ_T^2\right)^{\frac{1}{2}}\left[G_T^{(n)}-\gamma(\theta)Q_T^{(n)}\right]\xrightarrow{P}0 \tag{5.4}$$

as a natural generalization of (5.3). If (5.4) holds, we could say that G_T is *asymptotically first order efficient for θ within $\mathcal{H}$.*

Now, if $\{G_T\}$ is an asymptotic quasi-score sequence of estimating functions within $\mathcal{H}$, we have

$$E\left(G_T^{(n)}\right)^2\sim E\left(Q_T^{(n)}\right)^2=\left(EQ_T^2\right)^{-1} \tag{5.5}$$

as $T\to\infty$. Also, under the standard regularity conditions

$$EG_T^{(n)}Q_T^{(n)}=-(EQ_T^2)^{-1}.$$

Thus, in this case we have

$$\left(EQ_T^2\right)\left[E\left(G_T^{(n)}\right)^2-2EG_T^{(n)}Q_T^{(n)}+E\left(Q_T^{(n)}\right)^2\right]^2\to 0$$

and (5.4) holds with $\gamma(\theta)=1$ because L_2 convergence holds. It is thus clear that the asymptotic quasi-score condition is a sensible (sufficient) condition for Rao's asymptotic first order efficiency. It has the advantages of natural extensions to deal with the quasi-likelihood setting, the case where T_n is not a linear function of θ, and also the vector case.

There are several different but equivalent versions of Definition 5.1 that illuminate the concept. These are given in Proposition 5.4. As pointed out in Chapter 2, for the ordinary quasi-score case, a definition such as that given above is of limited practical use. However, the following propositions provide a rather easy route to checking whether a sequence of estimating functions has the asymptotic quasi-score property.

Proposition 5.1 The sequence of estimating functions $\{\boldsymbol{Q}_T\}\in\mathcal{H}$ is AQS if there exists a $p\times p$ matrix $\boldsymbol{k}_T$ such that

$$\lim_{T\to\infty}\boldsymbol{k}_T E\boldsymbol{G}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\boldsymbol{k}_T'=\boldsymbol{K}=\lim_{T\to\infty}\boldsymbol{k}_T E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\boldsymbol{k}_T' \tag{5.6}$$

for all $\{\boldsymbol{G}_T\}\in\mathcal{H}$, where $\boldsymbol{K}$ is some nondegenerate $p\times p$ matrix. In particular, if

$$\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{-\frac{1}{2}}E\boldsymbol{G}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{-\frac{T}{2}}\to\boldsymbol{I} \tag{5.7}$$

as $T\to\infty$ for all $\{\boldsymbol{G}_T\}\in\mathcal{H}$, or if there exists constant $\boldsymbol{\alpha}_T$ such that

$$\lim_{T\to\infty}\boldsymbol{\alpha}_T^{-1}E\boldsymbol{G}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}=\boldsymbol{K}$$

for all $\{\boldsymbol{G}_T\}\in\mathcal{H}$ and some nondegenerate matrix $\boldsymbol{K}$, then $\{\boldsymbol{Q}_T\}$ is AQS.

Proof. Suppose that (5.6) holds for some $\{\boldsymbol{k}_T\}$. It is easy to see that $\boldsymbol{k}_T$ is nondegenerate when T is large enough. Given any $\{\boldsymbol{G}_T\} \in \mathcal{H}$, let

$$\boldsymbol{A}_T \equiv \boldsymbol{k}_T \left[E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime} - E\boldsymbol{G}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime} \left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime} \right)^{-1} \left(E\boldsymbol{G}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime} \right)^{\prime} \right] \boldsymbol{k}_T^{\prime};$$

then $\boldsymbol{A}_T$ is nonnegative definite using a standard argument based on properties of covariance matrices (e.g. Heyde and Gay (1989, p. 227)). We may write

$$\begin{aligned}
\boldsymbol{A}_T &= \boldsymbol{k}_T \left(E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime} - E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime} \right) \boldsymbol{k}_T^{\prime} \\
&\quad + \Bigg(\boldsymbol{k}_T E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\boldsymbol{k}_T^{\prime} - \left(\boldsymbol{k}_T E\boldsymbol{G}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\boldsymbol{k}_T^{\prime} \right) \left(\boldsymbol{k}_T E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\boldsymbol{k}_T^{\prime} \right)^{-1} \\
&\quad \cdot \left(\boldsymbol{k}_T E\boldsymbol{G}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\boldsymbol{k}_T^{\prime} \right)^{\prime} \Bigg) \\
&= \boldsymbol{k}_T \left(E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime} - E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime} \right) \boldsymbol{k}_T^{\prime} + \boldsymbol{D}_T,
\end{aligned}$$

say. It follows from (5.6) that $\boldsymbol{D}_T \to \mathbf{0}$ as $T \to \infty$. Thus

$$\boldsymbol{k}_T \left(E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime} - E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime} \right) \boldsymbol{k}_T^{\prime}$$

is asymptotically nonnegative definite. This is equivalent, as shown in Proposition 5.4 below, to saying that $\{\boldsymbol{Q}_T\}$ is AQS within $\mathcal{H}$. The two particular cases are accomplished by taking $\boldsymbol{k}_T = (E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime})^{-\frac{1}{2}}$ and $\boldsymbol{k}_T = \boldsymbol{\alpha}_T^{-\frac{1}{2}}$, respectively. This completes the proof.

Remark 5.4 (5.6) and (5.7) are actually equivalent.

On the grounds that one seeks to maximize $\mathcal{E}(\boldsymbol{G}_T)$, i.e., to minimize the estimating function variance $E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime}$, for $\boldsymbol{G}_T \in \mathcal{H}$, it would be sensible to avoid those estimating functions for which asymptotic relative efficiency comparisons are arbitrarily poor. With this in mind we give the following definition.

Definition 5.2 The sequence of estimating function $\{\boldsymbol{G}_T\} \in \mathcal{H}$ is *unacceptable* if there exists a sequence $\{\boldsymbol{Q}_T\} \in \mathcal{H}$ such that

$$\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime} \right)^{-\frac{1}{2}} E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime} \left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime} \right)^{-\frac{T}{2}} - \boldsymbol{I}$$

is unbounded as $T \to \infty$.

If we denote by $\mathcal{H}_U \subseteq \mathcal{H}$ the subset of all *unacceptable* estimating functions, the complementary set $\mathcal{H} - \mathcal{H}_U$ consists of all the *acceptable* estimating functions.

Proposition 5.2 A sequence of estimating functions $\{\boldsymbol{G}_T\} \in \mathcal{H}$ is unacceptable if and only if there exists a sequence of estimating functions $\{\boldsymbol{Q}_T\} \in \mathcal{H}$ and vectors $\{\boldsymbol{c}_T\}$ such that

$$\overline{\lim_{T\to\infty}} \; \frac{\boldsymbol{c}_T' E \boldsymbol{G}_T^{(n)} \boldsymbol{G}_T^{(n)'} \boldsymbol{c}_T}{\boldsymbol{c}_T' E \boldsymbol{Q}_T^{(n)} \boldsymbol{Q}_T^{(n)'} \boldsymbol{c}_T} = \infty.$$

The proof is fairly straightforward from the definition of unacceptability and we omit the details.

Proposition 5.3 Suppose $\{\boldsymbol{Q}_T\} \in \mathcal{H}$ is AQS where $\mathcal{H}$ is a linear space, then (5.7) holds for all $\{\boldsymbol{G}_T\} \in \mathcal{H} - \mathcal{H}_U$. On the set $\mathcal{H} - \mathcal{H}_U$, (5.6) is both necessary and sufficient for the AQS property.

Proof. For $\boldsymbol{\beta}_T$ an arbitrary $p \times p$ matrix, consider the sequence of estimating functions $\{(E\dot{\boldsymbol{G}}_T)\,\boldsymbol{\beta}_T\,\boldsymbol{Q}_T^{(n)} + \boldsymbol{G}_T\}$ that belongs to $\mathcal{H}$ since $\mathcal{H}$ is a linear space. Because $\{\boldsymbol{Q}_T\}$ is AQS within $\mathcal{H}$, by the definition of AQS and the fact that $E\boldsymbol{Q}_T^{(n)} = \boldsymbol{I}$, we have, upon writing $\boldsymbol{L} = E\,\boldsymbol{Q}_T^{(n)}\,\boldsymbol{Q}_T^{(n)'}$,

$$\begin{aligned}
&\underline{\lim}_{T\to\infty}\, \boldsymbol{k}_T'\, \boldsymbol{L}^{\frac{1}{2}} \left(\left(E\,\dot{\boldsymbol{G}}_T \boldsymbol{\beta}_T + E\,\dot{\boldsymbol{G}}_T \right)^{-1} \right. \\
&\quad \cdot E\left(\left((E\dot{\boldsymbol{G}}_T)\boldsymbol{\beta}_T \boldsymbol{Q}_T^{(n)} + \boldsymbol{G}_T \right) \left(\boldsymbol{Q}_T^{(n)'} \boldsymbol{\beta}_T{}' E\dot{\boldsymbol{G}}_T' + \boldsymbol{G}_T' \right) \right) \\
&\quad \cdot \left. \left(\boldsymbol{\beta}_T{}' E\dot{\boldsymbol{G}}_T' + E\dot{\boldsymbol{G}}_T' \right)^{-1} \right) \boldsymbol{L}^{-\frac{T}{2}}\, \boldsymbol{k}_T - \boldsymbol{k}_T' \boldsymbol{k}_T \geq 0,
\end{aligned} \tag{5.8}$$

for arbitrary sequence of unit vectors $\{\boldsymbol{k}_T\}$. For ease of notation, set $\boldsymbol{M} = E\boldsymbol{G}_T^{(n)} \boldsymbol{Q}_T^{(n)'}$ and $\boldsymbol{N} = E\boldsymbol{G}_T^{(n)} \boldsymbol{G}_T^{(n)'}$. The quantity on the left side of (5.8) for fixed T can be written as

$$\begin{aligned}
&\boldsymbol{k}_T' \boldsymbol{L}^{-\frac{1}{2}} (\boldsymbol{\beta}_T + \boldsymbol{I})^{-1} (E\dot{\boldsymbol{G}}_T)^{-1} \Big(E\Big(\Big((E\dot{\boldsymbol{G}}_T)\boldsymbol{\beta}_T \boldsymbol{Q}_T^{(n)} + \boldsymbol{G}_T \Big) \\
&\quad \cdot \left(\boldsymbol{Q}_T^{(n)'} \boldsymbol{\beta}_T{}' E\dot{\boldsymbol{G}}_T' + \boldsymbol{G}_T' \right) \Big) \\
&\quad - (E\dot{\boldsymbol{G}}_T)(\boldsymbol{\beta}_T + \boldsymbol{I})\boldsymbol{L}(\boldsymbol{\beta}_T{}' + \boldsymbol{I})E\dot{\boldsymbol{G}}_T' \Big) (E\dot{\boldsymbol{G}}_T')^{-1} (\boldsymbol{\beta}_T{}' + \boldsymbol{I})^{-1} \boldsymbol{L}^{-\frac{T}{2}} \boldsymbol{k}_T \\
= \;& \boldsymbol{k}_T' \boldsymbol{L}^{-\frac{1}{2}} (\boldsymbol{\beta}_T + \boldsymbol{I})^{-1} \Big(\boldsymbol{\beta}_T\, \boldsymbol{M}' + \boldsymbol{M}\, \boldsymbol{\beta}_T{}' \\
&\quad - \boldsymbol{\beta}_T \boldsymbol{L} - \boldsymbol{L}\, \boldsymbol{\beta}_T{}' + (\boldsymbol{N} - \boldsymbol{L}) \Big) (\boldsymbol{\beta}_T{}' + \boldsymbol{I})^{-1} \boldsymbol{L}^{-\frac{T}{2}} \boldsymbol{k}_T \\
= \;& \boldsymbol{k}_T' \boldsymbol{L}^{-\frac{1}{2}} (\boldsymbol{\beta}_T + \boldsymbol{I})^{-1} \boldsymbol{L}^{\frac{1}{2}} \left(\left(\boldsymbol{L}^{-\frac{1}{2}} \boldsymbol{\beta}_T \boldsymbol{L}^{\frac{1}{2}} \right) \left(\boldsymbol{L}^{-\frac{1}{2}} \boldsymbol{M}' \boldsymbol{L}^{-\frac{T}{2}} - \boldsymbol{I} \right) \right.
\end{aligned}$$

$$
\begin{aligned}
& + (\boldsymbol{L}^{-\frac{1}{2}}\boldsymbol{M}\,\boldsymbol{L}^{-\frac{T}{2}} - \boldsymbol{I})(\boldsymbol{L}^{-\frac{1}{2}}\boldsymbol{\beta}_T\boldsymbol{L}^{\frac{1}{2}})' + \boldsymbol{L}^{-\frac{1}{2}}\boldsymbol{N}\,\boldsymbol{L}^{-\frac{T}{2}} - \boldsymbol{I}\Big) \\
& \cdot \boldsymbol{L}^{\frac{T}{2}}(\boldsymbol{\beta}_T{}' + \boldsymbol{I})^{-1}\boldsymbol{L}^{-\frac{T}{2}}\boldsymbol{k}_T \\
= \;& \boldsymbol{k}_T'\boldsymbol{C}_T\Big((\boldsymbol{C}_T^{-1} - \boldsymbol{I})(\boldsymbol{L}^{-\frac{1}{2}}\boldsymbol{M}\,\boldsymbol{L}^{-\frac{T}{2}} - \boldsymbol{I}) \\
& + (\boldsymbol{L}^{-\frac{1}{2}}\boldsymbol{M}'\boldsymbol{L}^{-\frac{T}{2}} - \boldsymbol{I})'(\boldsymbol{C}_T^{-1} - \boldsymbol{I})' + \boldsymbol{L}^{-\frac{1}{2}}\boldsymbol{N}\,\boldsymbol{L}^{-\frac{T}{2}} - \boldsymbol{I}\Big)\boldsymbol{C}_T'\,\boldsymbol{k}_T \\
& \qquad (\text{ where } \quad \boldsymbol{C}_T = \boldsymbol{L}^{-\frac{1}{2}}(\boldsymbol{\beta}_T + \boldsymbol{I})^{-1}\boldsymbol{L}^{\frac{1}{2}} \;) \\
= \;& \boldsymbol{k}_T'(\boldsymbol{I} - \boldsymbol{C}_T)\Big(\boldsymbol{L}^{-\frac{1}{2}}\boldsymbol{M}'\,\boldsymbol{L}^{-\frac{T}{2}} - \boldsymbol{I}\Big)\boldsymbol{C}_T'\,\boldsymbol{k}_T \qquad (5.9) \\
& + \boldsymbol{k}_T'\,\boldsymbol{C}_T\Big(\boldsymbol{L}^{-\frac{1}{2}}\boldsymbol{M}'\boldsymbol{L}^{-\frac{T}{2}} - \boldsymbol{I}\Big)'(\boldsymbol{I} - \boldsymbol{C}_T)'\boldsymbol{k}_T \\
& + \boldsymbol{k}_T'\boldsymbol{C}_T\Big(\boldsymbol{L}^{-\frac{1}{2}}\boldsymbol{N}\,\boldsymbol{L}^{-\frac{T}{2}} - \boldsymbol{I}\Big)\boldsymbol{C}_T'\,\boldsymbol{k}_T.
\end{aligned}
$$

Since $\{\boldsymbol{G}_T\} \in \mathcal{H} - \mathcal{H}_U$ and $\boldsymbol{L}^{-\frac{1}{2}}\,\boldsymbol{N}\,\boldsymbol{L}^{-\frac{T}{2}} - \boldsymbol{I}$ is bounded, there exists a constant A large enough such that

$$\boldsymbol{x}'\left(\boldsymbol{L}^{-\frac{1}{2}}\,\boldsymbol{N}\,\boldsymbol{L}^{-\frac{T}{2}} - \boldsymbol{I}\right)\boldsymbol{x} \le A\|\boldsymbol{x}\|^2 \qquad (5.10)$$

for any p dimensional vector $\boldsymbol{x}$.

Suppose (5.7) does not hold. Then there exist two subsequences of vectors, denoted by $\{\boldsymbol{x}_T\}$ and $\{\boldsymbol{y}_T\}$ for simplicity in notation, such that $\boldsymbol{y}_T = (\boldsymbol{L}^{-\frac{1}{2}}\boldsymbol{M}'\,\boldsymbol{L}^{-\frac{T}{2}} - \boldsymbol{I})\boldsymbol{x}_T$, $\|\boldsymbol{y}_T\| = 1$ and $\|\boldsymbol{x}_T\| \le A$. Because $\boldsymbol{k}_T$ is an arbitrary unit vector, we can choose $\boldsymbol{k}_T = \boldsymbol{y}_T$. Moreover, since $\boldsymbol{\beta}_T$ is an arbitrary matrix, so is $\boldsymbol{C}_T$. Hence we may choose $\boldsymbol{C}_T$ so that $\boldsymbol{C}_T'\,\boldsymbol{y}_T = \lambda\,\boldsymbol{x}_T$, where $\lambda = -\frac{1}{2A+A^2}$. Using (5.10), the quantity in (5.9) is less than or equal to

$$
\begin{aligned}
& 2\lambda - 2\lambda^2 < \boldsymbol{x}_T, \boldsymbol{y}_T > + A\,\lambda^2\|\boldsymbol{x}_T\|^2 \\
\le \;& 2\lambda + 2\lambda^2 A + \lambda^2 A^2 = \lambda < 0,
\end{aligned}
$$

which contradicts (5.8). This completes the proof of (5.7) and the remaining part of the proof follows immediately from Proposition 5.1.

The next proposition gives five equivalent versions of the definition of asymptotic quasi-score.

Proposition 5.4 The following conditions are equivalent.

1) $\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{-\frac{1}{2}} E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime}\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{-\frac{T}{2}} - \boldsymbol{I}$ is asymptotic non-negative definite for all $\{\boldsymbol{G}_T\} \in \mathcal{H}$.

2) $\boldsymbol{I} - \left(E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime}\right)^{-\frac{1}{2}} E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\left(E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime}\right)^{-\frac{T}{2}}$ is asymptotic nonnegative definite for all $\{\boldsymbol{G}_T\} \in \mathcal{H}$.

3) For any nonzero p-dimensional vector $\{\boldsymbol{c}_T\}$,

$$\underline{\lim}_{T\to\infty} \frac{\boldsymbol{c}_T' E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime}\boldsymbol{c}_T}{\boldsymbol{c}_T' E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\boldsymbol{c}_T} \geq 1$$

for all $\{\boldsymbol{G}_T\} \in \mathcal{H}$.

4) There exists a sequence of $p\times p$ matrices $\{\boldsymbol{k}_T\}$ such that $\boldsymbol{k}_T E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\boldsymbol{k}_T' \to \boldsymbol{K}$ nondegenerate, and $\boldsymbol{k}_T E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime}\boldsymbol{k}_T' - \boldsymbol{k}_T E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\boldsymbol{k}_T'$ is asymptotic nonnegative definite for all $\{\boldsymbol{G}_T\} \in \mathcal{H}$.

5) There exists a sequence of $p \times p$ matrices $\{\boldsymbol{k}_T\}$ such that $\boldsymbol{k}_T\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{-1}\boldsymbol{k}_T' \to \boldsymbol{K}$ nondegenerate, and $\boldsymbol{k}_T\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{-1}\boldsymbol{k}_T' - \boldsymbol{k}_T\left(E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime}\right)^{-1}\boldsymbol{k}_T'$ is asymptotic nonnegative definite.

Proof. 1) ⇒ 3) Let $\boldsymbol{x}_T$ be an arbitrary unit vector. Then,

$$\underline{\lim}_{T\to\infty} \boldsymbol{x}_T'\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{-\frac{1}{2}} E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime}\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{-\frac{T}{2}}\boldsymbol{x}_T \geq 1.$$

Let $\boldsymbol{c}_T = a\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{-\frac{T}{2}}\boldsymbol{x}_T$; then

$$\underline{\lim}_{T\to\infty} \frac{\boldsymbol{c}_T' E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime}\boldsymbol{c}_T}{\boldsymbol{c}_T' E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\boldsymbol{c}_T} \geq 1.$$

Here $\boldsymbol{c}_T$ is an arbitrary vector because $\boldsymbol{x}_T$ is an arbitrary unit vector and a is an arbitrary nonzero constant.

3) ⇒ 1). Let $\boldsymbol{x}_T = \dfrac{\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{\frac{T}{2}}\boldsymbol{c}_T}{\left\|\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{\frac{T}{2}}\boldsymbol{c}_T\right\|}$. Then,

$$\underline{\lim}_{T\to\infty} \boldsymbol{x}_T'\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{-\frac{1}{2}} E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime}\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\right)^{-\frac{T}{2}}\boldsymbol{x}_T$$

$$= \underline{\lim}_{T\to\infty} \frac{\boldsymbol{c}_T' E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)\prime}\boldsymbol{c}_T}{\boldsymbol{c}_T' E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\boldsymbol{c}_T} \geq 1.$$

Here $\boldsymbol{x}_T$ can be regarded as an arbitrary unit vector because $\boldsymbol{c}_T$ is an arbitrary vector.

2) ⇔ 3) can be shown in a similar fashion.

1) ⇒ 4) and 2) ⇒ 5) are obvious.

4) ⇒ 3) For an arbitrary vector $\boldsymbol{c}_T$,

$$\underline{\lim}_{T\to\infty} \frac{\boldsymbol{c}_T' \left(E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)'} - E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'} \right) \boldsymbol{c}_T}{\boldsymbol{c}_T' \boldsymbol{k}_T^{-1} \boldsymbol{k}_T^{-1'} \boldsymbol{c}_T} \geq 0,$$

so that

$$\underline{\lim}_{T\to\infty} \left(\frac{\boldsymbol{c}_T' E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)'} \boldsymbol{c}_T}{\boldsymbol{c}_T' E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'} \boldsymbol{c}_T} - 1 \right) \frac{\boldsymbol{c}_T' E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'} \boldsymbol{c}_T}{\boldsymbol{c}_T' \boldsymbol{k}_T^{-1} \boldsymbol{k}_T^{-1'} \boldsymbol{c}_T} \geq 0,$$

and since

$$\underline{\lim}_{T\to\infty} \frac{\boldsymbol{c}_T' E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'} \boldsymbol{c}_T}{\boldsymbol{c}_T' \boldsymbol{k}_T^{-1} \boldsymbol{k}_T^{-1'} \boldsymbol{c}_T} \geq \lambda_{min}(\boldsymbol{K}) > 0,$$

because $\boldsymbol{K}$ is nondegenerate and thus is positive definite, we have

$$\underline{\lim}_{T\to\infty} \frac{\boldsymbol{c}_T' E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)'} \boldsymbol{c}_T}{\boldsymbol{c}_T' E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'} \boldsymbol{c}_T} \geq 1.$$

5) ⇒ 3) can be proved similarly.

The final proposition indicates that two sequences of asymptotic quasi-score estimating functions are asymptotically close under appropriate norms.

Proposition 5.5 Suppose $\{\tilde{\boldsymbol{G}}_T\} \in \mathcal{H}$ is AQS, $\mathcal{H}$ is a linear space, and there exists a sequence of matrices $\{\boldsymbol{k}_T\}$ such that

$$\lim_{T\to\infty} \boldsymbol{k}_T E\tilde{\boldsymbol{G}}_T^{(n)}\tilde{\boldsymbol{G}}_T^{(n)'} \boldsymbol{k}_T' = \boldsymbol{K}$$

where $\boldsymbol{K}$ is some nondegenerate matrix. Then the following statements are equivalent:

(i) $\{\boldsymbol{Q}_T\} \in \mathcal{H}$ is AQS,

(ii) $\lim_{T\to\infty} \boldsymbol{k}_T E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'} \boldsymbol{k}_T' = \boldsymbol{K}$,

(iii) $\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'}\right)^{-\frac{1}{2}} E\tilde{\boldsymbol{G}}_T^{(n)}\tilde{\boldsymbol{G}}_T^{(n)'} \left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'}\right)^{-\frac{T}{2}} \to \boldsymbol{I}$ as $T \to \infty$,

(iv) $\mathrm{trace}\left(\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'}\right)^{-\frac{1}{2}} E\tilde{\boldsymbol{G}}_T^{(n)}\tilde{\boldsymbol{G}}_T^{(n)'} \left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'}\right)^{-\frac{T}{2}}\right) \to p$ as $T \to \infty$,

(v) $\dfrac{\det\left(E\tilde{\boldsymbol{G}}_T^{(n)}\tilde{\boldsymbol{G}}_T^{(n)'}\right)}{\det\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'}\right)} \to 1$ as $T \to \infty$,

(vi) $\dfrac{\boldsymbol{c}_T' E\tilde{\boldsymbol{G}}_T^{(n)}\tilde{\boldsymbol{G}}_T^{(n)\prime}\boldsymbol{c}_T}{\boldsymbol{c}_T' E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)\prime}\boldsymbol{c}_T} \to 1$ as $T \to \infty$ for all p-dimensional vectors $\boldsymbol{c}_T \neq 0$,

(vii) $\boldsymbol{k}_T E\left(\left(\tilde{\boldsymbol{G}}_T^{(n)} - \boldsymbol{Q}_T^{(n)}\right)\left(\tilde{\boldsymbol{G}}_T^{(n)\prime} - \boldsymbol{Q}_T^{(n)\prime}\right)\right)\boldsymbol{k}_T' \to 0$ as $T \to \infty$.

The proof of this proposition is straightforward and we omit the details.

It is worthwhile to point out that under a small perturbation within $\mathcal{H}$, an asymptotic quasi-score is still asymptotic quasi-score. (This can be seen from (vii) in Proposition 5.5.) One can see that the relative difference of the information about the parameter contained in two sequences of asymptotic quasi-score estimating functions is arbitrarily small under suitable normalization.

5.3 Examples

5.3.1 Generalized Linear Model

A simple application involves the generalized linear model where the parameter is a p-vector $\boldsymbol{\beta}$. The distribution of the response y_n is assumed to belong to a natural exponential family, which may be written as

$$f(y_n|\,\theta_n) = c(y_n)\,\exp(\theta_n y_n - b(\theta_n))$$

and θ_n and $\boldsymbol{Z}_n'\boldsymbol{\beta}$ are related by a link function, $\boldsymbol{Z}_n$ being a p-vector of covariates. We assume y_n and θ_n are one dimensional for simplicity. Suppose the link is such that $\theta_n = u(\boldsymbol{Z}_n'\boldsymbol{\beta})$ for some differentiable function u. The score function is

$$\tilde{\boldsymbol{G}}_T = \sum_{i=1}^{T}\boldsymbol{Z}_i\dot{u}(\boldsymbol{Z}_i'\boldsymbol{\beta})(y_i - \dot{b}(u(\boldsymbol{Z}_i'\boldsymbol{\beta}))),$$

which is the optimal estimating function within $\mathcal{H}_T = \{\sum_{i=1}^T \boldsymbol{C}_i(y_i - \dot{b}(u(\boldsymbol{Z}_i\boldsymbol{\beta})))\}$ for all p-vectors $\boldsymbol{C}_i$, $i = 1, 2, ...$

Now suppose the assumption of the exponential family distribution of y_n and the link function u is slightly relaxed to only assuming that y_n has mean $\dot{b}(u(\boldsymbol{Z}_n'\boldsymbol{\beta}))$ and finite variance σ_n^2. This guarantees that $\{\tilde{G}_T\}$ is within $\mathcal{H}$. Now $\tilde{G}_T$ is no longer necessarily a score or quasi-score estimating function. However, under the assumption $E\tilde{\boldsymbol{G}}_T^{(n)}\tilde{\boldsymbol{G}}_T^{(n)\prime} \to 0$ and $\frac{\ddot{b}(u(Z_T\beta))}{\sigma_T^2} \to 1$ as $T \to \infty$, one can still show that $\{\tilde{\boldsymbol{G}}_T\}$ remains an AQS sequence within $\cup_{T=1}^{\infty}\mathcal{H}_T$.

5.3.2 Heteroscedastic Autoregressive Model

Our second example concerns a d-dimensional heteroscedastic autoregressive model. Let $\{\boldsymbol{X}_k, k = 1, 2, ..., T\}$ be a sample from a d-dimensional process satisfying

$$\boldsymbol{X}_k = \boldsymbol{\beta}\boldsymbol{X}_{k-1} + \boldsymbol{\epsilon}_k,$$

where $\boldsymbol{\beta}$ is a $d \times d$ matrix to be estimated and $\{\boldsymbol{\epsilon}_k; k = 1, 2, ...\}$ is a d-dimensional martingale difference sequence such that $E\boldsymbol{\epsilon}_k = \mathbf{0}$ and $E\boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_k' = \boldsymbol{\Lambda}_k$. One can conveniently write this model as

$$\boldsymbol{X}_k = (\boldsymbol{X}_{k-1}' \otimes \boldsymbol{I}_d)\boldsymbol{\theta} + \boldsymbol{\epsilon}_k,$$

where $\boldsymbol{\theta} = \text{vec}\,(\boldsymbol{\beta}) = (\beta_{11}, ..., \beta_{d1}, \beta_{21}, ..., \beta_{d2}, ..., \beta_{d1}, ...\beta_{dd})'$, the vector obtained from β by stacking its columns one on top of the other and for matrices $\boldsymbol{A} = (\boldsymbol{A}_{ij})$ and $\boldsymbol{B}$ the Kronecker product, $\boldsymbol{A} \otimes \boldsymbol{B}$ denotes the matrix whose (i, j), the block element, is the matrix $\boldsymbol{A}_{ij}\boldsymbol{B}$. The quasi-score estimating function based on the martingale difference sequence $\{\boldsymbol{\epsilon}_k\}$ is

$$\boldsymbol{Q}_T(\boldsymbol{\theta}) = \sum_{k=1}^{T} (\boldsymbol{X}_{k-1}' \otimes \boldsymbol{I}_d)' \boldsymbol{\Lambda}_k^{-1} (\boldsymbol{X}_k - (\boldsymbol{X}_{k-1}' \otimes \boldsymbol{I}_d)\boldsymbol{\theta}),$$

which contains the generally unknown $\boldsymbol{\Lambda}_k$. Let

$$\boldsymbol{G}_T(\boldsymbol{\theta}) = \sum_{k=1}^{T} (\boldsymbol{X}_{k-1}' \otimes \boldsymbol{I}_d)' (\boldsymbol{X}_k - (\boldsymbol{X}_{k-1}' \otimes \boldsymbol{I}_d)\boldsymbol{\theta}).$$

We shall show that $\{\boldsymbol{G}_T\}$ is a sequence of asymptotic quasi-score estimating functions under the conditions that $\boldsymbol{\Lambda}_k \to \boldsymbol{\Lambda}$, a nondegenerate matrix. That is, $\boldsymbol{G}_T$ is a quasi-score for the model when $\boldsymbol{\Lambda}_k$ is constant in k but unknown, and $\{\boldsymbol{G}_T\}$ remains as AQS when $\boldsymbol{\Lambda}_k$ is not constant but converges to a constant.

Now, writing $\boldsymbol{V}_k$ for $E(\boldsymbol{X}_{k-1}\boldsymbol{X}_{k-1}')$, we have

$$\begin{aligned}
E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'} &= \left(\sum_{k=1}^{T} \boldsymbol{V}_k \otimes \boldsymbol{\Lambda}_k^{-1}\right)^{-1}, \\
E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)'} &= \left(\sum_{k=1}^{T} \boldsymbol{V}_k \otimes \boldsymbol{I}_d\right)^{-1} \left(\sum_{k=1}^{T} \boldsymbol{V}_k \otimes \boldsymbol{\Lambda}_k\right) \left(\sum_{k=1}^{T} \boldsymbol{V}_k \otimes \boldsymbol{I}_d\right)^{-1},
\end{aligned}$$

it being easily checked that

$$\begin{aligned}
E\dot{\boldsymbol{Q}}_T &= -\sum_{k=1}^{T} \boldsymbol{V}_k \otimes \boldsymbol{\Lambda}_k^{-1} = -E\boldsymbol{Q}_T\boldsymbol{Q}_T', \\
E\dot{\boldsymbol{G}}_T &= -\sum_{k=1}^{T} \boldsymbol{V}_k \otimes \boldsymbol{I}_d, \\
E\boldsymbol{G}_T\boldsymbol{G}_T' &= \sum_{k=1}^{T} \boldsymbol{V}_k \otimes \boldsymbol{\Lambda}_k,
\end{aligned}$$

since for a $d \times d$ matrix $\boldsymbol{M}$,

$$E\left[\left(\boldsymbol{X}_{k-1}' \otimes \boldsymbol{I}_d\right)' \boldsymbol{M} \left(\boldsymbol{X}_{k-1}' \otimes \boldsymbol{I}_d\right)\right] = \boldsymbol{V}_k \otimes \boldsymbol{M}.$$

Since $\boldsymbol{\Lambda}_k^{-1} \to \boldsymbol{\Lambda}^{-1}$, a positive definite matrix, for any d-dimensional vector sequence $\{\boldsymbol{b}_k\}$,

$$\frac{\boldsymbol{b}_k' \boldsymbol{\Lambda}_k^{-1} \boldsymbol{b}_k}{\boldsymbol{b}_k' \boldsymbol{\Lambda}^{-1} \boldsymbol{b}_k} \to 1. \tag{5.11}$$

Since $\boldsymbol{V}_k$ is nonnegative definite, we can write $\boldsymbol{V}_k = \boldsymbol{T}\boldsymbol{T}'$ for some matrix $\boldsymbol{T} = (t_{ij})_{d\times d}$. For any d^2-dimensional vector sequence $\{\boldsymbol{c}_k\}$, let $\boldsymbol{c}_k' = (\boldsymbol{c}_{k1}', \boldsymbol{c}_{k2}', ..., \boldsymbol{c}_{kd}')$ where $\boldsymbol{c}_{ki}$ is a d-dimensional vector for $1 \le i \le d$. Now we can write

$$\boldsymbol{c}_k'(\boldsymbol{V}_k \otimes \boldsymbol{\Lambda}_k^{-1})\,\boldsymbol{c}_k = \sum_{l=1}^{d}\left(\left(\sum_{i=1}^{d} t_{il}\,\boldsymbol{c}_{ki}\right)' \boldsymbol{\Lambda}_k^{-1}\left(\sum_{i=1}^{d} t_{il}\,\boldsymbol{c}_{ki}\right)\right).$$

Replacing $\boldsymbol{\Lambda}_k^{-1}$ by $\boldsymbol{\Lambda}^{-1}$, the above formula still holds. Therefore it follows from (5.11) that

$$\frac{\boldsymbol{c}_k'(\boldsymbol{V}_k \otimes \boldsymbol{\Lambda}_k^{-1})\,\boldsymbol{c}_k}{c_k'(\boldsymbol{V}_k \otimes \boldsymbol{\Lambda}^{-1})\,\boldsymbol{c}_k} \to 1. \tag{5.12}$$

Also, $\lambda_{min}(\mathcal{E}(\boldsymbol{Q}_T)) \to \infty$ gives

$$\boldsymbol{c}_T'(\sum_{k=1}^{T} \boldsymbol{V}_k \otimes \boldsymbol{\Lambda}_k^{-1})\,\boldsymbol{c}_T \to \infty \tag{5.13}$$

as $T \to \infty$. Combining (5.12) and (5.13), it follows that

$$\frac{\boldsymbol{c}_T'(\sum_{k=1}^{T} \boldsymbol{V}_k \otimes \boldsymbol{\Lambda}_k^{-1})\,\boldsymbol{c}_T}{\boldsymbol{c}_T'(\sum_{k=1}^{T} \boldsymbol{V}_k \otimes \boldsymbol{\Lambda}^{-1})\,\boldsymbol{c}_T} \to 1$$

as $T \to \infty$. But since $\{\boldsymbol{c}_T\}$ can be an arbitrary unit vector sequence, so we have

$$\frac{\det(\sum_{k=1}^{T} \boldsymbol{V}_k \otimes \boldsymbol{\Lambda}_k^{-1})}{\det(\sum_{k=1}^{T} \boldsymbol{V}_k \otimes \boldsymbol{\Lambda}^{-1})} \to 1$$

as $T \to \infty$, and

$$\begin{aligned} &\lim_{T\to\infty} \frac{\det(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'})}{\det(E\boldsymbol{G}_T^{(n)}\boldsymbol{G}_T^{(n)'})} \\ = &\lim_{T\to\infty} \frac{\det((\sum_{k=1}^{T} \boldsymbol{V}_k) \otimes \boldsymbol{\Lambda}^{-1})^{-1}}{(\det(\sum_{k=1}^{T} \boldsymbol{V}_k))^{-d}\det((\sum_{k=1}^{T} \boldsymbol{V}_k) \otimes \boldsymbol{\Lambda})(\det(\sum_{k=1}^{T} \boldsymbol{V}_k))^{-d}} \\ = &\lim_{T\to\infty} \frac{(\det(\sum_{k=1}^{T} \boldsymbol{V}_k))^{-d}(\det(\boldsymbol{\Lambda}))^{d}}{(\det(\sum_{k=1}^{T} \boldsymbol{V}_k))^{-d}(\det(\sum_{k=1}^{T} \boldsymbol{V}_k))^{d}(\det(\boldsymbol{\Lambda}))^{d}(\det(\sum_{k=1}^{T} \boldsymbol{V}_k))^{-d}} \\ = &\ 1. \end{aligned}$$

Because $\boldsymbol{Q}_T$ is a quasi-score, by Proposition 5.4 Condition (v), we know $\{\boldsymbol{G}_T\}$ is an AQS sequence of estimating functions.

5.3.3 Whittle Estimation Procedure

As a final illustration of the AQL methodology we shall show that the widely used estimation procedure based on the smoothed periodogram, which dates back to remarkable early work of Whittle (1951), provides an AQL-estimator under a broad range of conditions. These encompass many random processes or random fields with either short or long range dependence. Here we shall deal with the one-dimensional random process case. For the multivariate random process and random fields cases see Heyde and Gay (1989), (1993).

Suppose that the stationary random process $\{X_t\}$, with $E\,X_t = 0$, is observed at the times $t = 1, 2, \ldots, T$ and let

$$I_T(\lambda) = \frac{1}{2\pi T} \left| \sum_{i=1}^{T} X_j\, \boldsymbol{e}^{-ij\,\lambda} \right|^2$$

be the corresponding periodogram. The spectral density of the random process is $f(\lambda;\, \boldsymbol{\theta}, \sigma^2)$, where the one-step prediction variance is

$$\sigma^2 = 2\pi\, \exp\left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} \log f(\lambda;\, \boldsymbol{\theta},\, \sigma^2)\, d\lambda \right\}, \tag{5.14}$$

and we shall write

$$g(\lambda;\, \boldsymbol{\theta}) = 2\pi\, \sigma^{-2}\, f(\lambda;\, \boldsymbol{\theta},\, \sigma^2).$$

We shall assume that σ^2 is specified while $\boldsymbol{\theta}$, of dimension p, is to be estimated. It should be noted that (5.14) gives

$$\int_{-\pi}^{\pi} \log g(\lambda;\, \boldsymbol{\theta})\, d\lambda = 0,$$

which is equivalent to the fact that $\sigma^{-1} X_t$ has prediction variances independent of $\boldsymbol{\theta}$ (e.g., Rosenblatt (1985, Chapter III, Section 3)). We shall initially assume further that f is a continuous function of λ, including at $\lambda = 0$ (short range dependence), and is continuously differentiable in $\boldsymbol{\theta}$ as a function of $(\lambda, \boldsymbol{\theta})$.

We consider the class of estimating functions

$$\boldsymbol{G}_T = \int_{-\pi}^{\pi} \boldsymbol{A}(\lambda)\, [I_T(\lambda) - EI_T(\lambda)]\; d\lambda$$

constructed from smoothed periodograms where the smoothing function $\boldsymbol{A}(\lambda)$ (a vector of dimension p) is square integrable and symmetric about zero, that is,

$$\int_{-\pi}^{\pi} \boldsymbol{A}'(\lambda)\, \boldsymbol{A}(\lambda)\, d\lambda < \infty, \qquad \boldsymbol{A}(\lambda) = \boldsymbol{A}(-\lambda).$$

We shall show that under a considerable diversity of conditions, an asymptotic quasi-score sequence of estimating functions for $\boldsymbol{\theta}$ is given by

$$\boldsymbol{G}_T^* \;=\; \int_{-\pi}^{\pi} (g(\lambda;\, \boldsymbol{\theta}))^{-2}\, \frac{\partial g(\lambda; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\, [I_T(\lambda) - EI_T(\lambda)]\, d\lambda$$

$$= \int_{-\pi}^{\pi} (g(\lambda;\, \boldsymbol{\theta}))^{-2}\, \frac{\partial g(\lambda; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\, [I_T(\lambda) - f(\lambda;\, \boldsymbol{\theta}, \sigma^2)]\, d\lambda + o(1)$$

as $T \to \infty$. The corresponding AQL estimator $\boldsymbol{\theta}_T$ obtained from the estimating equation

$$\boldsymbol{G}_T^*(\boldsymbol{\theta}) = \mathbf{0}$$

is then asymptotically equivalent to the Whittle estimator obtained by choosing $\boldsymbol{\theta}$ to minimize

$$\int_{-\pi}^{\pi} \left[\log f(\lambda;\, \boldsymbol{\theta}, \sigma^2) + I_T(\lambda)(f(\lambda; \boldsymbol{\theta}, \sigma^2))^{-1}\right]\, d\lambda, \tag{5.15}$$

i.e., to solve

$$\int_{-\pi}^{\pi} (g(\lambda;\, \boldsymbol{\theta}))^{-2}\, \frac{\partial g(\lambda; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\, [I_T(\lambda) - f(\lambda;\, \boldsymbol{\theta}, \sigma^2)]\, d\lambda = 0.$$

The idea to minimize (5.15) comes from the observation that if the random process X_t has a Gaussian distribution, it is plausible that T^{-1} times the log likelihood of the data could be approximated by

$$-\frac{1}{2}\log 2\pi\, \sigma^2 - (2\, \sigma^2)^{-1} \int_{-\pi}^{\pi} I_T(\lambda)(g(\lambda;\, \boldsymbol{\theta}))^{-1}\, d\lambda$$

$$= -\log 2\,\pi - \frac{1}{2(2\pi)} \int_{-\pi}^{\pi} \left[\log f(\lambda;\, \boldsymbol{\theta}, \sigma^2) + I_T(\lambda)(f(\lambda; \boldsymbol{\theta}, \sigma^2))^{-1}\right]\, d\lambda.$$

It is, in fact, now a commonly used strategy in large sample estimation problems to make the Gaussian assumption, maximize the corresponding likelihood, and then show that the estimator still makes good sense without the Gaussian assumption (e.g., Hannan (1970, Chapter 6, Section 6; 1973)).

Our first specific illustration concerns the linear process

$$X_t = \sum_{u_1=-\infty}^{\infty} g_u e_{t-u},$$

with

$$\begin{aligned} Ee_t &= 0, \\ Ee_s e_t &= \begin{cases} \sigma^2 & \text{if} \quad s = t, \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

and

$$\sum_{u_1=-\infty}^{\infty} g_u^2 < \infty.$$

There are various somewhat different regularity conditions under which we could proceed and since the discussion is intended to be illustrative rather than to present minimal conditions, we shall, for convenience, use results from Section 6, Chapter IV of Rosenblatt (1985). The same conclusions, can be deduced under significantly different conditions using results of various other authors (e.g., via Theorem 4 of Parzen (1957) or Theorem 6.2.4, p. 427 of Priestley (1981)).

We suppose that $\{X_t\}$ is a strongly mixing stationary random process with $EX_t^8 < \infty$ and cumulants up to eighth order absolutely summable. Then, from the discussion leading up to Theorem 7, p. 118 of Rosenblatt (1985), we find that if

$$\begin{aligned} \boldsymbol{G}_T &= \int_{-\pi}^{\pi} \boldsymbol{A}(\lambda)[I_T(\lambda) - EI_T(\lambda)]\, d\lambda, \\ \boldsymbol{G}_T^* &= \int_{-\pi}^{\pi} \boldsymbol{A}^*(\lambda)[I_T(\lambda) - EI_T(\lambda)]\, d\lambda, \end{aligned}$$

then

$$\begin{aligned} TE\boldsymbol{G}_T\boldsymbol{G}_T^{*'} &\to (2\pi)\Big\{2\int_{-\pi}^{\pi} \boldsymbol{A}(\lambda)(\boldsymbol{A}^*(\lambda))' f^2(\lambda)\, d\lambda \\ &\qquad + \int_{-\pi}^{\pi} f_4(\lambda, -\mu, \mu)\boldsymbol{A}(\lambda)(\boldsymbol{A}^*(\lambda))'\, d\lambda\, d\mu\Big\} \end{aligned} \tag{5.16}$$

as $T \to \infty$, with $f(\lambda)$ now being written instead of $f(\lambda;\, \boldsymbol{\theta}, \sigma^2)$. Here $f_4(\lambda, \mu, \eta)$ is the fourth order cumulant spectral density

$$f_4(\lambda, \mu, \eta) = (2\pi)^{-3} \sum_{a,b,d} c_{a,b,d}\, e^{-i(a\lambda + b\,\mu + d\,\eta)}$$

with

$$c_{a,b,d} = \text{cum}(X_t, X_{t+a}, X_{t+b}, X_{t+d}) = (\kappa - 3)\sigma^4 \sum_u g_u g_{u+a} g_{u+b} g_{u+d},$$

$(\kappa - 3)\sigma^4$ being the fourth cumulant of X_t (which is zero in the case where the process is Gaussian). But, analogously to the discussion of Chapter II, Section 4 of Rosenblatt (1985, pp. 46, 47), we find that

$$f_4(\lambda, -\mu, \mu) = (2\pi)^{-1}(\kappa - 3)\sigma^4 f(\lambda) f(\mu),$$

and hence the second term on the right-hand side of (5.16) is

$$(2\pi)^{-1}(\kappa - 3)\,\sigma^4 \left(\int_{-\pi}^{\pi} \boldsymbol{A}(\lambda) f(\lambda)\, d\lambda\right)\left(\int_{-\pi}^{\pi} \boldsymbol{A}^*(\mu) f(\mu)\, d\mu\right)',$$

which is zero if the random field is Gaussian or if the class of smoothing functions $\{\boldsymbol{A}(\lambda)\}$ is chosen so that

$$\int_{-\pi}^{\pi} \boldsymbol{A}(\lambda) f(\lambda)\, d\lambda = 0.$$

We shall henceforth suppose that either (or both) of these conditions hold and then

$$T\, E\boldsymbol{G}_T\boldsymbol{G}_T^{*'} \to 2(2\pi)\int_{-\pi}^{\pi}\boldsymbol{A}(\lambda)(\boldsymbol{A}^*(\lambda))'f^2(\lambda)\,d\lambda \tag{5.17}$$

as $T \to \infty$.

Also,

$$\begin{aligned} E\dot{\boldsymbol{G}}_T &= -\int_{-\pi}^{\pi}\boldsymbol{A}(\lambda)\left(\frac{\partial}{\partial\theta}\,EI_T(\lambda)\right)'\,d\lambda \\ &\to -\int_{-\pi}^{\pi}\boldsymbol{A}(\lambda)(\dot{f}(\lambda))'\,d\lambda, \end{aligned} \tag{5.18}$$

provided $\dot{f}(\lambda)$ is square integrable over $[-\pi, \pi]$, since

$$\frac{\partial}{\partial\theta}\,EI_T(\lambda) \to \dot{f}(\lambda)$$

in the L_2 norm by standard Fourier methods (e.g. Hannan (1970, p. 508)).

Then, from (5.17) and (5.18) we see that (5.6) holds with $k_T = T$ if

$$\boldsymbol{A}^*(\lambda) = \dot{f}(\lambda)(f(\lambda))^{-2},$$

which gives the required AQL property.

Under the same conditions and a number of variants thereof it can be further deduced that if $\bar{\boldsymbol{\theta}}_T$ is the estimator obtained from the estimating equation $\boldsymbol{G}_T(\boldsymbol{\theta}) = \mathbf{0}$, then (see e.g. the discussion of the proof of Theorem 8, p. 119 of Rosenblatt (1985))

$$T^{1/2}(\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} N_p\left(\mathbf{0},\,(\boldsymbol{A}'\boldsymbol{B}^{-1}\boldsymbol{A})^{-1}\right),$$

where

$$\begin{aligned} \boldsymbol{A} &= -\int_{-\pi}^{\pi}\boldsymbol{A}(\lambda)(\dot{f}(\lambda))'\,d\lambda, \\ \boldsymbol{B} &= -\int_{-\pi}^{\pi}\boldsymbol{A}(\lambda)\boldsymbol{A}'(\lambda)f^2(\lambda)\,d\lambda \end{aligned}$$

and that the inverse covariance matrix $\boldsymbol{A}'\boldsymbol{B}^{-1}\boldsymbol{A}$ is maximized in the partial order of nonnegative definite matrices for the AQL case where $\boldsymbol{A}(\lambda)$ takes the value

$$\boldsymbol{A}^*(\lambda) = \dot{f}(\lambda)(f(\lambda))^{-2}.$$

These considerations place the optimality results for random processes of Kulperger (1985, Theorem 2.1 and Corollary) and Kabaila (1980, Theorem 3.1) into more general perspective. The matrix maximization result mentioned above appears in the latter paper.

The next specific example on the Whittle procedure concerns a random process exhibiting long range dependence. We shall consider the case where the spectral density

$$f(x;\, \boldsymbol{\theta}, \sigma^2) \sim \sigma^2 |x|^{-\alpha(\boldsymbol{\theta})} L_{\boldsymbol{\theta}}(x)$$

as $x \to 0$, where $0 < \alpha(\boldsymbol{\theta}) < 1$ and $L_{\boldsymbol{\theta}}(x)$ varies slowly at zero. A typical example is provided by the fractional Brownian process. This is a stationary Gaussian process with zero mean and covariance

$$E(X_n X_{n+k}) = \frac{1}{2}\, c \left\{ |k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H} \right\} \sim c\, H(2H-1) k^{2H-2}$$

as $k \to \infty$, where H is a parameter satisfying $\frac{1}{2} < H < 1$ and $c > 0$, while

$$f(x;\, H) \sim \frac{1}{2}\, cF(H) |x|^{1-2H}$$

as $x \to 0$ with

$$f(H) = \left\{ \int_{-\pi}^{\pi} (1 - \cos x) |x|^{-1-2H} \, dx \right\}^{-1}.$$

The Whittle estimation procedure (5.15) for $\boldsymbol{\theta}$ in this context and subject to the assumption of a Gaussian distribution has been studied in some detail by Fox and Taqqu (1986), (1987) and Dahlhaus (1988). In particular, it follows from Theorem 1 of Fox and Taqqu (1987) that if

$$\boldsymbol{G}_T = \int_{-\pi}^{\pi} \boldsymbol{A}(x) [I_T(x) - EI_T(x)] \, dx$$

with $\|\boldsymbol{A}(x)\| = O(|x|)^{-\beta-\delta}$ as $x \to 0$ for some $\beta < 1$ and each $\delta > 0$ and $\alpha + \beta < \frac{1}{2}$, then

$$T E \boldsymbol{G}_T \boldsymbol{G}_T' \to 4\pi \int_{-\pi}^{\pi} f^2(x) \boldsymbol{A}(x) \boldsymbol{A}'(x) \, dx \tag{5.19}$$

as $T \to \infty$; see also Proposition 1 of Fox and Taqqu (1986). Furthermore, provided that

$$\frac{\partial f}{\partial \theta_j} \in L^a, \quad 1 \le j \le p, \quad 1 \le a < \alpha^{-1},$$

then

$$\frac{\partial}{\partial \theta_j}\, EI_T(x) \to \frac{\partial f}{\partial \theta_j}$$

in the L^a norm (e.g., Hannan (1970, p. 508)) and from Hölder's inequality we find that

$$E\dot{\boldsymbol{G}}_T = -\int_{-\pi}^{\pi} \boldsymbol{A}(x) \left(\frac{\partial}{\partial \boldsymbol{\theta}}\, EI_T(x) \right)' dx \to -\int_{-\pi}^{\pi} \boldsymbol{A}(x) (\dot{f}(x))' \, dx. \tag{5.20}$$

Then, from (5.19) and (5.20),

$$4\pi T^{-1}(E\dot{\boldsymbol{G}}_T)'(E\boldsymbol{G}_T\boldsymbol{G}_T')^{-1}(E\dot{\boldsymbol{G}}_T) \to \left(\int_{-\pi}^{\pi} \boldsymbol{A}(x)(\dot{f}(x))'\,dx\right) \tag{5.21}$$

$$\cdot \left(\int_{-\pi}^{\pi} f^2(x)\boldsymbol{A}(x)\boldsymbol{A}'(x)\,dx\right)^{-1} \left(\int_{-\pi}^{\pi} \boldsymbol{A}(x)\dot{f}(x))'\,dx\right)$$

and the right-hand side of (5.21) is maximized in the partial order of nonnegative definite matrices when $\boldsymbol{A}(x)$ takes the value

$$\boldsymbol{A}^*(x) = \dot{f}(x)(f(x))^{-2}$$

(Kabaila (1980, Theorem 3.1)). This gives the AQL property for the corresponding $\{\boldsymbol{G}_T^*\}$ by direct application of Definition 5.1.

5.3.4 Addendum to the Example of Section 5.1

Earlier approaches to this estimation problem have used conditional least squares or *ad hoc* methods producing essentially similar results. This amounts to using the estimating function (5.1) with the terms $\sigma^2 X_{t-1} + \eta^2$ removed. For references and details of the approach see Hall and Heyde (1980, Chapter 6.3). The essential point is that quasi-likelihood or asymptotic quasi-likelihood will offer advantages in asymptotic efficiency and these may be substantial.

The character of the limiting covariance matrices associated with different estimators precludes direct general comparison other than via inequalities. We therefore give a numerical example as a concrete illustration. This concerns the Smoluchowski model for particles in a fluid in which the offspring distribution is assumed to be Poisson and the immigration distribution to be Bernoulli; for a discussion see Heyde and Seneta (1972). The data we use is a set of 505 observations from Fürth (1918, Tabelle 1). We shall compare the asymptotic covariance matrices which come from the methods (i) conditional least squares (e.g., Hall and Heyde (1980, Chapter 6.3); Winnicki (1988, Theorem 3.B)), (ii) Wei and Winnicki's weighted conditional least squares (e.g., Winnicki (1988, Theorem 3.C)), (iii) asymptotic quasi-likelihood. All quantities are estimated from the data, and since the covariance matrices are all functions m and λ, common point estimates of m and λ (those which come from (i)) are used as a basis for the comparison. Methods (ii), (iii) produce slightly different point estimates.

We find that $\hat{m} = 0.68$, $\hat{\lambda} = 0.51$ and that the estimated asymptotic covariance matrices for $(505)^{\frac{1}{2}}(\hat{m} - m, \hat{\lambda} - \lambda)'$ in the cases (i), (ii), (iii) are, respectively,

$$\begin{pmatrix} 0.81 & -1.04 \\ -1.04 & 2.05 \end{pmatrix}, \quad \begin{pmatrix} 0.76 & -0.74 \\ -0.74 & 1.57 \end{pmatrix}, \quad \begin{pmatrix} 0.75 & -0.73 \\ -0.73 & 1.56 \end{pmatrix}.$$

The reduction in variance estimates obtained by using (iii) rather than (i) or (ii) in this particular case is not of practical significance. Approximate 95%

confidence intervals for m are 0.68 ± 0.08 in each case, and for λ, (i) gives 0.51 ± 0.12 while (ii) and (iii) improve this to 0.51 ± 0.11. Although quasi-likelihood does not perform better than the simpler methods here, it nevertheless remains as the benchmark for minimum width confidence intervals.

5.4 Bibliographic Notes

The bulk of this chapter follows Chen and Heyde (1995), a sequel to Heyde and Gay (1989). The branching process discussion is from Heyde and Lin (1992). If $\{\boldsymbol{G}_T\}$ is an asymptotic quasi-score sequence under Definition 5.2, then it is also under the definition given in Heyde and Gay (1989), which is relatively weaker. In particular, the two definitions are equivalent if the parameter $\boldsymbol{\theta}$ is one dimensional or if $\lambda_{max}\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'}\right)/\lambda_{min}\left(E\boldsymbol{Q}_T^{(n)}\boldsymbol{Q}_T^{(n)'}\right)$ is bounded for all T where $\lambda_{max}(\cdot)$ $(\lambda_{min}(\cdot))$ denotes the largest (smallest) eigenvalue.

A corresponding theory to that of this chapter, based on conditional covariances rather than ordinary ones, should be able to be developed straightforwardly. This would be easier to use in some stochastic process problems.

5.5 Exercises

1. Suppose that $\{X_1, X_2, \dots, X_T\}$ is a sample from a normal distribution $N(\mu, \sigma^2)$, where σ^2 is to be estimated and μ is a nuisance parameter. Recall that $\sum_{t=1}^T (X_t - \mu)^2$ has a $\sigma^2\chi_T^2$ distribution and $\sum_{t=1}^T (X_t - \bar{X})^2$ a $\sigma^2\chi_{T-1}^2$ distribution, $\bar{X}$ being the sample mean. Show that

$$\sum_{t=1}^T [(X_t - \bar{X})^2 - \sigma^2]$$

can be interpreted as an AQS estimating function for σ^2.

2. Let $\{X_1, X_2, \dots, X_T\}$ be a sample from the process

$$X_t = \theta\, X_{t-1} + u_t,$$

where $\theta > 1$ is to be estimated and if $\mathcal{F}_t = \sigma(X_t, \dots, X_1)$ are the past-history σ-fields,

$$E(u_t \,|\, \mathcal{F}_{t-1}) = 0, \quad E(u_t^2 \,|\, \mathcal{F}_{t-1}) = \eta^2\, X_{t-1}^2 + \sigma^2\, X_{t-1},$$

η^2 and σ^2 being nuisance parameters. Consider the family of estimating functions

$$\mathcal{H} = \{\sum_{t=1}^T \alpha_t (Z_t - \theta\, X_{t-1})\}$$

where α_t's are $\mathcal{F}_{t-1}$-measurable and show that, on the set $\{X_t \to \infty\}$, $\sum_{t=1}^T X_{t-1}^{-1}(X_t - \theta\, X_{t-1})$ is an AQS estimating function.

3. For a regression model of the form

$$\boldsymbol{Y} = \boldsymbol{\mu}(\boldsymbol{\theta}) + \boldsymbol{e}$$

with $E\boldsymbol{e} = \boldsymbol{0}$ and $\boldsymbol{V} = E\boldsymbol{e}\,\boldsymbol{e}'$ (such as is studied in Section 2.4), suppose that the quasi-score estimating function from the space

$$\mathcal{H} = \{\boldsymbol{A}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta}))\}$$

for $p \times p$ matrices $\boldsymbol{A}$ not depending on $\boldsymbol{\theta}$ for which $E(\boldsymbol{A}\,\dot{\boldsymbol{\mu}})$ and $E(\boldsymbol{A}\,\boldsymbol{e}\,\boldsymbol{e}'\,\boldsymbol{A}')$ are nonsingular and $E(\boldsymbol{e}\,\boldsymbol{e}' \mid \dot{\boldsymbol{\mu}}) = \boldsymbol{V}$ is $\dot{\boldsymbol{\mu}}'\,\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$ (see, e.g., Theorem 2.3). Suppose that $\boldsymbol{V}$ is not known and is replaced by an approximate covariance matrix $\boldsymbol{W}$. Show that $\dot{\boldsymbol{\mu}}'\,\boldsymbol{W}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$ is an asymptotic quasi-score estimating function iff

$$\det(E\dot{\boldsymbol{\mu}}'\,\boldsymbol{W}^{-1}\boldsymbol{V}\,\boldsymbol{W}^{-1}\dot{\boldsymbol{\mu}})/\det(E\dot{\boldsymbol{\mu}}'\,\boldsymbol{V}^{-1}\dot{\mu}) \to 1$$

as the sample size $T \to \infty$. This gives a necessary and sufficient condition for a GEE (see Section 2.4.3) to be fully efficient asymptotically.

Chapter 6

Combining Estimating Functions

6.1 Introduction

It is often possible to collect a number of estimating functions, each with information to offer about an unknown parameter. These can be combined much more readily than the estimators therefrom and, indeed, simple addition is often appropriate. This chapter addresses the issue of how to perform the combination optimally.

Pragmatic considerations led to the combination of likelihood scores. Likelihood based inference encounters a range of possible problems including computational difficulty, information on only part of the model, questionable distributional assumptions, etc. As a compromise, the method of composite likelihoods, essentially the summing of any available likelihood scores of relevance was suggested (e.g., Besag (1975), Lindsay (1988)).

A good example concerns observations on a lattice of points in the plane where the conditional distributions of the form $f(y_i \mid y_{[i]})$ are specified, the $y_{[i]}$ denoting a prescribed set of variables, for example the nearest neighbors of i.

Let $\boldsymbol{y}$ be the vector of the y_i's in dictionary order and write $\boldsymbol{W} = (w_{ij})$ for the matrix of 0's and 1's such that $w_{ii} = 0$, $w_{ij} = w_{ji} = 1$ if i and j are neighbors, while $\boldsymbol{w}_i$ denotes the ith column of $\boldsymbol{W}$.

Then, if $y_i \mid y_{[i]}$ has a $N(\beta\, \boldsymbol{w}_i'\, \boldsymbol{y},\, \tau^2)$ distribution for parameters β, τ^2, there is a consistent joint distribution of the form

$$\boldsymbol{y} \overset{\mathrm{d}}{=} N(\boldsymbol{0},\, \sigma^2(\boldsymbol{I} - \beta \boldsymbol{W})^{-1}), \quad \sigma^2 = \sigma^2(\tau, \beta).$$

Taking σ^2 as known and setting it equal to 1 for convenience, it is easily seen that the score function associated with $\boldsymbol{y}$ is

$$\boldsymbol{y}'\, \boldsymbol{W}\, \boldsymbol{y} + \frac{d}{d\beta} \log \det (\boldsymbol{I} - \beta\, \boldsymbol{W}).$$

There are fundamental computational difficulties in estimating β from this.

On the other hand, the composite score obtained from summing the conditional scores of the y_i's is

$$\boldsymbol{y}'\, \boldsymbol{W}\, \boldsymbol{y} - \beta\, \boldsymbol{y}'\, \boldsymbol{W}^2\, \boldsymbol{y},$$

and this is simple to use. This kind of approach is easy to formalize into a framework of composite quasi-likelihood, and this will be the topic of the next section.

The setting that we shall consider is a general one in which maximum likelihood may be difficult or impossible to use; the context may be a semiparametric one, for example. However, we suppose that we are able to piece together a set of quasi-likelihood estimating functions, possibly based on conditional or marginal information. The problem is how to combine these most efficiently for estimation purposes.

6.2 Composite Quasi-Likelihoods

Let $\{\boldsymbol{X}_t,\ t \le T\}$ be a sample of discrete or continuous data that is randomly generated with values in r-dimensional Euclidean space whose distibution involves a "parameter" $\boldsymbol{\theta}$ taking values in an open subset $\boldsymbol{\Theta}$ of p-dimensional Euclidean space. The true value of the "parameter" is $\boldsymbol{\theta}_0$ and this is to be estimated.

Suppose that the possible probability measures for $\{\boldsymbol{X}_t\}$ are $\{P_{\boldsymbol{\theta}}\}$ and that each $(\Omega,\ \mathcal{F},\ P_{\boldsymbol{\theta}})$ is a complete probability space.

We shall as usual confine attention to the class $\mathcal{G}$ of zero mean square integrable estimating functions $\boldsymbol{G}_T = \boldsymbol{G}_T(\{\boldsymbol{X}_t,\ t \le T\},\ \boldsymbol{\theta})$ for which $E\boldsymbol{G}_T(\boldsymbol{\theta}) = \boldsymbol{0}$ and $E\dot{\boldsymbol{G}}_T(\boldsymbol{\theta})$ and $E\boldsymbol{G}_T(\boldsymbol{\theta})\,\boldsymbol{G}'_T(\boldsymbol{\theta})$ are nonsingular for each $P_{\boldsymbol{\theta}}$. Here $\boldsymbol{G}_T$ is a vector of dimension p.

We consider a class of $\mathcal{K} \subseteq \mathcal{G}$ of estimating functions $\boldsymbol{G}_T$ that are a.s. differentiable with respect to the components of $\boldsymbol{\theta}$ and such that

$$E\dot{\boldsymbol{G}}_T = \left(\frac{E\partial G_{T,i}}{\partial \theta_j} \right)$$

and $E\boldsymbol{G}_T\boldsymbol{G}'_T$ are nonsingular, the prime denoting transpose.

We shall suppose that the setting is such that k distinct estimating functions $G_{i,T}(\theta),\ 1 \le i \le k$, are available. Our first approach to the composition problem is to seek a quasi-score estimating function from within the set

$$\mathcal{K} = \left\{ \sum_{i=1}^{k} \boldsymbol{\alpha}_{i,T}(\boldsymbol{\theta})\, \boldsymbol{G}_{i,T}(\boldsymbol{\theta}) \right\}$$

where the weighting matrices $\boldsymbol{\alpha}_{i,T}$ are $p \times p$ constants. The discussion here follows Heyde (1989a).

Suppressing the dependence on T and $\boldsymbol{\theta}$ for convenience, we write

$$\boldsymbol{H} = \sum_{i=1}^{k} \boldsymbol{\alpha}_i\, \boldsymbol{G}_i = \boldsymbol{\alpha}'\, \boldsymbol{G}, \qquad \boldsymbol{H}^* = \sum_{i=1}^{k} \boldsymbol{\alpha}^*_i\, \boldsymbol{G}_i = \boldsymbol{\alpha}^{*'}\, \boldsymbol{G},$$

where

$$\boldsymbol{\alpha}' = \left(\boldsymbol{\alpha}_1 \,\vdots\, \dots \,\vdots\, \boldsymbol{\alpha}_k \right), \qquad \boldsymbol{\alpha}^{*'} = \left(\boldsymbol{\alpha}^*_1 \,\vdots\, \dots \,\vdots\, \boldsymbol{\alpha}^*_k \right),$$

$$\boldsymbol{G}' = \left(\boldsymbol{G}'_1 \,\vdots\, \dots \,\vdots\, \boldsymbol{G}'_k \right),$$

so that

$$E\dot{\boldsymbol{H}} = \sum_{i=1}^{k} \boldsymbol{\alpha}_i \, E\dot{\boldsymbol{G}}_i = \boldsymbol{\alpha}' \, E\dot{\boldsymbol{G}},$$

$$E\boldsymbol{H}\,\boldsymbol{H}^{*'} = \sum_{i=1}^{k} \sum_{j=1}^{k} \boldsymbol{\alpha}_i \, E\boldsymbol{G}_i \, \boldsymbol{G}_j' \, \boldsymbol{\alpha}_j^{*'} = \boldsymbol{\alpha} \, E\boldsymbol{G} \, \boldsymbol{G}' \, \boldsymbol{\alpha}^{*'}.$$

Then, from Theorem 2.1 we see that $\boldsymbol{H}^*$ is a quasi-score estimating function within $\mathcal{K}$ if

$$\boldsymbol{\alpha}^{*'} = (E\boldsymbol{G}\,\boldsymbol{G}')^{-1} \, E\dot{\boldsymbol{G}}.$$

This provides a general solution, although it may be less than straightforward to calculate $\boldsymbol{\alpha}^*$ in practice. Note in particular that since each individual $\boldsymbol{G}_i \in \mathcal{K}$, the formulation ensures that

$$\mathcal{E}(\boldsymbol{H}^*) - \mathcal{E}(\boldsymbol{G}_i)$$

is nnd, $1 \le i \le k$, so that composition is generally advantageous.

Various simplifications are sometimes possible. For example, if the $\boldsymbol{G}_i$ are standardized quasi-score (or true score) estimating functions, then we can suppose that each $\boldsymbol{G}_i$ satisfies

$$-E\dot{\boldsymbol{G}}_i = E\boldsymbol{G}_i \, \boldsymbol{G}_i', \quad 1 \le i \le k.$$

Circumstances under which equal weighting is obtained are especially important. Clearly this holds in the particular case when $E\boldsymbol{G}_i \, \boldsymbol{G}_j' = 0, \ i \ne j$, that is, the estimating functions are orthogonal. For an application to estimation in hidden-Markov random fields see Section 8.4.

If the estimating functions are not orthogonal, then we can adjust them using Gram-Schmidt orthogonalization to get a new set $\boldsymbol{K}_i, \ 1 \le i \le k$ of mutually orthogonal estimating functions. For example, in the case $k = 2$ we can replace $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ by $\boldsymbol{H}_1 = \boldsymbol{G}_1$ and

$$\boldsymbol{H}_2 = \boldsymbol{G}_2 - (E\boldsymbol{G}_2 \, \boldsymbol{G}_1')(E\boldsymbol{G}_1 \, \boldsymbol{G}_1')^{-1} \, \boldsymbol{G}_1$$

and $\boldsymbol{H}_1, \ \boldsymbol{H}_2$ are orthogonal.

There is also another approach to the problem of composition that is closely related but quite distinct, and this is treated in the next section. It is applicable only to the case of martingale estimating functions but it is nevertheless of use in a wide variety of contexts. As has been noted earlier, many models admit a semimartingale description, which naturally leads to martingale estimating functions and, furthermore, true score functions are martingales under mild conditions. It is principally spatial processes that are excluded.

6.3 Combining Martingale Estimating Functions

Section 6.2 deals with the general question of finding a quasi-score by combination of functions. However, as noted earlier in Section 2.5, it is often useful to adopt a martingale setting. For example, in many cases there is actually an underlying likelihood score $\boldsymbol{U}$, say, which is typically unknown. Then $\boldsymbol{U}$ will be a square integrable martingale under minor regularity conditions. This suggests the use of a martingale family of estimating functions as the most appropriate basis for approximation of $\boldsymbol{U}$. Indeed, if the family is large enough to contain $\boldsymbol{U}$, then the quasi-score estimating function on that space will be $\boldsymbol{U}$.

Again we let $(\boldsymbol{X}_t,\ 0 \leq t \leq T)$ be a sample from a process taking values in r-dimensional Euclidean space whose distribution depends on a parameter $\boldsymbol{\theta}$ belonging to an open subset of p-dimensional Euclidean space. The chosen setting is one of continuous time but the theory also applies to the discrete time case. This is dealt with by replacing a discrete time process $(\boldsymbol{X}_n,\ n \geq 0)$ by a continuous version $(\boldsymbol{X}_t^c,\ t \geq 0)$ defined through $\boldsymbol{X}_t^c = \boldsymbol{X}_n,\ n \leq t < n+1$. The discussion here follows Heyde (1987).

Suppose that the possible probability measures for $(\boldsymbol{X}_t)$ are $(P_{\boldsymbol{\theta}})$ and that each $(\Omega,\ \mathcal{F},\ P_{\boldsymbol{\theta}})$ is a complete probability space. The past-history σ-fields $(\mathcal{F}_t,\ t \geq 0)$ are assumed to be a standard filtration. That is, $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}$ for $s \leq t$, $\mathcal{F}_0$ being augmented by sets of measure zero of $\mathcal{F}$ and $\mathcal{F}_t = \mathcal{F}_{t+}$, where $\mathcal{F}_{t+} = \bigcap_{s>t} \mathcal{F}_s$.

Let $\mathcal{M}$ denote the class of square integrable estimating functions $(\boldsymbol{G}_T, \mathcal{F}_T)$ with $\boldsymbol{G}_T = \boldsymbol{G}_T\{(\boldsymbol{X}_t,\ 0 \leq t \leq T),\ \boldsymbol{\theta}\}$ that are martingales for each $P_{\boldsymbol{\theta}}$ and whose elements are almost surely differentiable with respect to the components of $\boldsymbol{\theta}$. Here $\boldsymbol{G}_T$ is a vector of dimension d not necessarily equal p.

Now for each martingale estimating function $\boldsymbol{G}_T \in \mathcal{M}$ there is an associated family of martingales $\int_0^T \boldsymbol{a}_s(\boldsymbol{\theta})\, d\boldsymbol{G}_s(\boldsymbol{\theta})$ where $\boldsymbol{a}_s(\boldsymbol{\theta})$ is a predictable matrix and the quasi-score estimating function from this family is

$$\int_0^T (d\bar{\dot{\boldsymbol{G}}}_s)'(d\langle \boldsymbol{G}\rangle_s)^-\, d\boldsymbol{G}_s.$$

As usual the prime denotes transpose and the minus generalized inverse. For $n \times 1$ vector valued martingales $\boldsymbol{M}_T$ and $\boldsymbol{N}_T$, the $n \times n$ process $\langle \boldsymbol{M}, \boldsymbol{N}'\rangle_T$, called the mutual quadratic characteristic, is the predictable increasing process such that $\boldsymbol{M}_T\, \boldsymbol{N}_T' - \langle \boldsymbol{M}, \boldsymbol{N}'\rangle_T$ is an $n \times n$ martingale. Also, we write $\langle \boldsymbol{M}\rangle_T$ for $\langle \boldsymbol{M}, \boldsymbol{M}'\rangle_T$, the quadratic characteristic of $\boldsymbol{M}_T$. Finally, $\dot{\boldsymbol{M}}_T$ is the $n \times p$ matrix obtained by differentiating the elements of $\boldsymbol{M}_T$ with respect to those of $\boldsymbol{\theta}$ and $d\bar{\dot{\boldsymbol{M}}}_t = E(d\dot{\boldsymbol{M}}_t \mid \mathcal{F}_{t-})$.

If we have two basic martingales $\boldsymbol{H}_T$ and $\boldsymbol{K}_T$, which belong to $\mathcal{M}$ and are such that one is not absolutely continuous with respect to the other, then each gives rise to a quasi-score estimating function, namely,

$$\int_0^T (d\bar{\dot{\boldsymbol{H}}}_t)'(d\langle \boldsymbol{H}\rangle_t)^-\, d\boldsymbol{H}_t$$

and

$$\int_0^T (d\bar{\dot{\boldsymbol{K}}}_t)'(d\langle \boldsymbol{K}\rangle_t)^-\, d\boldsymbol{K}_t,$$

respectively, and these may be regarded as competitors. If all the relevant quantities are known, a better estimating function may be obtained by combining them. We shall discuss the best procedure for combining the estimating functions and the gains that may be expected.

In the particular case where the X_i are independent random variables with finite mean $\mu_i(\boldsymbol{\theta})$ and variance $\sigma_i^2(\boldsymbol{\theta})$, natural martingales to consider for the estimation of $\boldsymbol{\theta}$ are

$$\sum_{i=1}^{n} \{X_i - \mu_i(\boldsymbol{\theta})\}$$

and

$$\sum_{i=1}^{n} \left\{(X_i - \mu_i(\boldsymbol{\theta}))^2 - \sigma_i^2(\boldsymbol{\theta})\right\}$$

and the optimal linear combination of these has been considered by Crowder (1987) and Firth (1987). Here we treat the problem of combination generally.

Given $\boldsymbol{H}_T$ and $\boldsymbol{K}_T$ we combine them into a new $2p \times 1$ vector martingale $\boldsymbol{J}_T = (\boldsymbol{H}_T', \boldsymbol{K}_T')'$ and the quasi-score estimating function for this is

$$\int_0^T (d\bar{\boldsymbol{J}}_t)'(d\langle \boldsymbol{J}\rangle_t)^- \, d\boldsymbol{J}_t. \tag{6.1}$$

Now

$$\langle \boldsymbol{J}\rangle_t = \begin{pmatrix} \langle \boldsymbol{H}\rangle_t & \langle \boldsymbol{H}, \boldsymbol{K}'\rangle_t \\ \langle \boldsymbol{H}, \boldsymbol{K}'\rangle_t' & \langle \boldsymbol{K}\rangle_t \end{pmatrix} \tag{6.2}$$

and if $\boldsymbol{A}$ and $\boldsymbol{D}$ are symmetric matrices and $\boldsymbol{A}$ is nonsingular,

$$\boldsymbol{M}^- = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}' & \boldsymbol{D} \end{pmatrix}^- = \begin{pmatrix} \boldsymbol{A}^{-1} + \boldsymbol{F}\,\boldsymbol{E}^-\boldsymbol{F}' & -\boldsymbol{F}\,\boldsymbol{E}^- \\ -\boldsymbol{E}^-\boldsymbol{F}' & \boldsymbol{E}^- \end{pmatrix} \tag{6.3}$$

where

$$\boldsymbol{E} = (\boldsymbol{D} - \boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B}), \quad \boldsymbol{F} = \boldsymbol{A}^{-1}\boldsymbol{B},$$

and $\boldsymbol{M}$ is nonsingular if $\boldsymbol{A}$ and $\boldsymbol{E}$ are nonsingular. Also, if $\boldsymbol{E}$ is nonsingular

$$(\boldsymbol{A}^{-1} + \boldsymbol{F}\,\boldsymbol{E}^{-1}\boldsymbol{F}')^{-1} = \boldsymbol{A} - \boldsymbol{B}\,\boldsymbol{D}^{-1}\boldsymbol{B}'. \tag{6.4}$$

The result (6.3) and its supplement follows from Exercise 2.4, p. 32 and a minor modification of Exercise 2.7, p. 33 of Rao (1973) while (6.4) comes from Exercise 2.9, p. 33 of the same reference.

Suppose that $d\langle \boldsymbol{H}\rangle_t$ and $d\langle \boldsymbol{K}\rangle_t$ are nonsingular. Then, using (6.2) – (6.4) and after some algebra, we find that (6.1) can be expressed as

$$\int_0^T \left\{(d\bar{\boldsymbol{H}}_t)'(d\langle \boldsymbol{H}\rangle_t)^{-1}(\boldsymbol{I} - \boldsymbol{R}_t\boldsymbol{S}_t)^{-1} - (d\bar{\boldsymbol{K}}_t)'(d\langle \boldsymbol{K}\rangle_t)^{-1}(\boldsymbol{I} - \boldsymbol{S}_t\boldsymbol{R}_t)^{-1}\boldsymbol{S}_t\right\} d\boldsymbol{H}_t$$

$$+\int_0^T \left\{(d\bar{\boldsymbol{K}}_t)'(d\langle \boldsymbol{K}\rangle_t)^{-1}(\boldsymbol{I} - \boldsymbol{S}_t\boldsymbol{R}_t)^{-1} - (d\bar{\boldsymbol{H}}_t)'(d\langle \boldsymbol{H}\rangle_t)^{-1}(\boldsymbol{I} - \boldsymbol{R}_t\boldsymbol{S}_t)^{-1}\boldsymbol{R}_t\right\} d\boldsymbol{K}_t \tag{6.5}$$

where

$$\begin{aligned} \boldsymbol{R}_t &= (d\langle \boldsymbol{H}, \boldsymbol{K}' \rangle_t)(d\langle \boldsymbol{K} \rangle_t)^{-1}, \\ \boldsymbol{S}_t &= (d\langle \boldsymbol{H}, \boldsymbol{K}' \rangle_t)'(d\langle \boldsymbol{H} \rangle_t)^{-1} = (d\langle \boldsymbol{K}, \boldsymbol{H}' \rangle_t)(d\langle \boldsymbol{H} \rangle_t)^{-1}. \end{aligned}$$

Note that if $\langle \boldsymbol{H}, \boldsymbol{K}' \rangle_t \equiv 0$, the quasi-score estimating function (6.5) becomes

$$\int_0^T (d\bar{\boldsymbol{H}}_t)'(d\langle \boldsymbol{H} \rangle_t)^{-1}\, d\boldsymbol{H}_t + \int_0^T (d\bar{\boldsymbol{K}}_t)'(d\langle \boldsymbol{K} \rangle_t)^{-1}\, d\boldsymbol{K}_t,$$

the sum of the quasi-score estimating functions for $\boldsymbol{H}$ and $\boldsymbol{K}$.

The quasi-score estimating function (6.5) can be conceived as arising from the martingales $\boldsymbol{H}$ and $\boldsymbol{K}$ in the following way. Based on $\boldsymbol{H}$, $\boldsymbol{K}$ we can consider the class of estimating functions

$$\int_0^T \boldsymbol{a}_t\, d\boldsymbol{H}_t + \int_0^T \boldsymbol{b}_t\, d\boldsymbol{K}_t$$

where $\boldsymbol{a}_t$, $\boldsymbol{b}_t$ are predictable matrices. Then, the corresponding quasi-score estimating function is

$$\int_0^T (\boldsymbol{a}_t\, d\bar{\boldsymbol{H}}_t + \boldsymbol{b}_t\, d\bar{\boldsymbol{K}}_t)'(d\langle \boldsymbol{a}\,\boldsymbol{H} + \boldsymbol{b}\,\boldsymbol{K} \rangle_t)^{-1}(\boldsymbol{a}_t\, d\boldsymbol{H}_t + \boldsymbol{b}_t\, d\boldsymbol{K}_t)$$

and the best of these, in the sense of minimum size asymptotic confidence intervals, is obtained by choosing for $\boldsymbol{a}_t$, $\boldsymbol{b}_t$ the values $\boldsymbol{a}_t^*$, $\boldsymbol{b}_t^*$, where $\boldsymbol{a}_t^*$, $\boldsymbol{b}_t^*$ are such that

$$\int_0^T (\boldsymbol{a}_t\, d\bar{\boldsymbol{H}}_t + \boldsymbol{b}_t\, d\bar{\boldsymbol{K}}_t)'(d\langle \boldsymbol{a}\,\boldsymbol{H} + \boldsymbol{b}\,\boldsymbol{K} \rangle_t)^{-1}(\boldsymbol{a}_t\, d\bar{\boldsymbol{H}}_t + \boldsymbol{b}_t\, d\bar{\boldsymbol{K}}_t)$$

is maximized. This result is given by (6.5).

The formula (6.5) can also be easily obtained through Gram-Schmidt orthogonalization. Note that for the martingales $\boldsymbol{G}_{1T} = \boldsymbol{H}_T$ and

$$\boldsymbol{G}_{2T} = \boldsymbol{K}_T - \int_0^T d\langle \boldsymbol{H}, \boldsymbol{K}' \rangle_t (d\langle \boldsymbol{H} \rangle_t)^{-1}\, d\boldsymbol{H}_t$$

we have $\langle \boldsymbol{G}_1, \boldsymbol{G}_2' \rangle \equiv 0$ so that, upon standardizing the estimating functions, we have as a combined quasi-score estimating function

$$\int_0^T (d\bar{\boldsymbol{H}}_t)'(d\langle \boldsymbol{H} \rangle_t)^{-1}\, d\boldsymbol{H}_t + \int_0^T (d\bar{\boldsymbol{G}}_{2t})'(d\langle \boldsymbol{G}_2 \rangle_t)^{-1}\, d\boldsymbol{G}_{2t},$$

which, after some algebra, reduces to (6.5).

In order to compare martingale estimating functions we shall use the martingale information introduced in Section 2.5. For $\boldsymbol{G} \in \mathcal{M}$ we define the martingale information $\boldsymbol{I}_G$ as

$$\boldsymbol{I}_G = \bar{\boldsymbol{G}}_T'(\langle \boldsymbol{G} \rangle_T)^{-1} \bar{\boldsymbol{G}}_T \tag{6.6}$$

when $\langle \boldsymbol{G} \rangle_T$ is nonsingular. Note that in the one-dimensional case where $\boldsymbol{G}_T = \boldsymbol{U}_T$, the score function of $(\boldsymbol{X}_t,\ 0 \le t \le T)$, $\boldsymbol{I}_G$ is the conditional Fisher information. Also, the quantity $\boldsymbol{I}_G$ occurs as a scale variable in the asymptotic distribution of the estimator $\boldsymbol{\theta}^*$ obtained from the estimating equation $\boldsymbol{G}_T(\boldsymbol{\theta}^*) = \mathbf{0}$.

The quantities $\boldsymbol{I}_G$ for competing martingales can be compared in the partial order of nonnegative definite matrices. Note that the quasi-score estimating function based on $\boldsymbol{G}$ is

$$Q\,S(G) = \int_0^T (d\bar{\boldsymbol{G}}_t)'(d\langle \boldsymbol{G} \rangle_t)^{-1} d\boldsymbol{G}_t$$

and

$$\boldsymbol{I}_{QS(G)} = \int_0^T (d\bar{\boldsymbol{G}}_t)'(d\langle \boldsymbol{G} \rangle_t)^{-1} d\bar{\boldsymbol{G}}_t. \tag{6.7}$$

In many cases there is one “natural” martingale suggested by the context such as through a semimartingale model representation as is described in Section 2.6. For example, this is certainly the case if $(\boldsymbol{X}_t)$ is representable in the form

$$\boldsymbol{X}_t = \int_0^t \boldsymbol{f}_s(\boldsymbol{\theta})\, d\lambda_s + \boldsymbol{m}_t(\boldsymbol{\theta}) \tag{6.8}$$

where (λ_t) is a real, monotone increasing, right continuous process with $\lambda_0 = 0$, $(\boldsymbol{m}_t(\boldsymbol{\theta}), \mathcal{F}_t) \in \mathcal{M}$ and $\{\boldsymbol{f}_t(\boldsymbol{\theta})\}$ is predictable. This framework has been extensively discussed in Hutton and Nelson (1986) and Godambe and Heyde (1987).

Various other martingales can easily be constructed from a basic martingale $(\boldsymbol{m}_t(\boldsymbol{\theta}), \mathcal{F}_t)$ to use in conjuction with it. The simplest general ones are, for $d = 1$,

$$\int_0^t (dm_s(\boldsymbol{\theta}))^2 - \langle m(\boldsymbol{\theta}) \rangle_t \tag{6.9}$$

(discrete time) and

$$m_t^2(\boldsymbol{\theta}) - \langle m(\boldsymbol{\theta}) \rangle_t \left(= 2 \int_0^t m_{s-}(\boldsymbol{\theta})\, dm_s(\boldsymbol{\theta}) \right), \tag{6.10}$$

the last result following from Ito’s formula. Generally if $H_n(x, y)$ is the Hermite-Chebyshev polynomial in x and y defined by the generating function

$$\exp\left(t\,x - \frac{1}{2}\,t^2 y \right) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\, H_n(x, y),$$

then, for each n, $H_n(m_t, \langle m \rangle_t)$ is a martingale. See for example, Chung and Williams (1983, Theorem 6.4, p. 114).

6.3.1 An Example

We consider the first order autoregressive process

$$X_i = \theta\, X_{i-1} + \varepsilon_i,$$

where the $\varepsilon_i = \varepsilon_i(\theta)$ are independent and identically distributed with $E\varepsilon_i = 0$, $E\varepsilon_i^2 = \sigma^2$, $E\varepsilon_i^4 < \infty$. Write $\sigma^{-3}E\varepsilon_i^3 = \gamma$ and $\sigma^{-4}E\varepsilon^4 - 3 = \kappa$ for the skewness and kurtosis, respectively, of the distribution of ε. We wish to estimate θ on the basis of sample $(X_i,\ 0 \le i \le T)$.

Now the natural martingale for the process (X_i), which of course has a representation of the form (6.8), is

$$H_j = \sum_{i=1}^{j}(X_i - \theta\, X_{i-1}), \quad j = 1, 2, \dots$$

and the associated martingale given by (6.9) is

$$K_j = \sum_{i=1}^{j}(X_i - \theta\, X_{i-1})^2 - j\sigma^2, \quad j = 1, 2, \dots .$$

Put $H_j = \sum_{i=1}^{j} h_i$, $K_j = \sum_{i=1}^{j} k_i$. The combined quasi-score estimating function given by (6.5) is then

$$\frac{1}{\sigma^3(\kappa+2)}\left(1 - \frac{\gamma^2}{\kappa+2}\right)^{-1}\sum_{i=1}^{T}\left\{\sigma(2\dot{\sigma}\gamma - (\kappa+2)X_{i-1})h_i + (X_{i-1}\gamma - 2\dot{\sigma})k_i\right\}, \tag{6.11}$$

where $\dot{\sigma} = d\sigma/d\theta$.

Note that the quasi-score estimating functions based separately on H and K are

$$-\frac{1}{\sigma^2}\sum_{i=1}^{T} X_{i-1}h_i, \qquad -\frac{2\dot{\sigma}}{(\kappa+2)\sigma^3}\sum_{i=1}^{T} k_i,$$

respectively. Thus, if $\dot{\sigma} = 0$, the K martingale will contribute to the estimation of θ in combination with H if $\gamma \neq 0$ even though it is useless in its own right.

If the ε_i are normally distributed, then $\gamma = 0$, $\kappa = 0$ and (6.11) reduces to minus twice the score function. That is, quasi-likelihood and maximum likelihood estimation are the same.

To compare the various estimators we calculate their corresponding martingale informations. We find, after some algebra, that

$$\begin{aligned} I_{QS(H)} &= \sigma^{-2}\sum_{i=1}^{T} X_{i-1}^2, \\ I_{QS(K)} &= \frac{4T(\dot{\sigma})^2}{(\kappa+2)\sigma^2} \end{aligned}$$

and, writing $Q\,S(H,K)$ for the combined quasi-score estimating function (6.11),

$$\begin{aligned} I_{QS(H,K)} &= \frac{1}{\sigma^2(\kappa+2)}\left(1-\frac{\gamma^2}{\kappa+2}\right)^{-1} \\ &\quad \cdot \sum_{i=1}^{T}\Big\{((\kappa+2)X_{i-1}-2\sigma\gamma)X_{i-1}+2\dot{\sigma}(2\dot{\sigma}-X_{i-1}\gamma)\Big\} \\ &= \left(1-\frac{\gamma^2}{\kappa+2}\right)^{-1}\left\{I_{QS(H)}+I_{QS(K)}-\frac{4\dot{\sigma}\gamma}{\sigma^2(\kappa+2)}\sum_{i=1}^{T}X_{i-1}\right\}. \end{aligned}$$

In the case where (X_i) is stationary ($|\theta|<1$) we have

$$\begin{aligned} I_{QS(H)} &\sim T\sigma^{-2}EX_1^2 = T(1-\theta^2)^{-1} \quad \text{a.s.} \quad \text{and} \\ I_{QS(H,K)} &\sim \left(1-\frac{\gamma^2}{\kappa+2}\right)^{-1}\left\{I_{QS(H)}+I_{QS(K)}\right\} \quad \text{a.s.} \end{aligned}$$

as $T\to\infty$. If, on the other hand, $|\theta|\geq 1$, then

$$\begin{aligned} I_{QS(K)} &= o(I_{QS(H)}) \quad \text{a.s.} \quad \text{and} \\ I_{QS(H,K)} &\sim \left(1-\frac{\gamma^2}{\kappa+2}\right)^{-1} I_{QS(H)} \quad \text{a.s.} \end{aligned}$$

as $T\to\infty$. Note that even in this latter case combining K with H is advantageous if $\gamma\neq 0$.

6.4 Application. Nested Strata of Variation

Suppose we have a nested model in which $e_{(R)}$ represents the error term for stratum R. Let $\boldsymbol{e}$ denote all the $e_{(R)}$ written in lexicographic order as a column vector. Then, the obvious approach to quasi-likelihood estimation based on the family of estimating function $\{\boldsymbol{C}\,\boldsymbol{e},\ \boldsymbol{C}\text{ constant matrix}\}$ is not practicable. Because of the nesting, the $\boldsymbol{e}$'s are correlated across strata making $E\boldsymbol{e}\,\boldsymbol{e}'$ difficult to invert.

The problem can be avoided, however, by working with appropriate residuals. Define, within each stratum R, residuals

$$\boldsymbol{r}_{(R)} = (\boldsymbol{I}-\boldsymbol{W}_{(R)})\,\boldsymbol{e}_{(R)} \tag{6.12}$$

where $\boldsymbol{W}_{(R)}$ is idempotent so that $\boldsymbol{W}_{(R)}\boldsymbol{r}_{(R)}=\boldsymbol{0}$. Then, if $\boldsymbol{W}_{(R)}$ is also chosen so that

$$\boldsymbol{V}_{(R)}\,\boldsymbol{W}'_{(R)} = \boldsymbol{W}_{(R)}\,\boldsymbol{V}_{(R)} \tag{6.13}$$

where $\boldsymbol{V}_{(R)} = E\boldsymbol{e}_{(R)}\,\boldsymbol{e}'_{(R)}$, we have that a quasi-score estimating function (QSEF) based on the set of estimating functions $\{\boldsymbol{C}\,\boldsymbol{r}_{(R)},\ \boldsymbol{C}\,\text{constant matrix}\}$ is

$$\left(E\dot{\boldsymbol{e}}_{(R)}\right)'\left(E\boldsymbol{e}_{(R)}\,\boldsymbol{e}'_{(R)}\right)^{-1}\boldsymbol{r}_{(R)}. \tag{6.14}$$

The result (6.14) follows because (dropping the suffix (R) for convenience) the QSEF is, from Theorem 2.1,

$$(E\dot{\boldsymbol{r}})'\,(E\boldsymbol{r}\,\boldsymbol{r}')^{-}\,\boldsymbol{r} = (E\dot{\boldsymbol{e}})'\,(\boldsymbol{I}-\boldsymbol{W})'\,((\boldsymbol{I}-\boldsymbol{W})\,\boldsymbol{V}\,(\boldsymbol{I}-\boldsymbol{W})')^{-}\,\boldsymbol{r}, \tag{6.15}$$

the minus denoting generalized inverse. Further, in view of (6.13),

$$\begin{aligned}(\boldsymbol{I}-\boldsymbol{W})\,\boldsymbol{V}\,(\boldsymbol{I}-\boldsymbol{W})' &= \boldsymbol{V}-\boldsymbol{W}\,\boldsymbol{V}-\boldsymbol{V}\,\boldsymbol{W}'+\boldsymbol{W}\,\boldsymbol{V}\,\boldsymbol{W}' \\ &= (\boldsymbol{I}-\boldsymbol{W})\,\boldsymbol{V}\end{aligned}$$

since $\boldsymbol{W}$ is idempotent and it is easily checked that

$$((\boldsymbol{I}-\boldsymbol{W})\,\boldsymbol{V})^{-} = \boldsymbol{V}^{-1}\,(\boldsymbol{I}-\boldsymbol{W}). \tag{6.16}$$

Also from (6.13),

$$(\boldsymbol{I}-\boldsymbol{W})'\,\boldsymbol{V}^{-1} = \boldsymbol{V}^{-1}\,(\boldsymbol{I}-\boldsymbol{W}), \tag{6.17}$$

so that using (6.16), (6.17) and $(\boldsymbol{I}-\boldsymbol{W})\,\boldsymbol{r}=\boldsymbol{r}$, the right hand side of (6.15) reduces to

$$(E\dot{\boldsymbol{e}})'\,\boldsymbol{V}^{-1}\,\boldsymbol{r}$$

as required.

Now it is necessary to combine the QSEF from each of the strata and for this it is desirable that the residuals be orthogonal across strata. That is, we need

$$E\boldsymbol{r}_{(R)}\,\boldsymbol{r}'_{(S)} = \boldsymbol{0} \qquad \forall\, R,\, S.$$

These requirements are not difficult to achieve in practice and then the combined QSEF is the sum over strata

$$\sum_{R}\,(E\boldsymbol{e}_{(R)}\,\boldsymbol{e}'_{(R)})^{-1}\,(E\dot{\boldsymbol{e}}_{(R)})'\,\boldsymbol{r}_{(R)}. \tag{6.18}$$

As a concrete example we shall consider a model introduced by Morton (1987) to deal with several nested strata of extra Poisson variation in a generalized linear model where multiplicative errors are associated with each stratum. The motivation for this modeling came from a consideration of insect trap catches, details of which are given in the cited paper.

Suppose that there are three nested strata labeled 2, 1, 0 with respective subscripts i, j, k having arbitrary ranges. Stratum 0, the bottom stratum, has scaled Poisson errors; stratum 1 introduces errors Z_{ij} and the top stratum 2 has further errors Z_i. It is assumed that the Z's are mutually independent with

$$EZ_{ij}=1, \quad EZ_i=1, \quad \text{var } Z_{ij}=\sigma_1^2, \quad \text{var } Z_i=\sigma_2^2$$

and the model can be written as

$$X_{ijk} = \mu_{ijk}\, Z_{ij}\, Z_i + e_{ijk}$$

with the e_{ijk} uncorrelated and

$$E\left(e_{ijk}\,\middle|\, Z_{ij},\, Z_i\right) = 0, \quad \text{var}\left(e_{ijk}\,\middle|\, Z_{ij},\, Z_i\right) = \phi\, \mu_{ijk}\, Z_{ij}\, Z_i$$

where ϕ is an unknown scale parameter.

From these assumptions it is straightforward to calculate the covariance matrix $\boldsymbol{V}$ for the data vector $\boldsymbol{x} = (X_{ijk})'$, the elements being written in a row in lexicographic order. Define $\psi_1 = \sigma_1^2(\sigma_2^2 + 1)/\phi$ and $\psi_2 = \sigma_2^2/\phi$. Then, the elements of $\boldsymbol{V}$ are

$$\begin{aligned}
\text{var}\; X_{ijk} &= \phi\, \mu_{ijk}\{1 + (\psi_1 + \psi_2)\, \mu_{ijk}\},\\
\text{cov}\;(X_{ijk}, X_{ijn}) &= \phi\, (\psi_1 + \psi_2)\, \mu_{ijk}\, \mu_{ijn} \quad (k \neq n),\\
\text{cov}\;(X_{ijk}, X_{imn}) &= \phi\, \psi_2\, \mu_{ijk}\, \mu_{imn} \quad (j \neq m),\\
\text{cov}\;(X_{ijk}, X_{lmn}) &= 0 \quad (i \neq l).
\end{aligned}$$

If $\boldsymbol{\mu} = (\mu_{ijk})'$ depends on a vector $\boldsymbol{\theta}$ of unknown parameters, then the quasi-score estimating function can be written as $(\partial\boldsymbol{\mu}/\partial\boldsymbol{\theta})'\, \boldsymbol{V}^{-1}\, (\boldsymbol{x} - \boldsymbol{\mu})$. However, this is typically not a tractable form to use because of its dimension and the complexity of $\boldsymbol{V}$.

We shall use the theory developed above to obtain a useful version of the quasi-score estimating function by focusing on the sources of variation in each of the strata separately and optimally combining the resulting estimating functions. This contrasts with the approach of Morton (1987) where ideas based on partitioning of a sum of squares are used (in the spirit of the Wedderburn approach to quasi-likelihood).

We write for the error terms in the three strata,

$$\begin{aligned}
e_{ijk} &= X_{ijk} - \mu_{ijk}\, Z_{ij}\, Z_i,\\
e_{ij} &= X_{ij\cdot} - \mu_{ij\cdot}\, Z_i,\\
e_i &= X_{i\cdot\cdot} - \mu_{i\cdot\cdot},
\end{aligned}$$

where the weighted totals are

$$\begin{aligned}
X_{ij\cdot} &= \sum_k X_{ijk}, \quad \mu_{ij\cdot} = \sum_k \mu_{ijk}, \quad e_{ij\cdot} = \sum_k e_{ijk},\\
X_{i\cdot\cdot} &= \sum_j w_{ij}\, X_{ij\cdot}, \quad \mu_{i\cdot\cdot} = \sum_j w_{ij}\, \mu_{ij\cdot}, \quad e_{i\cdot\cdot} = \sum_j w_{ij}\, e_{ij\cdot},\\
e_{i\cdot} &= \sum_j w_{ij}\, e_{ij},
\end{aligned}$$

with $w_{ij} = \phi\,\mu_{ij\cdot}\,/\mathrm{var}\ e_i = 1/(1+\psi_1\,\mu_{ij\cdot})$, and the residuals are

$$\begin{aligned}
r_{ijk} &= X_{ijk} - \frac{\mu_{ijk}}{\mu_{ij\cdot}}\,X_{ij\cdot} = e_{ijk} - \frac{\mu_{ijk}}{\mu_{ij\cdot}}\,e_{ij\cdot}, \\
r_{ij} &= X_{ij\cdot} - \frac{\mu_{ij\cdot}}{\mu_{i\cdot\cdot}}\,X_{i\cdot\cdot} = e_{ij} - \frac{\mu_{ij\cdot}}{\mu_{i\cdot\cdot}}\,e_{i\cdot}, \\
r_i &= X_{i\cdot\cdot} - \mu_{i\cdot\cdot} = e_{i\cdot\cdot}
\end{aligned}$$

We find that

$$\mathrm{var}\ e_i = \phi\,\mu_{i\cdot\cdot}\,(1+\psi_2\,\mu_{i\cdot\cdot}), \quad \mathrm{var}\ e_{ij} = \phi\,\mu_{ij\cdot}\,(1+\psi_1\,\mu_{ij\cdot})$$

and it is easily checked that the conditions (6.12) and (6.13) are satisfied for the three strata.

To check orthogonality across strata we note that

$$\begin{aligned}
\mathrm{cov}\ (r_{ijk},\, X_{ij\cdot}) &= \mathrm{cov}\ (X_{ijk},\, X_{ij\cdot}) - \frac{\mu_{ijk}}{\mu_{ij\cdot}}\ \mathrm{var}\ X_{ij\cdot} \\
&= \phi\,\mu_{ijk}\,\mu_{ij\cdot}\,(\psi_1+\psi_2) + \phi\,\mu_{ijk} \\
&\quad - \frac{\mu_{ijk}}{\mu_{ij\cdot}}\,\phi\{\mu_{ij\cdot}^2(\psi_1+\psi_2) + \mu_{ij\cdot}\} \\
&= 0
\end{aligned}$$

and hence cov $(r_{ijk}, r_{ij}) = 0$, cov $(r_{ijk}, r_i) = 0$. Also,

$$\begin{aligned}
\mathrm{cov}\ (r_{ij},\, X_{i\cdot\cdot}) &= \mathrm{cov}\ (X_{ij\cdot},\, X_{i\cdot\cdot}) - \frac{\mu_{ij\cdot}}{\mu_{i\cdot\cdot}}\ \mathrm{var}\ X_{i\cdot\cdot} \\
&= \phi\,\psi_2\,\mu_{ij\cdot}\,\mu_{i\cdot\cdot} + \phi\,\mu_{ij\cdot} \\
&\quad - \frac{\mu_{ij\cdot}}{\mu_{i\cdot\cdot}}\,(\phi\,\psi_2\,\mu_{i\cdot\cdot}^2 + \phi\,\mu_{i\cdot\cdot}\} \\
&= 0
\end{aligned}$$

and hence cov $(r_{ij}, r_i) = 0$. All other cross strata covariances of residuals are easily seen to be zero.

The combined QSEF is then given by (6.18) as

$$\sum_{ijk} \frac{\partial\mu_{ijk}}{\partial\theta}\,\frac{r_{ijk}}{\mu_{ijk}} + \sum_{ij} \frac{\partial\mu_{ij\cdot}}{\partial\theta}\,\frac{r_{ij}}{\mu_{ij\cdot}\,(1+\psi_1\,\mu_{ij\cdot})} + \sum_{i} \frac{\partial\mu_{i\cdot\cdot}}{\partial\theta}\,\frac{r_i}{\mu_{i\cdot\cdot}\,(1+\psi_2\,\mu_{i\cdot\cdot})}.$$

For another example of a linear model with multiplicative random effects on which quasi-likelihood methods have been used, see Firth and Harris (1991). Their approach is similar to that of Morton (1987).

6.5 State-Estimation in Time Series

An important class of time series problems involve state space models of which a simple example is the type

$$F_{1t}(Y_t, Y_{t-1}, X_t) = \epsilon_t, \qquad F_{2t}(X_t, X_{t-1}) = \delta_t, \tag{6.19}$$

where the Y's are observed but not the X's, the F's are known functional forms and the $\{\epsilon_t\}$ and $\{\delta_t\}$ are each uncorrelated sequences of random variables that are mutually orthogonal. In *filtering* we require the estimation of X_T given the observations $(Y_0, \dots, Y_T)$ and in *smoothing* the estimation of $(X_0, \dots, X_T)$ given $(Y_0, \dots, Y_T)$. Other parameters in the system are generally assumed known in the context of filtering or smoothing. The joint estimation of the parameters and the X's is known as system identification.

In this section we shall indicate the estimating function approach to filtering and smoothing and in particular show how the celebrated Kalman filter follows from optimal combination of estimating functions. The material follows the ideas of Naik-Nimbalkar and Rajarshi (1995). A closely related approach, based on an extension of the E-M algorithm for missing data, is given in Chapter 7.4.4.

To focus the discussion we shall specialize (6.19) to the linear case

$$Y_t = \alpha_t\, X_t + \epsilon_t, \qquad X_t - \mu = \beta_t(X_{t-1} - \mu) + \delta_t, \tag{6.20}$$

where μ, α's, β's are known constants and if $\mathcal{G}_t$ and $\mathcal{H}_t$ are the σ-fields generated by $X_t, \dots, X_0, Y_t, \dots, Y_0$ and $X_t, \dots, X_0, Y_{t-1}, \dots, Y_0$, respectively, then almost surely,

$$E\left(\epsilon_t \,\middle|\, \mathcal{H}_t\right) = 0, \qquad E\left(\delta_t \,\middle|\, \mathcal{G}_{t-1}\right) = 0,$$

$$E\left(\epsilon_s\, \epsilon_t \,\middle|\, \mathcal{H}_t\right) = 0 = E\left(\epsilon_s\, \delta_t \,\middle|\, \mathcal{H}_t\right), \quad s \neq t.$$

Now the model (6.20) suggests estimating functions

$$G_{1T} = Y_T - \alpha_T\, X_T, \qquad G_{2T} = X_T - X_{T|T-1},$$

where $X_{T|T-1}$ is the minimum variance unbiased estimator of X_T based on $Y_{T-1}, \dots, Y_0$. Further, G_{1T} and G_{2T} are uncorrelated since

$$E G_{1T}\, G_{2T} = E\left(G_{2T}\, E\left(\epsilon_T \,\middle|\, \mathcal{H}_T\right)\right) = 0,$$

G_{2T} being $\mathcal{H}_T$-measurable. Then, according to Section 6.2 we obtain a composite quasi-score by standardizing the estimating functions G_{1T}, G_{2T} and adding. This leads to the estimating equation

$$\frac{Y_T - \alpha_T\, X_T}{E\epsilon_T^2}\,(-\alpha_T) + \frac{X_T - X_{T|T-1}}{E(X_T - X_{T|T-1})^2} = 0 \tag{6.21}$$

for X_T. The solution of (6.21) gives Zehnwirth's (1988) extended *Kalman filter*, the classical filter corresponding to the case where $E(\epsilon_T^2 \,|\, \mathcal{H}_T)$ is nonrandom.

We now consider the case of smoothing. Here we can conveniently use the estimating functions

$$\begin{aligned} H_{1T} &= \sum_{t=1}^{T} c_t (Y_t - \alpha_t\, X_t) \\ H_{2T} &= \sum_{t=1}^{T} d_t (X_t - \mu - \beta_t (X_{t-1} - \mu)) \end{aligned}$$

where c_t is $\mathcal{H}_{t-1}$-measurable and d_t is $\mathcal{G}_{t-1}$-measurable. Then, using the ideas of Section 6.3 and noting that H_{1T}, H_{2T} are orthogonal, the quasi-score estimating equations for $X_0, \ldots, X_T$ are

$$\begin{aligned} &\frac{Y_t - \alpha_t\, X_t}{E\left(\epsilon_t^2 \,\middle|\, \mathcal{H}_t\right)} (-\alpha_t) + \frac{X_t - \mu - \beta_t (X_{t-1} - \mu)}{E\left(\delta_t^2 \,\middle|\, \mathcal{G}_{t-1}\right)} \\ &\qquad + \frac{X_{t+1} - \mu - \beta_{t+1} (X_t - \mu)}{E\left(\delta_{t+1}^2 \,\middle|\, \mathcal{G}_t\right)} (-\beta_{t+1}) = 0, \quad t = 1, 2, \ldots, T-1, \\ &\frac{Y_T - \alpha_T\, X_T}{E\left(\epsilon_T^2 \,\middle|\, \mathcal{H}_T\right)} (-\alpha_T) + \frac{X_T - \mu - \beta_T (X_{T-1} - \mu)}{E\left(\delta_T^2 \,\middle|\, \mathcal{G}_{T-1}\right)} = 0. \end{aligned}$$

Further examples and references can be found in Naik-Nimbalkar and Rajarshi (1995).

6.6 Exercises

1. (*Generalized linear model*). Take $\boldsymbol{X} = (X_1, \ldots, X_n)'$ as a sample of independent random variables, $\{\mathcal{F}\}$ a class of distributions and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ a vector parameter such that

$$E_F\, X_i = \mu_i(\boldsymbol{\theta}(F)), \qquad \text{var}\ X_i = \phi(F)\, V_i(\boldsymbol{\theta}(F))$$

for all $F \in \mathcal{F}$ where μ_i, V_i, $i = 1, 2, \ldots, n$ are *specified* real functions of the indicated variables. Suppose that the skewness (γ) and kurtosis (κ) are known:

$$\gamma = E_F (X_i - \mu_i)^3 / (\phi V_i)^{3/2}, \qquad \kappa = E_F (X_i - \mu_i)^4 / (\phi V_i)^2 - 3$$

for all $F \in \mathcal{F}$ and $i = 1, 2, \ldots, n$. Let

$$h_{1i} = X_i - \mu, \qquad h_{2i} = (X_i - \mu_i)^2 - \phi V_i - \gamma (\phi V_i)^{1/2} (X_i - \mu_i)$$

and write $\boldsymbol{h}_r = (h_{r1}, \ldots, h_{rn})'$, $r = 1, 2$. Show that the quasi-score estimating function for the family

$$\mathcal{G} = \left\{ \boldsymbol{\alpha}' \boldsymbol{h}_1 + \boldsymbol{\beta}' \boldsymbol{h}_2; \quad \boldsymbol{\alpha}, \boldsymbol{\beta}\ \text{nonrandom}\, n \times p\, \text{matrices} \right\}$$

is $\boldsymbol{\alpha}^{*'}\boldsymbol{h}_1 + \boldsymbol{\beta}^{*'}\boldsymbol{h}_2$ where

$$\begin{aligned}\boldsymbol{\alpha}^* &= \left(-(\phi V_i)^{-1}(\partial\mu_i/\partial\theta_j)\right),\\ \boldsymbol{\beta}^* &= \left(\frac{(\phi V_i)^{1/2}(\partial\mu_i/\partial\theta_j)+\phi(\partial V_i/\partial\theta_j)}{(\phi V_i)^2(\kappa+2-\gamma^2)}\right)\end{aligned}$$

(Crowder (1987), Firth (1987), Godambe and Thompson (1989)).

2. Consider the random effects model

$$X_{i,j} = \mu + a_i + e_{ij} \quad (i = 1,2,\dots,n,\ j = 1,2,\dots,n_i),$$

where a_i and e_{ij} are normally distributed with zero means and variances σ_a^2 and σ_i^2, respectively. Take $\mu = 0$ for convenience. Use the space of estimating functions

$$\mathcal{H} = \left\{\sum_{i=1}^{n} \boldsymbol{A}_i\,\boldsymbol{Y}_i\right\}$$

where the $\boldsymbol{A}_i$ are $(n+1)\times(n_i + n_i(n_i-1)/2)$ nonrandom matrices and the $\boldsymbol{Y}_i$ are the $n_i + n_i(n_i-1)/2$ vectors

$$\begin{aligned}\boldsymbol{Y}_i &= (X_{i1}^2 - EX_{i1}^2,\dots,X_{in_i}^2 - EX_{in_i}^2, X_{i1}X_{i2} - EX_{i1}X_{i2},\dots,\\ &\qquad X_{i1}X_{in_1} - EX_{i1}X_{in_1},\dots,X_{in_i-1}X_{in_i} - E(X_{in_i-1}X_{in_i}))'\end{aligned}$$

to find a quasi-score estimating function for estimation of σ_a^2 and σ_i^2, $i = 1,2,\dots,n$ (Lin (1996)).

Chapter 7

Projected Quasi-Likelihood

7.1 Introduction

There is a considerable emphasis in theoretical statistics on conditional inference. Principles such as that inference about a parameter should be conditioned on an ancillary statistic have received much attention. This discussion requires full specification of distributions and is consequently not directly available in quasi-likelihood context. However, conducting inference conditionally on a statistic T is equivalent to projecting onto the subspace orthogonal to that generated by T and such projections can conveniently be carried out for spaces of estimating functions. Indeed, projection based methods seem to be the natural vehicle for dealing with conditioning in an estimating function context, and it has recently become clear that quasi-likelihood extensions of much conditional inference can be obtained via projection.

It is worthwhile to trace the development of this subject from conditional likelihoods to conditional score functions to projected quasi-likelihoods. The principles of conditioning date back to Fisher but with extensive subsequent developments; see Cox and Hinkley (1974, Chapter 2) for a discussion and references. The first step to an estimating function environment was taken by Godambe (1976) who showed that the conditional score function is an optimal estimating function for a parameter of interest when the conditioning statistic is complete and sufficient for the nuisance parameters. This work was generalized by Lindsay (1982) to deal with partial likelihood factorizations and cases where the conditioning statistic may depend on the parameter of interest. The next steps in the evolution, from a likelihood to a quasi-likelihood environment, are now emerging and elements of this development are the subject of this chapter.

Conditioning problems may involve either the parameter to be estimated or the data that are observed. We shall illustrate by dealing with parameters subject to constraint, nuisance parameters, and missing data (for which we provide a P-S (Project-Solve) generalization of the E-M (Expectation-Maximization) algorithm).

7.2 Constrained Parameter Estimation

Suppose that the parameter $\boldsymbol{\theta}$ is subject to the linear constraint $\boldsymbol{F}'\boldsymbol{\theta} = \boldsymbol{d}$, where $\boldsymbol{F}$ is a $p \times q$ matrix not depending on the data or $\boldsymbol{\theta}$. For nonlinear or inequality constraints the extension of the theory is indicated briefly in Section 7.2.3. The overall discussion follows Heyde and Morton (1993).

For an arbitrary positive-definite symmetric matrix $\boldsymbol{V}$, we define the projection matrix

$$\boldsymbol{P}_V = \boldsymbol{F}\,(\boldsymbol{F}'\,\boldsymbol{V}^{-1}\boldsymbol{F})^{-}\,\boldsymbol{F}'\,\boldsymbol{V}^{-1},$$

where $\boldsymbol{A}^{-}$ denotes a generalized inverse of $\boldsymbol{A}$. Necessarily (Rao (1973, p. 26)) $\boldsymbol{P}_V\boldsymbol{F} = \boldsymbol{F}$, and (Rao (1973, p. 47)) $\boldsymbol{P}_V$ is unique.

In what follows, $\boldsymbol{V}$ is independent of the data but may depend on $\boldsymbol{\theta}$, although such dependence will be suppressed in the notation. An important property of these projections is that, for any $\boldsymbol{V}$ and $\boldsymbol{W}$,

$$(\boldsymbol{I} - \boldsymbol{P}_W)(\boldsymbol{I} - \boldsymbol{P}_V) = \boldsymbol{I} - \boldsymbol{P}_W. \tag{7.1}$$

We shall consider three methods of estimation of $\boldsymbol{\theta}$ subject to constraint. For Methods 2 and 3 we use standardization of the type employed in Chapter 2 which ensures that the likelihood score property $E\boldsymbol{G}\,\boldsymbol{G}' = -E\dot{\boldsymbol{G}}$ holds.

Method 1: Projection of free estimator. Given any estimating function $\boldsymbol{G} \in \mathcal{G}$, let $\hat{\boldsymbol{\theta}}$ solve $\boldsymbol{G}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}$. For arbitrary $\boldsymbol{V}$ the projected estimator is

$$\tilde{\boldsymbol{\theta}} = (\boldsymbol{I} - \boldsymbol{P}_V)'\,\hat{\boldsymbol{\theta}} + \boldsymbol{V}^{-1}\boldsymbol{F}\,(\boldsymbol{F}'\,\boldsymbol{V}^{-1}\,\boldsymbol{F})^{-}\boldsymbol{d}. \tag{7.2}$$

Provided that $\boldsymbol{d}$ is consistent in the sense that

$$\boldsymbol{F}'\,\boldsymbol{V}^{-1}\boldsymbol{F}\,(\boldsymbol{F}'\,\boldsymbol{V}^{-1}\boldsymbol{F})^{-}\boldsymbol{d} = \boldsymbol{d},$$

then the constraint $\boldsymbol{F}'\,\tilde{\boldsymbol{\theta}} = \boldsymbol{d}$ is ensured.

Method 2: Projection of standardized functions. For any standardized estimating function $\boldsymbol{G} \in \mathcal{G}$ let $\tilde{\boldsymbol{\theta}}$ solve the equations

$$(\boldsymbol{I} - \boldsymbol{P}_V)\,\boldsymbol{G}(\tilde{\boldsymbol{\theta}}) = \boldsymbol{0}, \qquad \boldsymbol{F}'\,\tilde{\boldsymbol{\theta}} = \boldsymbol{d}. \tag{7.3}$$

Note that from (7.1) the choice of $\boldsymbol{V}$ is immaterial, for if we multiply (7.3) by $\boldsymbol{I} - \boldsymbol{P}_W$, we get $(\boldsymbol{I} - \boldsymbol{P}_W)\,\boldsymbol{G}(\tilde{\boldsymbol{\theta}}) = \boldsymbol{0}$. In practice we may like to choose $\boldsymbol{V} = \boldsymbol{I}$.

Method 3: Analogue of Lagrange multipliers. If we had a quasi-likelihood q, then by the method of Lagrange multipliers we would maximize $q + \boldsymbol{\lambda}'(\boldsymbol{F}'\,\boldsymbol{\theta} - \boldsymbol{d})$, where $\boldsymbol{\lambda}$ is determined by the constraint. This suggests the following method. Let $\boldsymbol{G} \in \mathcal{G}$ be any standardized estimating function. We solve for $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\lambda}$ the equations

$$\boldsymbol{G}(\tilde{\boldsymbol{\theta}}) + \boldsymbol{F}\boldsymbol{\lambda} = \boldsymbol{0}, \qquad \boldsymbol{F}'\,\tilde{\boldsymbol{\theta}} = \boldsymbol{d}. \tag{7.4}$$

This method is available even when $\boldsymbol{G}$ is not the derivative of a function q or even a quasi-score. Define $\boldsymbol{V}_0 = \mathcal{E}(\boldsymbol{G})$ and $\boldsymbol{P}_0$ to be projector with $\boldsymbol{V} = \boldsymbol{V}_0$. The information criterion is modified to

$$\mathcal{E}_F(\boldsymbol{G}) = (\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{V}_0(\boldsymbol{I} - \boldsymbol{P}_0)'.$$

Its symmetric generalized inverse is

$$\mathcal{E}_F(\boldsymbol{G})^- = (\boldsymbol{I} - \boldsymbol{P}_0)'\boldsymbol{V}_0^{-1}(\boldsymbol{I} - \boldsymbol{P}_0), \tag{7.5}$$

which is shown below to be the asymptotic variance for the projected estimator with the optimum choice of $\boldsymbol{V} = \boldsymbol{V}_0$. Our optimality criterion will be to minimize (7.5). With this choice, it is seen that the three methods are asymptotically equivalent and that optimality occurs within $\mathcal{H}$ when $\boldsymbol{G}$ is a quasi-score estimating function, standardized for the purpose of (7.3) and (7.4).

7.2.1 Main Results

Theorem 7.1 Let $\tilde{\boldsymbol{\theta}}$ be the projected estimator (7.2). Suppose that $\hat{\boldsymbol{\theta}}$ has asymptotic variance $\mathcal{E}(\boldsymbol{G})^{-1}$. Then $\tilde{\boldsymbol{\theta}}$ has asymptotic variance

$$(\boldsymbol{I} - \boldsymbol{P}_V)'\,\mathcal{E}(\boldsymbol{G})^{-1}(\boldsymbol{I} - \boldsymbol{P}_V) \geq \mathcal{E}_F(\boldsymbol{G})^-;$$

equality holds when $\boldsymbol{V} = \mathcal{E}(\boldsymbol{G})$.

Proof. The formula for the asymptotic variance $\tilde{\boldsymbol{\theta}}$ is obvious from the definition of $\tilde{\boldsymbol{\theta}}$. To prove the inequality, we use (7.1) and the fact that $\boldsymbol{V}_0^{-1}\boldsymbol{P}_0 = \boldsymbol{P}_0' V_0^{-1}$ and $\boldsymbol{P}_0^2 = \boldsymbol{P}_0$. The difference is

$$\begin{aligned}
&(\boldsymbol{I} - \boldsymbol{P}_V)'\,\boldsymbol{V}_0^{-1}(\boldsymbol{I} - \boldsymbol{P}_V) - (\boldsymbol{I} - \boldsymbol{P}_0)'\,V_0^{-1}(\boldsymbol{I} - \boldsymbol{P}_0) \\
&\quad = (\boldsymbol{I} - \boldsymbol{P}_V)'\,\{\boldsymbol{V}_0^{-1} - (\boldsymbol{I} - \boldsymbol{P}_0)'\,\boldsymbol{V}_0^{-1}(\boldsymbol{I} - \boldsymbol{P}_0)\}(\boldsymbol{I} - \boldsymbol{P}_V) \\
&\quad = (\boldsymbol{I} - \boldsymbol{P}_V)'\,\{\boldsymbol{V}_0^{-1} - \boldsymbol{V}_0^{-1}(\boldsymbol{I} - \boldsymbol{P}_0)\}(\boldsymbol{I} - \boldsymbol{P}_V) \\
&\quad = (\boldsymbol{I} - \boldsymbol{P}_V)'\,\boldsymbol{V}_0^{-1}\boldsymbol{P}_0(\boldsymbol{I} - \boldsymbol{P}_V) \\
&\quad = (\boldsymbol{I} - \boldsymbol{P}_V)'\,\boldsymbol{P}_0'\,\boldsymbol{V}_0^{-1}\boldsymbol{P}_0(\boldsymbol{I} - \boldsymbol{P}_V) \\
&\quad = (\boldsymbol{P}_0 - \boldsymbol{P}_V)'\,\boldsymbol{V}_0^{-1}(\boldsymbol{P}_0 - \boldsymbol{P}_V) \geq \boldsymbol{0},
\end{aligned}$$

since $\boldsymbol{V}_0^{-1}$ is positive definite.

Theorem 7.2 Let $\boldsymbol{G} \in \mathcal{G}$ be a standardized estimating function. Assume that with probability tending to 1 all methods (7.2)–(7.4) possess unique solutions for $\tilde{\boldsymbol{\theta}}$ in some neighborhood of $\boldsymbol{\theta}$. Then in this neighborhood, when solutions are unique:

(a) under mild regularity conditions the projected estimator (7.2) with $\boldsymbol{V} = \boldsymbol{V}_0$ and the projected function estimator (7.3) agree to first order;

(b) the projected function estimator (7.3) and the Lagrange analogue estimator (7.4) have identical solutions.

Proof. (a) Expand $\boldsymbol{G}(\hat{\boldsymbol{\theta}})$ about $\boldsymbol{\theta}$ to first order and assume that $\dot{\boldsymbol{G}}$ approximates $-\boldsymbol{V}_0$. Then

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \simeq -\dot{\boldsymbol{G}}(\boldsymbol{\theta})^{-1}\boldsymbol{G}(\boldsymbol{\theta}) \simeq \boldsymbol{V}_0^{-1}\boldsymbol{G}(\boldsymbol{\theta}).$$

Hence

$$\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta} \simeq (\boldsymbol{I} - \boldsymbol{P}_0)'\,\boldsymbol{V}_0^{-1}\boldsymbol{G}(\boldsymbol{\theta})$$

and

$$\begin{aligned} \boldsymbol{G}(\tilde{\boldsymbol{\theta}}) &\simeq \boldsymbol{G}(\boldsymbol{\theta}) + \dot{\boldsymbol{G}}(\boldsymbol{\theta})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \simeq \boldsymbol{G}(\boldsymbol{\theta}) - \boldsymbol{V}_0(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\ &\simeq \{\boldsymbol{I} - \boldsymbol{V}_0(\boldsymbol{I} - \boldsymbol{P}_0)'\,\boldsymbol{V}_0^{-1}\}\,\boldsymbol{G}(\boldsymbol{\theta}) = \boldsymbol{P}_0\,\boldsymbol{G}(\boldsymbol{\theta}). \end{aligned}$$

Thus $(\boldsymbol{I} - \boldsymbol{P}_0)\,\boldsymbol{G}(\tilde{\boldsymbol{\theta}}) \simeq \boldsymbol{0}$, so (7.3) is satisfied to first order. By uniqueness, the two estimators must agree to first order.

(b) Given that a solution to (7.3) exists, it also solves (7.4) with

$$\boldsymbol{\lambda} = -(\boldsymbol{F}'\,\boldsymbol{V}^{-1}\boldsymbol{F})^{-}\boldsymbol{F}'\,\boldsymbol{V}^{-1}\boldsymbol{G}.$$

By uniqueness the two solutions must be identical.

Corollary 7.1 Assume that $\boldsymbol{G}(\boldsymbol{\theta})$ is asymptotically normally distributed under variance based norming. Then, under mild regularity conditions, and for large samples, $\tilde{\boldsymbol{\theta}}$ is approximately normally distributed with mean $\boldsymbol{\theta}$ and singular variance matrix $\mathcal{E}_F(\boldsymbol{G})^{-}$, whichever method is used. Also, $(\boldsymbol{I} - \boldsymbol{P}_0)\,\boldsymbol{G}(\boldsymbol{\theta})$ is approximately normally distributed with mean zero and singular variance matrix $\mathcal{E}_F(\boldsymbol{G})$.

This Corollary may be used to construct approximate confidence regions for $\boldsymbol{\theta}$ (see Chapter 4). We would prefer to base them on the projected estimating function, since the asymptotic normality of $\boldsymbol{G}$ rather than $\hat{\boldsymbol{\theta}}$ is usually the origin of the theory, so that the further approximation due to the expansion of $\boldsymbol{G}$ is avoided. Then the confidence region for $\boldsymbol{\theta}$ is of the form

$$\boldsymbol{G}'\,\mathcal{E}_F(\boldsymbol{G})^{-}\boldsymbol{G} \le c, \quad \boldsymbol{F}'\,\boldsymbol{\theta} = \boldsymbol{d},$$

where c is a scalar obtained from the appropriate chi-squared or F distribution.

Theorem 7.3 Let $\boldsymbol{Q}$ be a quasi-score estimating function within $\mathcal{H}$. Then $\boldsymbol{Q}$ is optimal in that $\mathcal{E}_F(\boldsymbol{G})^{-} \ge \mathcal{E}_F(\boldsymbol{Q})^{-}$ for all $\boldsymbol{G} \in \mathcal{H}$.

Proof. Since $\mathcal{E}(\boldsymbol{Q}) \ge \mathcal{E}(\boldsymbol{G})$, $\mathcal{E}(\boldsymbol{G})^{-1} \ge \mathcal{E}(\boldsymbol{Q})^{-1}$. Hence

$$(\boldsymbol{I} - \boldsymbol{P}_V)'\,\{\mathcal{E}(\boldsymbol{G})^{-1} - \mathcal{E}(\boldsymbol{Q})^{-1}\}\,(\boldsymbol{I} - \boldsymbol{P}_V) \ge 0.$$

Upon inserting $\boldsymbol{V} = \mathcal{E}(\boldsymbol{G})$, we get that

$$\mathcal{E}_F(\boldsymbol{G})^{-} \ge (\boldsymbol{I} - \boldsymbol{P}_V)'\,\mathcal{E}(\boldsymbol{Q})^{-1}(\boldsymbol{I} - \boldsymbol{P}_V) \ge \mathcal{E}_F(\boldsymbol{Q})^{-}$$

by Theorem 7.1.

Naturally, the asymptotic properties of the corollary hold for $\boldsymbol{Q}$.

7.2.2 Examples

Example 1: Linear regression. Consider the usual regression model

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad E(\boldsymbol{\epsilon}) = \boldsymbol{0}, \quad E(\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}') = \sigma^2\,\boldsymbol{I},$$

where $\boldsymbol{X}$ is the design matrix with full rank and $\boldsymbol{F}'\,\boldsymbol{\theta} = \boldsymbol{d}$. In this case the quasi-score estimating function $\boldsymbol{Q} = \boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\,\boldsymbol{\theta})$ gives the usual normal equation $\boldsymbol{Q} = \boldsymbol{0}$. The optimal choice for $\boldsymbol{V}$ is $\boldsymbol{V}_0 = \sigma^2\,\boldsymbol{X}'\,\boldsymbol{X}$, so

$$\boldsymbol{P}_0 = \boldsymbol{F}\,\{\boldsymbol{F}'\,(\boldsymbol{X}'\,\boldsymbol{X})^{-1}\,\boldsymbol{F}\}^{-}\,\boldsymbol{F}'\,(\boldsymbol{X}'\,\boldsymbol{X})^{-1}.$$

Standardization of $\boldsymbol{Q}$ gives $\boldsymbol{Q}^{(s)} = \sigma^{-2}\boldsymbol{Q}$, which would be the likelihood score if $\boldsymbol{\epsilon}$ were normally distributed. The free estimator is

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\,\boldsymbol{X})^{-1}\boldsymbol{X}'\,\boldsymbol{Y},$$

so its projection is

$$\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - (\boldsymbol{X}'\,\boldsymbol{X})^{-1}\boldsymbol{F}\,\{\boldsymbol{F}'\,(\boldsymbol{X}'\,\boldsymbol{X})^{-1}\boldsymbol{F}\}^{-}\,(\boldsymbol{F}'\,\hat{\boldsymbol{\theta}} - \boldsymbol{d}),$$

which is identical to that obtained by projecting $\boldsymbol{Q}^{(s)}$ or using the Lagrange multiplier method. The result has been long known: see Judge and Takayama (1966), who used Lagrange multipliers with least squares.

Example 2: Two Poisson processes, constrained total. Let $N_i = \{N_i(t),\, 0 \le t \le T\}$ be independent Poisson processes with constant rates $\theta_i\,(i = 1, 2)$, which are constrained so that $\theta_1 + \theta_2 = 1$. We have that $\mathcal{H}$ is the class of estimating functions of the form

$$\left[\int_0^T b_1(s)\{dN_1(s) - \theta_1\,ds\}, \int_0^T b_2(s)\{dN_2(s) - \theta_2\,ds\}\right],$$

with b_1, b_2 being predictable. An application of Theorem 2.1 can be used to find the quasi-score estimating function

$$\boldsymbol{Q} = \left\{\theta_1^{-1}N_1(T) - T, \theta_2^{-1}N_2(T) - T\right\}'.$$

This is already standardized and in fact is the likelihood score. Here $\boldsymbol{F} = (1, 1)'$, $\boldsymbol{V}_0 = \operatorname{diag}(\theta_1^{-1}, \theta_2^{-1})$, so that

$$\boldsymbol{P}_0 = (\theta_1 + \theta_2)^{-1}\begin{pmatrix} \theta_1 & \theta_2 \\ \theta_1 & \theta_2 \end{pmatrix}.$$

Using (7.3) we obtain the estimator

$$\tilde{\theta}_i = N_i(T)/\{N_1(T) + N_2(T)\} \quad (i = 1, 2),$$

which is identical to the solution of (7.2) and (7.4).

For an example in this form a likelihood and Lagrange multiplier approach is available, but this could easily have been precluded by the use of counting processes, which were not specified to be Poisson and for which only the first and second moment properties were assumed.

Example 3: Linear combination of estimating functions. Suppose that we have $n > p$ estimating functions $\boldsymbol{H} = \{h_i(\boldsymbol{X}, \boldsymbol{\theta})\}$ such that $E(\boldsymbol{H}) = \boldsymbol{0}$ and $E(\boldsymbol{H}\,\boldsymbol{H}') = \boldsymbol{V}_H$ exist. We wish to reduce the dimension of $\boldsymbol{H}$ to p by choosing optimal linear combinations. That is, $\mathcal{H}$ consists of estimating functions of the form $\boldsymbol{G} = \boldsymbol{C}\,\boldsymbol{H}$, where $\boldsymbol{C}$ is a $p \times n$ matrix depending only on $\boldsymbol{\theta}$. The standardized quasi-score estimating function is

$$\boldsymbol{Q}^{(s)} = -E(\dot{\boldsymbol{H}})'\,\boldsymbol{V}_H^{-1}\boldsymbol{H}. \tag{7.6}$$

Suppose now that we wish to reduce the dimensions of $\boldsymbol{H}$ in the presence of linear constraints $\boldsymbol{F}'\,\boldsymbol{\theta} = \boldsymbol{d}$. By (7.4), the theory is modified to solve for $\tilde{\boldsymbol{\theta}}$, $\boldsymbol{\lambda}$ from

$$E(\dot{\boldsymbol{H}})'\,\boldsymbol{V}_H^{-1}\boldsymbol{H} - \boldsymbol{F}\,\boldsymbol{\lambda} = 0, \quad \boldsymbol{F}'\,\tilde{\boldsymbol{\theta}} = \boldsymbol{d}.$$

The theory extends to the case where $\boldsymbol{C}$ may contain incidential parameters (Morton (1981a)). An important application concerns functional and structural relationships where the elements of $\boldsymbol{H}$ are the contrasts that are free of the incidential parameters (Morton (1981b)).

7.2.3 Discussion

We have presented the theory assuming that the constraints are linear in order to avoid obscuring the simple arguments by messy details. If we had nonlinear constraints, say $\boldsymbol{\phi}(\boldsymbol{\theta}) = \boldsymbol{c}$, then the Lagrange multiplier analogue replaces $\boldsymbol{F}'$ by $\boldsymbol{F}(\boldsymbol{\theta})' = \partial\boldsymbol{\phi}/\partial\boldsymbol{\theta}'$, which we assume exists and has full rank. If the projection methods are used, then this substitution is also made and in (7.2) we must also approximate $\boldsymbol{d}(\boldsymbol{\theta}) = \boldsymbol{F}(\boldsymbol{\theta}_0)'\,\boldsymbol{\theta}$ iteratively. In (7.3) and (7.4), $\boldsymbol{F}'\,\boldsymbol{\theta} = \boldsymbol{d}$ is replaced by $\boldsymbol{\phi}(\boldsymbol{\theta}) = \boldsymbol{c}$.

Inequality constraints could in principle be handled by Lagrange multipliers. If the jth constraint is $(\boldsymbol{F}'\,\boldsymbol{\theta})_j \le d_j$, the corresponding multiplier λ_j would be set equal to zero if strict inequality holds and would be free if equality holds.

In many problems it would be natural to transform the parameter $\boldsymbol{\theta}$ to $(\boldsymbol{\phi}', \boldsymbol{\psi}')'$, where $\boldsymbol{\phi}$ has $r = \text{rank}(\boldsymbol{F})$ elements, which define the constraints, and $\boldsymbol{\psi}$ is free. We then need to reduce $\boldsymbol{G}$ to dimensions $p - r$. Picking any $p - r$ elements of $\boldsymbol{G}$ would not be optimal; the best linear reduction would be achieved in the manner of (7.6), which would be equivalent to projecting $\boldsymbol{G}$ as in (7.3). Algebraic elimination of $\boldsymbol{\psi}$ from $\boldsymbol{G}$ would be equivalent to the Lagrange multiplier method. For nonlinear problems this could lead to an estimating function outside the class $\mathcal{G}$. In the case where $\mathcal{G}$ is the class of functions that are linear in $\boldsymbol{\theta}$, all three methods agree with the parameter transformation method.

In the linear regression model (Example 1), suppose that $\boldsymbol{\psi} = \boldsymbol{F}'\boldsymbol{\theta}$ is not fixed but is a nuisance parameter. It can be shown that, in the terminology of Chapter 3 that $\boldsymbol{Q}_1 = (\boldsymbol{I} - \boldsymbol{P}_0)\,\boldsymbol{Q}$ is locally E-ancillary for $\boldsymbol{\psi}$ and if $\boldsymbol{Q}_2 = \boldsymbol{P}_0\,\boldsymbol{Q}$, then $E\boldsymbol{Q}_1\,\boldsymbol{Q}_2' = \boldsymbol{0}$ and hence $\boldsymbol{Q}_2$ is locally E-sufficient for $\boldsymbol{\psi}$.

In the general situation, assuming that linear approximation is adequate, the decomposition

$$\boldsymbol{Q} = \boldsymbol{P}_0\,\boldsymbol{Q} + (\boldsymbol{I} - \boldsymbol{P}_0)\,\boldsymbol{Q}$$

has the first term locally E-sufficient for $\boldsymbol{\psi}$ and the second term locally E-ancillary for $\boldsymbol{\psi}$. Thus, the method of projection of standardized functions (7.3) discards the first order sufficient information about $\boldsymbol{\psi}$ and replaces it by the known constraint $\boldsymbol{\psi} = \boldsymbol{d}$.

7.3 Quasi-Likelihood in the Presence of Nuisance Parameters

In this section we use methods analogous to those of the last section to treat the case where the basic parameter $\boldsymbol{\theta}$ contains a (vector) nuisance component.

Suppose that the parameter $\boldsymbol{\theta}$ is partitioned with

$$\boldsymbol{\theta}' = (\boldsymbol{\phi}', \boldsymbol{\psi}'),$$

$\boldsymbol{\phi}$ being the component of interest and $\boldsymbol{\psi}$ a nuisance parameter and that $\boldsymbol{G}$ is a standardized estimating function chosen for the estimation of $\boldsymbol{\theta}$. Standardization means that the likelihood score property

$$E\dot{\boldsymbol{G}} = E(\partial \boldsymbol{G}/\partial \boldsymbol{\theta}) = -E\boldsymbol{G}\,\boldsymbol{G}'$$

holds. We write

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{G}_\phi \\ \boldsymbol{G}_\psi \end{pmatrix}$$

and partition the information matrix

$$\boldsymbol{V}_G = E\boldsymbol{G}\,\boldsymbol{G}' = \mathcal{E}(\boldsymbol{G})$$

according to

$$\boldsymbol{V}_G = \begin{pmatrix} \boldsymbol{V}_{\phi\phi} & \boldsymbol{V}_{\phi\psi} \\ \boldsymbol{V}_{\psi\phi} & \boldsymbol{V}_{\psi\psi} \end{pmatrix}$$

where

$$\boldsymbol{V}_{\phi\phi} = E\boldsymbol{G}_\phi\,\boldsymbol{G}_\phi', \quad \boldsymbol{V}_{\psi\psi} = E\boldsymbol{G}_\psi\,\boldsymbol{G}_\psi', \quad \boldsymbol{V}_{\psi\phi} = \boldsymbol{V}_{\phi\psi}' = E(\boldsymbol{G}_\psi\,\boldsymbol{G}_\phi').$$

Write also

$$\boldsymbol{F}_\psi = \begin{pmatrix} \boldsymbol{V}_{\phi\psi} \\ \boldsymbol{V}_{\psi\psi} \end{pmatrix} = -E(\partial \boldsymbol{G}/\partial \boldsymbol{\psi}),$$

$$\boldsymbol{F}_\phi = \begin{pmatrix} \boldsymbol{V}_{\phi\phi} \\ \boldsymbol{V}_{\psi\phi} \end{pmatrix} = -E(\partial \boldsymbol{G}/\partial \boldsymbol{\phi}).$$

It turns out that the projection

$$\boldsymbol{P}_\psi = \boldsymbol{F}_\psi (\boldsymbol{F}'_\psi \, \boldsymbol{V}_G^{-1} \, \boldsymbol{F}_\psi)^{-1} \, \boldsymbol{F}'_\psi \, \boldsymbol{V}_G^{-1}$$

identifies the information about $\boldsymbol{\psi}$ for $\boldsymbol{\phi}$ given and the estimating equation

$$(\boldsymbol{I} - \boldsymbol{P}_\psi)\, \boldsymbol{G} = \boldsymbol{0} \tag{7.7}$$

is optimal for estimation of $\boldsymbol{\phi}$ in the presence of the nuisance parameter $\boldsymbol{\psi}$. In (7.7) the sensitive dependence of $\boldsymbol{G}$ on $\boldsymbol{\psi}$ has been removed in the sense that

$$E\left[\frac{\partial}{\partial \boldsymbol{\psi}} \, \{(\boldsymbol{I} - \boldsymbol{P}_\psi)\, \boldsymbol{G}\}\right] = \boldsymbol{0}$$

since $\boldsymbol{P}_\psi \, \boldsymbol{F}_\psi = \boldsymbol{F}_\psi$ and $\boldsymbol{G}$ has zero mean. This is, of course, a first order approach and $(\boldsymbol{I} - \boldsymbol{P}_\psi)\boldsymbol{G}$ will not in general be free of $\boldsymbol{\psi}$. There is a convenient interpretation in terms of E-ancillary and E-sufficient estimating functions as discussed in Chapter 3. It can be seen that $(\boldsymbol{I} - \boldsymbol{P}_\psi)\, \boldsymbol{G}$ is locally E-ancillary for $\boldsymbol{\psi}$ and $\boldsymbol{P}_\psi \, \boldsymbol{G}$ is locally E-sufficient for $\boldsymbol{\psi}$.

To appreciate (7.7) in more detail, first note that, in partitioned matrix form,

$$\boldsymbol{F}'_\psi \, \boldsymbol{V}_G^{-1} = \begin{pmatrix} 0 & \boldsymbol{I} \end{pmatrix},$$

so that

$$\begin{aligned} \boldsymbol{P}_\psi &= \boldsymbol{F}_\psi \, \boldsymbol{V}_{\psi\psi}^{-1} \begin{pmatrix} 0 & \boldsymbol{I} \end{pmatrix} \\ &= \begin{pmatrix} 0 & \boldsymbol{V}_{\phi\psi} \, \boldsymbol{V}_{\psi\psi}^{-1} \\ 0 & \boldsymbol{I} \end{pmatrix} \end{aligned}$$

and

$$(\boldsymbol{I} - \boldsymbol{P}_\psi)\, \boldsymbol{G} = \begin{pmatrix} \boldsymbol{G}_\phi - \boldsymbol{V}_{\phi\psi} \, \boldsymbol{V}_{\psi\psi}^{-1} \, \boldsymbol{G}_\psi \\ \boldsymbol{0} \end{pmatrix}.$$

The "information" concerning $\boldsymbol{\theta}$ associated with the estimation procedure is

$$E\left(\boldsymbol{G}_\phi - \boldsymbol{V}_{\phi\psi} \, \boldsymbol{V}_{\psi\psi}^{-1} \, \boldsymbol{G}_\psi\right)\left(\boldsymbol{G}_\phi - \boldsymbol{V}_{\phi\psi} \, \boldsymbol{V}_{\psi\psi}^{-1} \, \boldsymbol{G}_\psi\right)' = \boldsymbol{V}_{\phi\phi} - \boldsymbol{V}_{\phi\psi} \, \boldsymbol{V}_{\psi\psi}^{-1} \, \boldsymbol{V}_{\psi\phi}, \tag{7.8}$$

or alternatively we can think of it in partitioned matrix form as

$$\begin{aligned} \mathcal{E}_\phi(\boldsymbol{G}) &= (\boldsymbol{I} - \boldsymbol{P}_\psi)\, \boldsymbol{V}_G \, (\boldsymbol{I} - \boldsymbol{P}_\psi)' \\ &= \begin{pmatrix} \boldsymbol{V}_{\phi\phi} - \boldsymbol{V}_{\phi\psi} \, \boldsymbol{V}_{\psi\psi}^{-1} \, \boldsymbol{V}_{\psi\phi} \\ \boldsymbol{0} \end{pmatrix}. \end{aligned}$$

Of course there is no loss in efficiency if $\boldsymbol{G}_\phi$, $\boldsymbol{G}_\psi$ are orthogonal, for then $\boldsymbol{V}_{\phi\psi} = \boldsymbol{V}_{\psi\phi} = \boldsymbol{0}$. The formula (7.8) is a direct generalization of the familiar result obtained from partitioning the likelihood score (e.g., Bhaphkar (1989)) and the asymptotic variance of the estimator of $\boldsymbol{\phi}$ based on $\boldsymbol{G}$ is then

$$\left(\boldsymbol{V}_{\phi\phi} - \boldsymbol{V}_{\phi\psi}\, \boldsymbol{V}_{\psi\psi}^{-1}\, \boldsymbol{V}_{\psi\phi}\right)^{-1} = \boldsymbol{V}^{\phi\phi}, \tag{7.9}$$

where

$$\boldsymbol{V}_G^{-1} = \begin{pmatrix} \boldsymbol{V}^{\phi\phi} & \boldsymbol{V}^{\phi\psi} \\ \boldsymbol{V}^{\psi\phi} & \boldsymbol{V}^{\psi\psi} \end{pmatrix}.$$

Optimality of estimation for $\boldsymbol{\phi}$ thus focuses on maximizing $\boldsymbol{V}^{\phi\phi}$ in the partial order of nnd matrices and we have the following theorem.

Theorem 7.4 Let $\boldsymbol{Q}$ be a quasi-score estimating function within $\mathcal{H}$. Then $\boldsymbol{Q}$ is optimal for estimation of $\boldsymbol{\phi}$ in the sense that

$$\boldsymbol{V}_G^{\phi\phi} \geq \boldsymbol{V}_Q^{\phi\phi}$$

in the partial order for all $G \in \mathcal{H}$.

Proof. The result follows immediately from the fact that

$$\mathcal{E}(\boldsymbol{Q}) = \boldsymbol{V}_Q \geq \mathcal{E}(\boldsymbol{G}) = \boldsymbol{V}_G$$

for all $\boldsymbol{G} \in \mathcal{H}$ and then

$$\boldsymbol{V}_G^{-1} \geq \boldsymbol{V}_Q^{-1}.$$

As a simple example of the methodology we shall discuss the linear model

$$\boldsymbol{Z} = \boldsymbol{X}\,\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

say, where $\boldsymbol{Z}$ is a vector of dimension T, $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ is $T \times p$ matrix with $\boldsymbol{X}_1$ of dimension $T \times r$ and $\boldsymbol{X}_2$ of dimension $T \times (p-r)$. Also, $\boldsymbol{\theta} = (\boldsymbol{\phi}', \boldsymbol{\psi}')'$ where $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are, respectively, vectors of dimension r and $p-r$ and $\boldsymbol{\epsilon}$ is a vector of dimension T with zero mean and covariance matrix Σ.

In this case the standardized estimating function is

$$\begin{aligned} \boldsymbol{G} &= \boldsymbol{X}'\,\boldsymbol{\Sigma}^{-1}\,(\boldsymbol{Z} - \boldsymbol{X}_1\,\boldsymbol{\phi} - \boldsymbol{X}_2\,\boldsymbol{\psi}) \\ &= \begin{pmatrix} \boldsymbol{X}_1'\,\boldsymbol{\Sigma}^{-1}\,(\boldsymbol{Z} - \boldsymbol{X}_1\,\boldsymbol{\phi} - \boldsymbol{X}_2\,\boldsymbol{\psi}) \\ \boldsymbol{X}_2'\,\boldsymbol{\Sigma}^{-1}\,(\boldsymbol{Z} - \boldsymbol{X}_1\,\boldsymbol{\phi} - \boldsymbol{X}_2\,\boldsymbol{\psi}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{G}_\phi \\ \boldsymbol{G}_\psi \end{pmatrix}. \end{aligned}$$

Also,

$$\begin{aligned} \boldsymbol{V}_G = E\boldsymbol{G}\,\boldsymbol{G}' &= \boldsymbol{X}'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{X} \\ &= \begin{pmatrix} \boldsymbol{X}_1'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{X}_1 & \boldsymbol{X}_1'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{X}_1 & \boldsymbol{X}_2'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{X}_2 \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{V}_{\phi\phi} & \boldsymbol{V}_{\phi\psi} \\ \boldsymbol{V}_{\psi\phi} & \boldsymbol{V}_{\psi\psi} \end{pmatrix}. \end{aligned}$$

Then, the estimating function

$$\boldsymbol{G}_{\phi} - \boldsymbol{V}_{\phi\psi}\,\boldsymbol{V}_{\psi\psi}^{-1}\,\boldsymbol{G}_{\psi}$$

$$= \left[\boldsymbol{X}_1'\,\boldsymbol{\Sigma}^{-1} - (\boldsymbol{X}_1'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{X}_2)(\boldsymbol{X}_2'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\,\boldsymbol{\Sigma}^{-1}\right](\boldsymbol{Z} - \boldsymbol{X}_1\,\boldsymbol{\phi})$$

does not involve the nuisance parameter $\boldsymbol{\psi}$ and the estimator for $\boldsymbol{\phi}$ is obtained from equating this to zero.

This solution can, of course, be obtained directly from solving the ordinary estimating equation

$$\begin{bmatrix} \boldsymbol{G}_{\phi} \\ \boldsymbol{G}_{\psi} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}$$

and then eliminating $\boldsymbol{\psi}$ from the resultant equations. Indeed, many nuisance parameter problems are amenable to such a direct approach.

The focus in this section has been on projection but there is no need for this geometric interpretation to be emphasized. The first order theory seeks the best combination $\boldsymbol{G}_{\phi} - \boldsymbol{c}(\phi,\psi)\boldsymbol{G}_{\psi}$, which is easy to calculate analytically. Also, it is not difficult to go beyond the first order theory. Second order theory allows, incorporation, for example, of the result of Godambe and Thompson (1974) who dealt with the case in which the likelihood score is $U_{\theta} = (U_{\phi}, U_{\psi})'$, ϕ and ψ being scalars, and showed that an estimating function of the form

$$U_{\phi} - c(\phi,\psi)(U_{\psi}^2 - EU_{\psi}^2),$$

if $c(\phi,\psi)$ can be chosen to make it free of ψ, is the best choice for estimating ϕ. An example where this holds is in estimation of ϕ using a random sample of observations from the $N(\psi,\phi)$ distribution.

7.4 Generalizing the E-M Algorithm: The P-S Method

The E-M algorithm (e.g., Dempster et al. (1977)) is a widely used method for dealing with maximum likelihood calculations where there are missing or otherwise incomplete data. It involves the taking of the expectation of the complete-data likelihood with respect to the available data (the E-step) and then maximizing this over possible distributions (the M-step) and the procedure suggests a simple iterative computing algorithm. It has not been available in contexts where a likelihood is unknown or unavailable.

In this section we extend the E-M algorithm method to deal with estimation via estimating functions, in particular the quasi-score. The transitions to estimating functions is made since there are situations where no quasi-log-likelihood exists. The discussion here follows Heyde and Morton (1995).

In our approach the E-step is replaced by a step which *projects* the *quasi-score* rather than taking expectations of a log-likelihood. In many examples the projection is equivalent to *predicting* the missing data or terms. The predictor will not in general be a conditional expectation. The M-step is replaced by *solving* the projected quasi-score set equal to zero. The approach can reasonably be described as the projection-solution (P-S) method.

When the likelihood is available and the score function is included in the class of estimating functions permitted, the standard E-M procedure is recovered from the P-S method.

More broadly, however, we seek to make the point that there is no formal difference between QL estimation for incomplete data and QL estimation for complete data.

7.4.1 From Log-Likelihood to Score Function

As a prelude to generalization of the E-M algorithm formalism from log-likelihoods to estimating functions, we first show how attention can be transferred from operations on the log-likelihood to corresponding ones on its derivative, the likelihood score.

We denote the full data by $\boldsymbol{x}$ and the observed data by $\boldsymbol{y}$. The parameter of interest is $\boldsymbol{\theta}$, a $p\,(\geq 1)$ dimensional vector. The likelihood of $\boldsymbol{\theta}$ based on data $\boldsymbol{z}$ is denoted by $L(\boldsymbol{\theta}; \boldsymbol{z})$.

First note that, from the definition of conditional distributions,

$$\frac{\partial \log L(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} = \frac{\partial \log L(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \boldsymbol{\theta}} - \frac{\partial \log L(\boldsymbol{\theta}; \boldsymbol{x} \mid \boldsymbol{y})}{\partial \boldsymbol{\theta}}. \tag{7.10}$$

The second term on the right hand side of (7.10) has zero expectation conditional on $\boldsymbol{y}$ under the usual regularity conditions and so

$$\frac{\partial \log L(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} = E_\theta \left(\frac{\partial \log\, L(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \boldsymbol{\theta}} \,\middle|\, \boldsymbol{y} \right). \tag{7.11}$$

That is, the likelihood score based on data $\boldsymbol{y}$ is obtained by taking the expectation, conditional on $\boldsymbol{y}$ fixed, of the likelihood score based on the full data $\boldsymbol{x}$.

Now the M-step of the E-M algorithm is based on maximization of

$$E_{\theta_0} (\log\, L(\boldsymbol{\theta}; \boldsymbol{x}) \mid \boldsymbol{y}),$$

and it is clear that, under the usual regularity conditions,

$$\frac{\partial \log L(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} = \left[\frac{\partial}{\partial \boldsymbol{\theta}}\, E_{\theta_0} \left(\log\, L(\boldsymbol{\theta}; \boldsymbol{x}) \,\middle|\, \boldsymbol{y}\right) \right]_{\theta_0 = \theta}. \tag{7.12}$$

The iterative computing algorithm involving $\boldsymbol{\theta}^{(p+1)}$ as the value of $\boldsymbol{\theta}$ that maximizes $E_{\theta^{(p)}} (\log\, L(\boldsymbol{\theta}; \boldsymbol{x}) \,\big|\, \boldsymbol{y})$ then has as its obvious analogue the choice of $\boldsymbol{\theta}^{(p+1)}$ as the value of $\boldsymbol{\theta}$ which solves

$$\frac{\partial}{\partial \boldsymbol{\theta}}\, E_{\theta^{(p)}} \left(\log\, L(\boldsymbol{\theta}; \boldsymbol{x}) \,\middle|\, \boldsymbol{y}\right) = 0 \tag{7.13}$$

or, equivalently, to solve $\partial \log L(\boldsymbol{\theta}; \boldsymbol{y})/\partial \boldsymbol{\theta} = \mathbf{0}$ via

$$\left[E_{\theta^{(p)}} \left(\frac{\partial \log L(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \boldsymbol{\theta}} \,\middle|\, \boldsymbol{y} \right) \right]_{\theta = \theta^{(p+1)}} = \mathbf{0}, \tag{7.14}$$

using (7.11).

7.4.2 From Score to Quasi-Score

The Framework

Now suppose that likelihoods are unsuitable or unavailable and that, based on data $\boldsymbol{x}$, a family of zero mean, square integrable estimating functions $\mathcal{H}_x$ has been chosen within which the quasi-score estimating function is $Q(\boldsymbol{\theta}; \boldsymbol{x})$. The function $\boldsymbol{Q}$ may be a likelihood score but this will not be the case in general, nor will $\boldsymbol{Q}$ in general be expressable as the derivative with respect to $\boldsymbol{\theta}$ of a scalar function. As usual $\boldsymbol{Q}$ is chosen as the estimating function $\boldsymbol{G} \in \mathcal{H}_x$ for which the "information"

$$\mathcal{E}(\boldsymbol{G}) = (E\dot{\boldsymbol{G}})'(E\boldsymbol{G}\,\boldsymbol{G}')^{-1}\,E\dot{\boldsymbol{G}} \tag{7.15}$$

is maximized in the Loewner ordering (partial order of nonnegative definite matrices).

It should be recalled that estimating functions $\boldsymbol{G}$ and $\boldsymbol{M}\,\boldsymbol{G}$, $\boldsymbol{M}$ being any nonsingular matrix of the dimension of $\boldsymbol{G}$, give rise to the same estimators of $\boldsymbol{\theta}$. For each estimating function $\boldsymbol{G}$ we choose

$$\boldsymbol{M} = -(E\dot{\boldsymbol{G}})'(E\boldsymbol{G}\,\boldsymbol{G}')^{-1}$$

and denote $\boldsymbol{G}^{(s)} = \boldsymbol{M}\,\boldsymbol{G}$, the standardized version of $\boldsymbol{G}$. This satisfies the likelihood score property

$$E\boldsymbol{G}^{(s)}\,\boldsymbol{G}^{(s)'} = -E\dot{\boldsymbol{G}}^{(s)} = \mathcal{E}(\boldsymbol{G}^{(s)})\,(= \mathcal{E}(\boldsymbol{G})).$$

All estimating functions in the subsequent discussion of this section will be assumed to be in standardized form and we shall drop the superscripts for convenience. Note that the likelihood score is automatically in standardized form.

We shall subsequently stipulate that all families $\mathcal{G}$ of estimating functions under consideration, such as $\mathcal{H}_x$, are convex. Then, $\boldsymbol{Q}$ is a quasi-score estimating function within $\boldsymbol{G}$ iff

$$E\boldsymbol{G}\,\boldsymbol{Q}' = -E\dot{\boldsymbol{G}} \tag{7.16}$$

for all $\boldsymbol{G} \in \mathcal{G}$ (Theorem 2.1).

In the case where the likelihood score $\boldsymbol{U}$ exists, an equivalent formulation of quasi-score is obtained by defining $\boldsymbol{Q} \in \mathcal{H}_x$ to be a quasi-score estimating function if

$$E\{(\boldsymbol{Q} - \boldsymbol{U})(\boldsymbol{Q} - \boldsymbol{U})'\} = \inf_{G \in \mathcal{H}_x} E\{(\boldsymbol{G} - \boldsymbol{U})(\boldsymbol{G} - \boldsymbol{U})'\}, \tag{7.17}$$

the infimum being taken with respect to the partial order of nonnegative definite matrices. This follows easily from the formulation based on (7.15) upon observing that

$$E\{(\boldsymbol{G}-\boldsymbol{U})(\boldsymbol{G}-\boldsymbol{U})'\} = E\boldsymbol{G}\,\boldsymbol{G}' - E\boldsymbol{U}\,\boldsymbol{U}'$$

(Criterion 2.2 and Theorem 2.1). Also, we see in conjunction with (7.17) that

$$\begin{aligned} E\{(\boldsymbol{G}-\boldsymbol{U})(\boldsymbol{G}-\boldsymbol{U})'\} &= E\{(\boldsymbol{Q}-\boldsymbol{U})(\boldsymbol{Q}-\boldsymbol{U})'\} \\ &\quad + E\{(\boldsymbol{Q}-\boldsymbol{G})(\boldsymbol{Q}-\boldsymbol{G})'\}. \end{aligned} \tag{7.18}$$

The general situation with which we are concerned is when $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$ is observed rather than $\boldsymbol{x}$. We seek to adapt a quasi-score estimating function $\boldsymbol{Q}(\boldsymbol{\theta};\boldsymbol{x})$ to obtain $\boldsymbol{Q}^*(\boldsymbol{\theta};\boldsymbol{y})$ where $\boldsymbol{Q}^*$ is a quasi-score in a class $\mathcal{H}_y$, which is typically a suitable linear subspace of $\mathcal{H}_x$. Then,

$$E\{(\boldsymbol{Q}^*-\boldsymbol{U})(\boldsymbol{Q}^*-\boldsymbol{U})'\} = \inf_{G\in\mathcal{H}_y} E\{(\boldsymbol{G}-\boldsymbol{U})(\boldsymbol{G}-\boldsymbol{U})'\},$$

when $\boldsymbol{U}$ exists, and subtracting $E\{(\boldsymbol{Q}-\boldsymbol{U})(\boldsymbol{Q}-\boldsymbol{U})'\}$ from both sides, we have that

$$E\{(\boldsymbol{Q}^*-\boldsymbol{Q})(\boldsymbol{Q}^*-\boldsymbol{Q})'\} = \inf_{G\in\mathcal{H}_y} E\{(\boldsymbol{G}-\boldsymbol{Q})(\boldsymbol{G}-\boldsymbol{Q})'\}, \tag{7.19}$$

a formula whose use we advocate even when $\boldsymbol{U}$ does not exist. Thus, $\boldsymbol{Q}^*$ is the element of $\mathcal{H}_y$ with minimum dispersion distance from $\boldsymbol{Q} \in \mathcal{H}_x$.

We note that $\boldsymbol{Q}^* = E(\boldsymbol{Q}\,|\,\boldsymbol{y})$ provided this belongs to $\mathcal{H}_y$ as in the E-M algorithm. However, in general $\boldsymbol{Q}^*$ is just given as the least squares predictor (7.19). That is, the "expectation" step is replaced by a "projection" step in the E-M algorithm procedure. We can regard this as a generalization of the E-M technique.

The Algorithm

In general the algorithm proceeds as follows. Write $\|\boldsymbol{a}\|^2 = \boldsymbol{a}'\,\boldsymbol{a}$ for vector $\boldsymbol{a}$. Then, we define $\boldsymbol{H}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_0;\boldsymbol{y})$ such that

$$E_{\theta_0}\,\|\,\boldsymbol{H}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_0;\boldsymbol{y}) - \boldsymbol{Q}(\boldsymbol{\theta};\boldsymbol{x})\|^2 = \inf_{G\in\mathcal{H}_y} E_{\theta_0}\,\|\boldsymbol{G}(\boldsymbol{\theta};\boldsymbol{y}) - \boldsymbol{Q}(\boldsymbol{\theta};\boldsymbol{x})\|^2$$

and solve iteratively for $\boldsymbol{\theta}^{(p+1)}$ from

$$\boldsymbol{H}(\boldsymbol{\theta}^{(p+1)}\,\big|\,\boldsymbol{\theta}^{(p)};\boldsymbol{y}) = \boldsymbol{0}$$

starting from an initial guess $\boldsymbol{\theta}^{(0)}$. A sequence $\{\boldsymbol{\theta}^{(p)}\}$ is generated and, provided $\boldsymbol{\theta}^{(p)} \to \boldsymbol{\theta}$, we have in the limit

$$E_\theta\,\|\boldsymbol{H}(\boldsymbol{\theta}\,\big|\,\boldsymbol{\theta};\boldsymbol{y}) - \boldsymbol{Q}(\boldsymbol{\theta};\boldsymbol{x})\|^2 = \inf_{G\in\mathcal{H}_y} E_\theta\,\|\boldsymbol{G}((\boldsymbol{\theta};\boldsymbol{y}) - \boldsymbol{Q}(\boldsymbol{\theta};\boldsymbol{x})\|^2, \tag{7.20}$$

and

$$\boldsymbol{Q}^*(\boldsymbol{\theta};\boldsymbol{y}) = \boldsymbol{H}(\boldsymbol{\theta}\,\big|\,\boldsymbol{\theta};\boldsymbol{y})$$

satisfies (7.19) provided a solution does exist.

This last result is a straightforward consequence of the following lemma.

Lemma 7.1 If $\boldsymbol{a}$ and $\boldsymbol{b}$ are random vectors such that $E(\boldsymbol{a}\,\boldsymbol{a}' - \boldsymbol{b}\,\boldsymbol{b}')$ is nonnegative definite and $E\|\boldsymbol{a}\|^2 \le E\|\boldsymbol{b}\|^2$, then $E\boldsymbol{a}\,\boldsymbol{a}' = E\boldsymbol{b}\,\boldsymbol{b}'$.

For our application, suppose that $\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{\theta};\boldsymbol{y}) \in \mathcal{H}_y$ satisfies (7.20) and that a solution $\boldsymbol{Q}^* = \boldsymbol{Q}^*(\boldsymbol{\theta};\boldsymbol{y})$ to (7.19) exists. We take $\boldsymbol{a} = \boldsymbol{K} - \boldsymbol{Q}$ and $\boldsymbol{b} = \boldsymbol{Q}^* - \boldsymbol{Q}$ in the lemma. Since (7.19) gives nonnegativity of $E(\boldsymbol{a}\,\boldsymbol{a}' - \boldsymbol{b}\,\boldsymbol{b}')$ while (7.20) gives $E\|\boldsymbol{a}\|^2 \le E\|\boldsymbol{b}\|^2$ we obtain $E\boldsymbol{a}\,\boldsymbol{a}' = E\boldsymbol{b}\,\boldsymbol{b}'$. However, for $\mathcal{H}_y \subset \mathcal{H}_x$ we have $E\boldsymbol{a}'\,\boldsymbol{b} = 0$, so that

$$E\|\boldsymbol{a} - \boldsymbol{b}\|^2 = E\|\boldsymbol{a}\|^2 - E\|\boldsymbol{b}\|^2 = 0,$$

which gives $\boldsymbol{K} = \boldsymbol{Q}^*$ a.s.

Finally, to prove the lemma we note that, upon taking traces,

$$0 \le \operatorname{tr}\{E(\boldsymbol{a}\,\boldsymbol{a}' - \boldsymbol{b}\,\boldsymbol{b}')\} = \operatorname{tr}\{E\|\boldsymbol{a}\|^2 - E\|\boldsymbol{b}\|^2\} \le 0,$$

so that $\boldsymbol{A} = E(\boldsymbol{a}\,\boldsymbol{a}' - \boldsymbol{b}\,\boldsymbol{b}')$ is a symmetric and nonnegative definite matrix with $\operatorname{tr}\boldsymbol{A} = 0$. This forces $\boldsymbol{A} \equiv \boldsymbol{0}$ since the sum of squares of the elements of $\boldsymbol{A}$ is the sum of squares of its eigenvalues all of which must be zero as they are nonnegative and sum to zero.

Alternatives

It may be expected, by analogy with the arguments of Osborne (1992) for the score function, that the algorithm developed above will give a first order rate of convergence compared with the second order rate typical of Fisher's method of scoring. Thus, the latter would be preferred for solving $\boldsymbol{Q}^*(\boldsymbol{\theta};\boldsymbol{y}) = \boldsymbol{0}$ if the necessary derivative calculation could be straightforwardly accomplished. The E-M generalization, of course, has the virtue of avoiding any such requirement. Sufficient conditions for convergence of either method can, in principle, be formulated along the lines of Section 3 of Osborne (1992), but they involve technical assumptions that ensure smoothness and the applicability of a law of large numbers, and are not transparent, so they will not be discussed herein.

It should be noted that there is no formal requirement for $\boldsymbol{Q}$ in (7.19), (7.20) to be a quasi-score estimating function. It can, in principle, be replaced by any estimating function $\boldsymbol{H} \in \mathcal{H}_x$. However, the optimality properties associated with the quasi-score estimating function will then be lost.

In practice $\boldsymbol{Q}^*$ can usually be calculated with the aid of Theorem 2.1, only first and second moment properties being required.

7.4.3 Key Applications

Estimating Functions Linear in the Data

Since explicit expressions are not available in general we shall focus on the most common area of use of quasi-likelihood methods, namely, the situation of estimating functions which are linear in the data.

Let

$$\boldsymbol{\mu} = E_\theta \boldsymbol{x}, \quad \boldsymbol{\nu} = E_\theta \, \boldsymbol{y},$$
$$\boldsymbol{V}_x = E(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})', \quad \boldsymbol{V}_y = E(\boldsymbol{y} - \boldsymbol{\nu})(\boldsymbol{y} - \boldsymbol{\nu})'.$$

If $\mathcal{H}_x$ and $\mathcal{H}_y$ consist, respectively, of functions that are linear in $\boldsymbol{x}$ and $\boldsymbol{y}$, the quasi-score estimating functions may be taken as

$$\begin{aligned} \boldsymbol{Q}(\boldsymbol{\theta}; \boldsymbol{x}) &= \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right)' \boldsymbol{V}_x^{-1} (\boldsymbol{x} - \boldsymbol{\mu}), \\ \boldsymbol{Q}^*(\boldsymbol{\theta}; \boldsymbol{y}) &= \left(\frac{\partial \boldsymbol{\nu}}{\partial \boldsymbol{\theta}}\right)' \boldsymbol{V}_y^{-1} (\boldsymbol{y} - \boldsymbol{\nu}) \end{aligned} \tag{7.21}$$

(e.g. using Theorem 2.1). Also, if $\boldsymbol{y}$ belongs to a linear subspace of $\boldsymbol{x}, \boldsymbol{y} = \boldsymbol{C}' \boldsymbol{x}$ say, then we have

$$\boldsymbol{\nu} = \boldsymbol{C}' \boldsymbol{\mu}, \quad \boldsymbol{V}_y = \boldsymbol{C}' \boldsymbol{V}_x \boldsymbol{C}$$

and the best linear predictor of $\boldsymbol{x}$ given $\boldsymbol{y}$ (in the weighted least squares sense) is

$$\hat{\boldsymbol{x}}(\boldsymbol{y}) = \boldsymbol{\mu} + \boldsymbol{V}_x \boldsymbol{C} \boldsymbol{V}_y^{-1} (\boldsymbol{y} - \boldsymbol{\nu}),$$

so that

$$\boldsymbol{Q}^*(\boldsymbol{\theta}; \boldsymbol{y}) = \boldsymbol{Q}(\boldsymbol{\theta}; \hat{\boldsymbol{x}}(\boldsymbol{y})). \tag{7.22}$$

It should be noted that even in this particular case

$$\boldsymbol{Q}^*(\boldsymbol{\theta}; \boldsymbol{y}) \neq E_\theta(\boldsymbol{Q}(\boldsymbol{\theta}; \boldsymbol{x}) \,|\, \boldsymbol{y}) \tag{7.23}$$

in general, the right-hand side being what is suggested by the E-M prescription. Equality in (7.23) requires that $E(\boldsymbol{x} \,|\, \boldsymbol{y})$ be linear in $\boldsymbol{y}$ as holds, for example, for Normal or Poisson variables.

Another context where equality in (7.23) may not hold occurs when $\mathcal{H}_y$ is not a subset of $\mathcal{H}_x$. Then it can also happen that $\boldsymbol{Q}^* \notin \mathcal{H}_x$. For details involving exponential variables see the example concerned with totals from data with constant coefficient of variation below.

Generalization to Polynomials

The linear description developed above can be exploited to deal with quadratic, and indeed higher polynomial, functions of $\boldsymbol{x}$. To illustrate this, take $\mathcal{H}_x$ to include quadratic functions of $\boldsymbol{x}$ and again $\boldsymbol{y} = \boldsymbol{C}' \boldsymbol{x}$. We can regard $\mathcal{H}_x$ as linear in vectors containing all $\frac{1}{2} \, p(p+3)$ linear and quadratic elements

$$\boldsymbol{x}^* = (x_1, \ldots, x_p, x_1^2, x_1 x_2, \ldots, x_p^2)'$$

and $\mathcal{H}_y$ as linear, in a similarly augmented vector $\boldsymbol{y}^*$ with $\boldsymbol{y}^* = \boldsymbol{C}^{*'} \boldsymbol{x}^*$ where $\boldsymbol{C}^*$ is constructed so as to generate the quadratic functions of $\boldsymbol{y}$.

7.4.4 Examples

Screening Tests

Our first example deals with extension of a standard setting for the E-M algorithm concerned with developing screening tests for a disease given by Laird (1985). Our purpose is to compare the approach via the score or quasi-score estimating function to that of the familiar one via the log-likelihood which is given by Laird.

The problem here involves observed data consisting of a random sample of patients measured on screening tests, each of which gives a dichotomous result. The true disease status is unknown and for Test i the sensitivity has a distribution with mean S_i, $i = 1, 2$. It is assumed that the test results are conditionally independent given disease status and that false positives are not possible. We wish to estimate S_1, S_2 and the disease prevalence π.

Retaining Laird's notation, the observed data can be put into an array of the form

		Test 2	
		+	−
Test 1	+	y_{11}	y_{12}
	−	y_{21}	y_{22}

The complete data are x_{11}, x_{12}, x_{21}, x_{221}, x_{222} where $y_{ij} = x_{ij}$ unless $(i, j) = (2, 2)$ and $y_{22} = x_{221} + x_{222}$, with x_{221} being the number of diseased individuals with negative outcomes on both tests and x_{222} the number of nondiseased patients. We also write $x_{1+} = x_{11} + x_{12}$, $x_{+1} = x_{11} + x_{21}$, N for the sample size $\Sigma_{ij} y_{ij}$ and N_D for the number of diseased patients $(N - x_{222})$.

Laird has supposed that the test sensitivities are random and that, conditional on N_D fixed, x_{1+} and x_{+1} have, respectively, the independent binomial distribution $b(N_D, S_1)$ and $b(N_D, S_2)$ while x_{222} has the binomial distribution $b(N, 1-\pi)$. Then the quasi-score estimating function based on linear functions of the essential data x_{1+}, x_{+1} and N_D is

$$\begin{aligned} \boldsymbol{Q}(\boldsymbol{\theta}; \boldsymbol{x}) &= \frac{x_{1+} - E\left(x_{1+} \,\middle|\, N_D\right)}{\text{var}\left(x_{1+} \,\middle|\, N_D\right)} \frac{\partial E\left(x_{1+} \,\middle|\, N_D\right)}{\partial \boldsymbol{\theta}} \\ &\quad + \frac{x_{+1} - E\left(x_{+1} \,\middle|\, N_D\right)}{\text{var}\left(x_{+1} \,\middle|\, N_D\right)} \frac{\partial E\left(x_{+1} \,\middle|\, N_D\right)}{\partial \boldsymbol{\theta}} \end{aligned}$$

$$+\ \frac{N_D - EN_D}{\text{var}\ N_D}\,\frac{\partial EN_D}{\partial \boldsymbol{\theta}},$$

where $\boldsymbol{\theta} = (S_1, S_2, \pi)'$. We have, under Laird's assumptions,

$$\begin{aligned} &E\left(x_{1+} \,\middle|\, N_D\right) = N_D\, S_1, && \text{var}\ \left(x_{1+} \,\middle|\, N_D\right) = N_D\, S_1\,(1 - S_1),\\ &E\left(x_{+1} \,\middle|\, N_D\right) = N_D\, S_2, && \text{var}\ \left(x_{+1} \,\middle|\, N_D\right) = N_D\, S_2\,(1 - S_2),\\ &EN_D = N\,\pi, && \text{var}\ N_D = N\,\pi\,(1-\pi) \end{aligned}$$

so that

$$\begin{aligned} \boldsymbol{Q}(\boldsymbol{\theta};\boldsymbol{x}) &= \left(\frac{x_{1+} - N_D\, S_1}{S_1(1-S_1)},\ \frac{x_{+1} - N_D\, S_2}{S_2(1-S_2)},\ \frac{N_D - N\,\pi}{\pi(1-\pi)}\right)' \\ &= \left(\frac{x_{1+}}{S_1} - \frac{N_D - x_{1+}}{1-S_1},\ \frac{x_{+1}}{S_2} - \frac{N_D - x_{+1}}{1-S_2},\ \frac{N_D}{\pi} - \frac{x_{222}}{1-\pi}\right)', \end{aligned}$$

which is also the likelihood score estimating function as is easily seen from Laird's discussion. Then, the obvious choice for $\mathcal{H}_y$ yields the quasi-score

$$\begin{aligned} E_\theta(\boldsymbol{Q}(\boldsymbol{\theta};\boldsymbol{x}) \,|\, \boldsymbol{y}) &= \left(\frac{y_{1+}}{S_1} - \frac{E(N_D \,|\, \boldsymbol{y}) - y_{1+}}{1-S_1},\right. \\ &\qquad \left.\frac{y_{+1}}{S_2} - \frac{E(N_D \,|\, \boldsymbol{y}) - y_{+1}}{1-S_2},\ \frac{E(N_D \,|\, \boldsymbol{y})}{\pi} - \frac{N - E(N_D \,|\, \boldsymbol{y})}{1-\pi}\right)' \end{aligned}$$

and a simple conditional probability argument gives

$$E(N_D \,|\, \boldsymbol{y}) = \frac{N - y_{22}(1-\pi)}{(1-S_1)(1-S_2)\,\pi + (1-\pi)}. \tag{7.24}$$

If an algorithmic solution is sought, we set $N_D^{(p+1)} = E(N_D \,|\, \boldsymbol{y}, \boldsymbol{\theta}^{(p)})$ and we have

$$\begin{aligned} H(\boldsymbol{\theta}^{(p+1)} \,|\, \boldsymbol{\theta}^{(p)}, \boldsymbol{y}) &= \left(\frac{y_{1+}}{S_1^{(p+1)}} - \frac{N_D^{(p+1)} - y_{1+}}{1 - S_1^{(p+1)}},\right. \\ &\qquad \left.\frac{y_{+1}}{S_2^{(p+1)}} - \frac{N_D^{(p+1)} - y_{+1}}{1 - S_2^{(p+1)}},\ \frac{N_D^{(p+1)}}{\pi^{(p+1)}} - \frac{N - N_D^{(p+1)}}{1 - \pi^{(p+1)}}\right)' \end{aligned}$$

leading to

$$\pi^{(p+1)} = \frac{N_D^{(p+1)}}{N},\quad S_1^{(p+1)} = \frac{y_{1+}}{N_D^{(p+1)}},\quad S_2^{(p+1)} = \frac{y_{+1}}{N_D^{(p+1)}}$$

as obtained by Laird for the E-M algorithm solution. The iterative solution converges quite quickly in this case but an explicit solution to the estimating equation

$$E_\theta(\boldsymbol{Q}(\boldsymbol{\theta};\boldsymbol{x} \,\big|\, \boldsymbol{y}) = \boldsymbol{0}$$

is also available. After some algebra we find that

$$\hat{S}_1 = \frac{y_{11}}{y_{+1}},\ \hat{S}_2 = \frac{y_{11}}{y_{1+}},\ \hat{\pi} = \frac{1 - \frac{y_{22}}{N}}{1 - (1 - \frac{y_{11}}{y_{+1}})(1 - \frac{y_{11}}{y_{1+}})} = \frac{y_{+1}\, y_{1+}}{N\, y_{11}},$$

which gives the maximum likelihood estimator for $\boldsymbol{\theta}$ based on $\boldsymbol{y}$.

Note that the full distributional assumptions made by Laird are not needed in the above analysis. All that is required is the first and second (conditional) moment behavior of x_{1+}, x_{+1} and N_D. Suppose now that the data are not homogeneous and perhaps not really from a random sample. It is, nevertheless, still reasonable to take

$$E(x_{1+} \,|\, N_D) = N_D\, S_1, \qquad E(x_{+1} \,|\, N_D) = N_D\, S_2$$

and to suppose that

$$\frac{E\left(x_{221} \,|\, y_{22}\right)}{E\left(x_{222} \,|\, y_{22}\right)} = \frac{\pi\,(1 - S_1)(1 - S_2)}{1 - \pi},$$

which, together with

$$E(N_D \,|\, y) = N - E\left(x_{222} \,|\, y_{22}\right)$$

still gives (7.24).

The expressions for the conditional variances of $x_{+1} \,|\, N_D$ and $x_{1+} \,|\, N_D$ may need to be changed to incorporate a measure of dispersion. For example, if the individuals counted in x_{+1} (resp. x_{1+}) displayed sensitivity to Test 1 (resp. 2) with a distribution of beta type with mean S_1 (resp. S_2), then formulas

$$\text{var } (x_{1+} \,|\, N_D) = \phi_1\, N_D\, S_1(1 - S_1), \quad \text{var } (x_{+1} \,|\, N_D) = \phi_2\, N_D\, S_2(1 - S_2)$$

would be obtained for certain dispersion parameters ϕ_1, ϕ_2. If it were reasonable to assume that the dispersion parameters could be taken as the same, the original analysis would remain unchanged. If not, obvious adjustments could be made.

Totals from Data with Constant Coefficient of Variation

Here we consider a situation where the full data set is $\boldsymbol{x} = (x_{ij},\ 1, 2, \ldots, n;\ j = 1, 2)$ and it is the totals $y_i = x_{i1} + x_{i2}$ that are observed, so that $\boldsymbol{y} = (y_i,\ i = 1, 2, \ldots, n)$. The likelihood is not known but the x_{ij} are assumed to be independent with moments

$$E(x_{ij}) = \mu_{ij}, \qquad \text{var } x_{ij} = \phi\, \mu_{ij}^2$$

where $\mu_{ij} = \mu_{ij}(\boldsymbol{\theta})$ and ϕ is the coefficient of variation, assumed to be constant. The aim is to estimate $\boldsymbol{\theta}$, taken as scalar for convenience, and for simplicity we assume that $\phi = 1$ is known.

The class $\mathcal{H}_x$ consists of the linear functions of $\boldsymbol{x}$ and the usual quasi-score is (e.g., McCullagh and Nelder (1989))

$$Q(\theta;\boldsymbol{x}) = \sum_{ij} (x_{ij} - \mu_{ij})\, \mu_{ij}^{-2} \, \frac{\partial \mu_{ij}}{\partial \theta}.$$

Since we observe only the totals $\boldsymbol{y}$ we consider $\mathcal{H}_y$, the class of linear functions of $\boldsymbol{y}$. Then the corresponding quasi-score is

$$Q^*(\theta;\boldsymbol{y}) = \sum_i (y_i - (\mu_{i1} + \mu_{i2}))(\mu_{i1}^2 + \mu_{i2}^2)^{-2} \, \frac{\partial}{\partial \theta}\, (\mu_{i1} + \mu_{i2}). \qquad (7.25)$$

The linear predictor $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}$ given $\boldsymbol{y}$ is

$$\hat{x}_{ij} - \mu_{ij} = \frac{\mu_{ij}^2}{\mu_{i1}^2 + \mu_{i2}^2}\, (y_i - (\mu_{i1} + \mu_{i2}))$$

and we see that

$$Q^*(\theta;\boldsymbol{y}) = Q(\theta;\hat{\boldsymbol{x}}).$$

It should be noted that Q^* is not the expectation of Q, which is general is nonlinear in $\boldsymbol{y}$. For example, suppose that x_{ij} has the exponential density

$$f(x_{ij};\theta) = \begin{cases} \mu_{ij}^{-1} \exp(-x_{ij}/\mu_{ij}), & \text{if} \quad 0 < x_{ij} < \infty, \\ 0, & \qquad\qquad \text{otherwise.} \end{cases}$$

Then it is easily checked that

$$E(x_{ij} \,|\, y_i;\theta) = \nu_i - \frac{y_i \exp(-y_i/\nu_i)}{1 - \exp(-y_i/\nu_i)},$$

where

$$\nu_i^{-1} = \mu_{i1}^{-1} - \mu_{i2}^{-1}.$$

In this case $Q = U$ is the likelihood score for the full data but

$$U^* = E(U \,|\, \boldsymbol{y};\theta)$$

differs from Q^*, which is the projection of Q to be used when the exponential likelihood specification is doubtful or inappropriate.

There is some evidence that there is little loss in efficiency in using Q instead of U^* even when the exponential assumption is actually correct. To evaluate the asymptotic relative efficiency we must compare

$$\begin{aligned} \mathcal{E}(U^*) &= E(U^*)^2 & (7.26) \\ &= \sum_i E(E(x_{i1} \,|\, y_i) - \mu_{i1}) \left(\frac{\partial \mu_{ij}^{-1}}{\partial \theta} \right)^2 + (E(x_{i2} \,|\, y_i) - \mu_{i2}) \left(\frac{\partial \mu_{ij}^{-1}}{\partial \theta} \right)^2, \end{aligned}$$

and

$$\mathcal{E}(Q^*) = E(Q^*)^2 = \sum_i \left(\mu_{i1}^2 + \mu_{i2}^2\right)^{-1} \left(\frac{\partial \mu_{i1}}{\partial \theta} + \frac{\partial \mu_{i2}}{\partial \theta}\right)^2 . \tag{7.27}$$

The information measures (7.26) and (7.27) satisfy the inequality

$$E(U^*)^2 \geq E(Q^*)^2$$

but a concrete comparison requires more specific values for the μ_{ij}. For example, in the particular case where $\mu_{i2} = \lambda\, \mu_{i1}$ with λ fixed and taking $\delta = \lambda - 1 > 0$ without loss of generality we find that

$$E(Q^*)^2 = \frac{(1+\lambda)^2}{1+\lambda^2} \sum_i \left(\frac{\partial \mu_{i2}}{\partial \theta}\right)^2 \tag{7.28}$$

while $E(U^*)^2$ is given below. Indeed, temporarily dropping subscript i again, we find that

$$f(y; \mu_1, \mu_2) = \frac{1+\delta}{\delta\, \mu_2}\, e^{-y/\mu_2}\, (1 - e^{-\delta y/\mu_2}), \quad y > 0,$$

and, after some algebra,

$$\begin{aligned} E\left(-\frac{\partial^2}{\partial \theta^2} \log f(y; \mu_1, \mu_2)\right) &= \left(\frac{\partial \mu_2}{\partial \theta}\right)^2 \left(1 + \delta(1+\delta) \int_0^\infty z^2\, \frac{e^{-(1+\delta)z}}{1 - e^{-\delta z}}\, dz\right) \\ &= \left(\frac{\partial \mu_2}{\partial \theta}\right)^2 \left(1 + 2\,\delta(1+\delta) \sum_{r=1}^\infty \frac{1}{(1 + r\,\delta)^3}\right) \\ &= \left(\frac{\partial \mu_2}{\partial \theta}\right)^2 \left(1 + 2\,\frac{1+\delta}{\delta^2}\, \zeta\left(3, \frac{1}{\delta}\right)\right), \end{aligned}$$

where $\zeta(3, \frac{1}{\delta}) = \sum_{r=1}^\infty (r + \frac{1}{\delta})^{-3}$ is the generalized Riemann zeta-function. Then,

$$E(E(U \,|\, \boldsymbol{y}))^2 = \left(1 + 2\,\frac{1+\delta}{\delta^2}\, \zeta\left(3, \frac{1}{\delta}\right)\right) \sum_i \left(\frac{\partial \mu_{i2}}{\partial \theta}\right)^2 \tag{7.29}$$

and the asymptotic relative efficiency of Q^* is

$$ARE(Q^*)(\lambda) = \frac{(1+\lambda)^2}{1+\lambda^2}\, \frac{1}{1 + 2\,\lambda(\lambda-1)^{-2}\, \zeta\,(3, (\lambda-1)^{-1})}$$

using (7.27) and (7.28). It is easily seen that $ARE(Q^*)(\lambda)$ approaches unity as $\lambda \to 1$ or $\lambda \to \infty$. A plot of ARE$(Q^*)(\lambda)$ against $\lambda\,(1 < \lambda < 50)$ has been obtained with the aid of the package *MAPLE* and appears in Heyde and Morton (1995). The *ARE* always exceeds 0.975.

Time Series Smoothing with Missing Observations

The E-M algorithm itself is mainly useful in situations where the $\boldsymbol{y}$-likelihood is intractable but the $\boldsymbol{x}$-likelihood is simple (so that the M-step is easily achieved). The algorithm outlined in this section is similarly useful in situations where quasi-score $\boldsymbol{Q}$ is simple but $\boldsymbol{Q}^*$ is intractable. Many such examples can be generated in the context of time series observed subject to noise, and with missing data as well, when Gaussian distributional assumptions are questionable or inappropriate.

The example we discuss here comes from Shumway and Stoffer (1982) whose discussion has been succinctly summarized by Little and Rubin (1987, pp. 165-168). It concerns data on expeditures for physician services and is treated using the model:

$$x_{ij} = z_i + \epsilon_{ij}, \quad j = 1, 2, \tag{7.30}$$

$$z_i = \phi\, z_{i-1} + \eta_i, \tag{7.31}$$

where the ϵ_{ij} are independent with $E\epsilon_{ij} = 0$, var $\epsilon_{ij} = B_j$, the η_i are independent with $E\eta_i = 0$, var $\eta_i = Q$ and ϕ is an inflation factor. The x_{ij}'s are the observations but in practice only a subset y_{ij} have been observed $y_{ij} = x_{ij}$ for $i, j \in A$, say. The object is to estimate the z_i's, which are not assumed to be stationary.

Now Shumway and Stoffer have assumed, without comment, that the ϵ_{ij} and η_i are normally distributed and they have used the E-M algorithm to calculate the z_i, the corresponding likelihood being intractable. However, if normality is not assumed it is still reasonable to use linear estimating equations for this problem. Then, the quasi-scores are identical to those based on the normal likelihood and the E-M algorithm results are preserved in our approach. The inferences remain appropriate for nonnormal data having unchanged first and second moments.

7.5 Exercises

1. (Mixture densities) Suppose that $0 \leq \psi \leq 1$ and let $X_1, \ldots, X_n$ be independent and identically distributed with mixture density

   $$h(x; \phi, \psi) = \psi\, f(x; \phi) + (1 - \psi)\, g(x; \phi).$$

   Here ϕ is to be estimated.

   If the functions h, f, g are known, so that the score function $\boldsymbol{U} = (U_\phi, U_\psi)'$ can be used as a basis for the calculation, show that the optimal estimating equation for ϕ is

   $$U_\phi - [E(U_\phi\, U_\psi)/EU_\phi^2]\, U_\psi = 0$$

   (Small and McLeish (1989)).

2. (A form of Rao-Blackwellization) Let $\boldsymbol{Q}$ be a quasi-score estimating function. If $\boldsymbol{K}$ is an estimating function show that $E(\boldsymbol{K} \mid \boldsymbol{Q})$ is preferable to $\boldsymbol{K}$ in the usual sense that $\mathcal{E}(E(\boldsymbol{K} \mid \boldsymbol{Q})) \geq \mathcal{E}(\boldsymbol{K})$ in the partial order of nonnegative definite matrices (Small and McLeish (1994, p. 85)).

Chapter 8

Bypassing the Likelihood

8.1 Introduction

Estimation based on the application of maximum likelihood methods can involve quite formidable difficulty in calculation of the joint distribution of the observations and sometimes even in differentiating the likelihood to obtain the corresponding estimating equations. Notable examples are the derivation of the restricted (or residual) (REML) estimating functions for dispersion parameters associated with linear models and the derivation of the maximum likelihood estimators of parameters in diffusion type models. In the former case both the derivation of the likelihood and its differentiation are less than straightforward while for the latter the Radon-Nikodym derivative calculations are a significant obstacle. Quasi-likelihood methods, however, allow such estimators (or estimating functions) to be obtained quite painlessly and under more general conditions. In this chapter we shall illustrate the power and simplicity of the approach through three quite different examples.

8.2 The REML Estimating Equations

Suppose that the $n \times 1$ vector $\boldsymbol{y}$ has the multivariate normal distribution $MVN(\boldsymbol{X\beta}, \boldsymbol{V}(\boldsymbol{\theta}))$ with mean vector $\boldsymbol{X\beta}$ and covariance matrix $\boldsymbol{V}(\boldsymbol{\theta})$. For simplicity we take the rank of $\boldsymbol{X}$ as r, the dimension of the vector $\boldsymbol{\beta}$. Let $\boldsymbol{A}$ be any matrix with n rows and rank $n-r$ satisfying $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{0}$. Then, the distribution of $\boldsymbol{A}'\boldsymbol{y}$ is $MVN(0, \boldsymbol{A}'\boldsymbol{V}\boldsymbol{A})$ which may be singular and it can be shown, with some difficulty, that the likelihood function of $\boldsymbol{\theta}$ based on $\boldsymbol{A}'\boldsymbol{y}$ has the form of a constant multiplied by

$$\left(\prod_{i=1}^{n-r} \ell_i\right)^{1/2} \exp\left\{-\frac{1}{2}\,\boldsymbol{y}'\,\boldsymbol{V}^{-1}\boldsymbol{Q}\,\boldsymbol{y}'\right\}, \tag{8.1}$$

where $\boldsymbol{Q} = I - \boldsymbol{P}_{R(X)}^{V}$, $\boldsymbol{P}_{R(X)}^{V}$ denoting the projector $\boldsymbol{X}\,(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{V}^{-1}$ onto the subspace $R(\boldsymbol{X})$, the range space of the matrix $\boldsymbol{X}$ that is the orthogonal projection onto the orthogonal complement of $R(\boldsymbol{X})$ with respect to the inner product $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \boldsymbol{a}'\boldsymbol{V}^{-1}\boldsymbol{b}$, while $\ell_1, \dots, \ell_{n-r}$ are the nonzero eigenvalues of $\boldsymbol{V}^{-1}\boldsymbol{Q}$. (See Rao (1973, Eq. (8a.4.11) p. 528) for the case where $\boldsymbol{A}'\boldsymbol{V}\boldsymbol{A}$ has full rank.) The striking thing here is that the result (8.1) does not depend on $\boldsymbol{A}$. That is, *all* rank $n-r$ matrices that define error contrasts (i.e., for which $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{0}$) generate the same likelihood function in $\boldsymbol{\theta}$. This follows from the result

$$\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{V}\boldsymbol{A})^{-}\boldsymbol{A}' = \boldsymbol{V}^{-1}\boldsymbol{Q} \tag{8.2}$$

for all $\boldsymbol{A}$ and any g-inverse $(\boldsymbol{A}'\boldsymbol{V}\boldsymbol{A})^{-}$. The basic reference on this topic is Harville (1977), although detailed derivations are not included, the reader being referred to an unpublished technical report. For a more recent derivation see Verbyla (1990).

The next step is to differentiate (8.1) with respect to $\boldsymbol{\theta}$. Again this is less than straightforward, involving vector differentiation of matrices and determinants, but there finally emerge the REML estimating equations

$$\operatorname{tr}\left\{\boldsymbol{V}^{-1}\boldsymbol{Q}\,\frac{\partial \boldsymbol{V}}{\partial \theta_i}\right\} = \boldsymbol{y}'\left\{\boldsymbol{V}^{-1}\boldsymbol{Q}\,\frac{\partial \boldsymbol{V}}{\partial \theta_i}\,\boldsymbol{V}^{-1}\boldsymbol{Q}\right\}\boldsymbol{y}, \tag{8.3}$$

where tr denotes trace.

We shall now show how the result (8.3) can be derived painlessly, and under more general conditions, by quasi-likelihood methods. We shall no longer suppose that $\boldsymbol{y}$ has the multivariate normal distribution but only that its mean vector is $\boldsymbol{X\beta}$, its covariance matrix is $\boldsymbol{V}(\boldsymbol{\theta})$, that $Ey_i^2y_j^2 = Ey_i^2\,Ey_j^2$ and $Ey_i^3y_j = 0$ for $i \neq j$ and that each y_i has kurtosis 3.

For fixed $\boldsymbol{A}$, let $\boldsymbol{z} = \boldsymbol{A}'\boldsymbol{y}$ and, in the expectation that it is quadratic functions of the data that should be used to estimate covariances, consider the class of quadratic form estimating functions

$$\begin{aligned}\mathcal{H} = \{\boldsymbol{G} :\ \boldsymbol{G} = (\boldsymbol{G}(\boldsymbol{S}_i), \ldots, \boldsymbol{G}(\boldsymbol{S}_p))',\ \boldsymbol{G}(\boldsymbol{S}_i) = \boldsymbol{z}'\boldsymbol{S}_i\boldsymbol{z} - E(\boldsymbol{z}'\boldsymbol{S}_i\boldsymbol{z}),\\ \boldsymbol{S}_i \text{ symmetric matrix } i = 1, 2, \ldots, p\},\end{aligned}$$

$\boldsymbol{\theta}$ being of dimension p. Write $\boldsymbol{W} = \boldsymbol{A}'\boldsymbol{V}\boldsymbol{A}$.

Theorem 8.1 $\boldsymbol{G}^* = (\boldsymbol{G}(\boldsymbol{S}_i^*), \ldots, \boldsymbol{G}(\boldsymbol{S}_p^*))' \in \mathcal{H}$ is a quasi-score estimating function in $\mathcal{H}$ if

$$\boldsymbol{S}_i^* = \boldsymbol{W}^{-}\,\frac{\partial \boldsymbol{W}}{\partial \theta_i}\,\boldsymbol{W}^{-},\ i = 1, 2, \ldots, p, \tag{8.4}$$

for any g-inverse $\boldsymbol{W}^{-} = (\boldsymbol{A}'\boldsymbol{V}\boldsymbol{A})^{-}$. Furthermore, the $\boldsymbol{G}(\boldsymbol{S}_i^*)$ do not depend on $\boldsymbol{A}$ and the corresponding estimating equation $\boldsymbol{G}(\boldsymbol{S}_i^*) = \boldsymbol{0}$ is Equation (8.3).

Proof. Using standard covariance results for quadratic forms in multivariate normal variables (see, e.g., Karson (1982, p. 62)) that involve only the moment assumptions noted above, we have

$$\operatorname{cov}(\boldsymbol{G}(\boldsymbol{S}_i),\, \boldsymbol{G}(\boldsymbol{S}_j^*)) = 2\operatorname{tr}(\boldsymbol{W}\boldsymbol{S}_i\boldsymbol{W}\boldsymbol{S}_j^*), \tag{8.5}$$

while

$$E\boldsymbol{z}'\boldsymbol{S}_i\boldsymbol{z} = \operatorname{tr}(\boldsymbol{W}\boldsymbol{S}_i).$$

Also, it is easily checked that the (i,j) element of $E\dot{\boldsymbol{G}}$,

$$(E\dot{\boldsymbol{G}})_{ij} = -\operatorname{tr}\left(\frac{\partial \boldsymbol{W}}{\partial \theta_j}\,\boldsymbol{S}_i\right).$$

Then, when $\boldsymbol{S}_i^*$ is given by (8.4), we see that

$$\operatorname{cov}(\boldsymbol{G}(\boldsymbol{S}_i), \boldsymbol{G}(\boldsymbol{S}_j^*)) = -2(E\dot{\boldsymbol{G}}(\boldsymbol{S}))_{ij},$$

since $\operatorname{tr} \boldsymbol{BC} = \operatorname{tr} \boldsymbol{CB}$ and $\boldsymbol{A}\boldsymbol{W}^-\boldsymbol{W} = \boldsymbol{A}$, $\boldsymbol{W}\boldsymbol{W}^-\boldsymbol{A}' = \boldsymbol{A}'$ for any choice of generalized inverse (e.g., Lemma 2.2.6, p. 22 of Rao and Mitra (1971)). Consequently, the result of the theorem concerning the form of $\boldsymbol{S}_i^*$ follows from Theorem 2.1. The REML estimating equations (8.5) then follow immediately using $\boldsymbol{z} = \boldsymbol{A}'\boldsymbol{y}$ and the representation (8.2).

The theorem is new even in the case where the mean of $\boldsymbol{y}$ is known to be zero (i.e., the case $r = 0$). The results could, in principle, be adjusted to deal with y_i's having different moments from those of the corresponding multivariate normal variables. For a general discussion of variance function estimation see Davidian and Carroll (1987). The results in this section are from Heyde (1994b). For a discussion of consistency and asymptotic normality of REML estimators see Jiang (1996).

8.3 Estimating Parameters in Diffusion Type Processes

The standard method of estimation for parameters in the drift coefficient of a diffusion process involves calculation of a likelihood ratio (Radon-Nikodym derivative) and thence the maximum likelihood estimator(s). This is less than straightforward for more complicated models and indeed it is often not available at all because of the nonexistence of the Radon-Nikodym derivative. The methods of quasi-likelihood , however, allow estimators to be obtained straightforwardly under very general conditions. They can deal, in particular, with the situation in which the Brownian motion in a diffusion is replaced by a general square-integrable martingale. The approach, which is based on selection of an optimal estimating function from within a specified class of such functions, involves assumptions on only the first two conditional moments of the underlying process. Nevertheless, the quasi-likelihood estimators will ordinarily be true maximum likelihood estimators in a context where the Radon-Nikodym derivative is available. Furthermore, they will generally be consistent, asymptotically normally distributed, and can be used to construct minimum size asymptotic confidence zones for the unknown parameters among estimators coming from the specified class. In this section we illustrate these ideas through a general discussion and application to the Cox-Ingersoll-Ross model for interest rates and to a modification of the Langevin model for dynamical systems. The results are from Heyde (1994a).

The models which we shall consider in this section can all be written in the semimartingale form

$$d\boldsymbol{X}_t = d\boldsymbol{A}_t(\boldsymbol{\theta}) + d\boldsymbol{M}_t(\boldsymbol{\theta}), \tag{8.6}$$

where the finite variation process $\{\boldsymbol{A}_t\}$ can be interpreted as the *signal* and the local martingale $\{\boldsymbol{M}_t\}$ can be interpreted as the *noise*, as has been dicussed in Section 2.6. Then, the quasi-score estimating function based on the family of local martingale estimating functions

$$\mathcal{H} = \left\{ \int_0^T \boldsymbol{a}_t(\boldsymbol{\theta})\, d\boldsymbol{M}_t(\boldsymbol{\theta}), \quad \{\boldsymbol{a}_t\} \text{ predictable} \right\}$$

is easily seen from Theorem 2.1 to be

$$\int_0^T \left(E\left(d\dot{\boldsymbol{M}}_t(\boldsymbol{\theta}) \,\middle|\, \mathcal{F}_{t-} \right)\right)' (d\langle \boldsymbol{M}(\boldsymbol{\theta})\rangle_t)^- \, d\boldsymbol{M}_t(\boldsymbol{\theta}), \tag{8.7}$$

where $\{\mathcal{F}_t\}$ is a filtration of past-history σ-fields, $\langle \boldsymbol{M}(\boldsymbol{\theta})\rangle_t$ is the quadratic characteristic and the $-$ denotes the generalized inverse.

Now (8.7) can be rewritten as

$$\int_0^T \left(E\left(d\dot{\boldsymbol{A}}_t(\boldsymbol{\theta}) \,\middle|\, \mathcal{F}_{t-} \right)\right)' (d\langle \boldsymbol{M}(\boldsymbol{\theta})\rangle_t)^- \, (d\boldsymbol{A}_t(\boldsymbol{\theta}) - d\boldsymbol{X}_t)$$

from which it is clear that the quasi-score estimating function is unaffected if the local martingale noise $\{\boldsymbol{M}_t(\boldsymbol{\theta})\}$ is replaced by another whose quadratic characteristic is the same. The precise distributional form of the noise does not need to be known. In the commonly met situation where $\boldsymbol{M}_t(\boldsymbol{\theta}) = \sigma W_t$ with $\sigma > 0$ and W_t being standard Brownian motion, the results are robust to the extent that $\{\boldsymbol{M}_t(\boldsymbol{\theta})\}$ could be replaced by any local martingale $\{\boldsymbol{Z}_t(\boldsymbol{\theta})\}$, for example, one with independent increments, for which $\langle \boldsymbol{Z}\rangle_t = \sigma^2 t$ without changing the estimators.

In the particular case of a diffusion process the components on the right-hand side of the representation (8.6) can be written as

$$d\boldsymbol{A}_t(\boldsymbol{\theta}) = \boldsymbol{a}(t, \boldsymbol{X}_t, \boldsymbol{\theta})\, dt, \quad d\boldsymbol{M}_t = b^{\frac{1}{2}}(t, \boldsymbol{X}_t)\, dW_t, \tag{8.8}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are known vector and matrix functions, respectively, and $\{W_t\}$ is standard Brownian motion. Then, an appropriate Radon-Nikodym derivative of the measure induced by the process $\{\boldsymbol{X}_t,\ 0 \le t \le T\}$ with parameter $\boldsymbol{\theta}$ with respect to the corresponding measure for parameter $\boldsymbol{\theta}_0$ can be calculated and is given by

$$\exp\left\{ \int_0^T \boldsymbol{C}(t, \boldsymbol{X}_t)\, d\boldsymbol{X}_t - \int_0^T \boldsymbol{D}(t, \boldsymbol{X}_t)\, dt \right\}, \tag{8.9}$$

where

$$\boldsymbol{b}(t,x)\, \boldsymbol{C}(t,x) = \boldsymbol{a}(t,x,\boldsymbol{\theta}) - \boldsymbol{a}(t,x,\boldsymbol{\theta}_0),$$

$$\boldsymbol{D}(t,x) = (\boldsymbol{a}(t,x,\boldsymbol{\theta}_0))'\, \boldsymbol{C}(t,x) + \frac{1}{2}\, (\boldsymbol{C}(t,x))' \boldsymbol{b}(t,x)\, \boldsymbol{C}(t,x)$$

(e.g., Basawa and Prakasa Rao (1980, p. 219)).

From (8.9) it is easily checked that the likelihood score function (derivative of the logarithm of the Radon-Nikodym derivative with respect to $\boldsymbol{\theta}$) is given by (8.7). This means that the quasi-likelihood estimator is the maximum likelihood estimator for the model (8.8). However, as indicated above, the quasi-likelihood estimator is available much more generally.

The explanation for the quasi-score corresponding to the likelihood score for the model (8.8) is not hard to discern. A likelihood score is a martingale under modest regularity conditions and all square integrable martingales living on the same probability space as the Brownian motion in the noise term of the model (8.8) can be described as stochastic integrals with respect to the Brownian motion (see, e.g., Theorem 5.17 of Liptser and Shiryaev (1977)). The likelihood score will be one such martingale and will therefore be included in the relevant family $\mathcal{H}$ over which optimization takes place and it solves the optimization problem.

As a concrete illustration of the methodology we shall discuss the stochastic differential equation

$$dX_t = \alpha(\beta - X_t)\,dt + \sigma\sqrt{X_t}\,dW_t, \tag{8.10}$$

where $X_0 > 0,\ \alpha > 0,\ \beta > 0,\ \sigma > 0$. This form was proposed by Cox, Ingersoll and Ross (1985) as a model for interest rates and it has been widely used in finance.

In considering the model (8.10) we shall be concerned with the estimation of $\boldsymbol{\theta} = (\alpha, \beta)'$. The parameter σ can be regarded as known whenever Brownian motion and continuous sampling are involved. Indeed, σ can be calculated with probability one on the basis of knowledge of a path of the process on any finite time interval. This follows from the definitions of the quadratic variation process and stochastic integrals with respect to Brownian motion (e.g., Rogers and Williams (1987, Chapter IV, Section 4)) from which one obtains that, writing $t_i^{(n)} = \min(T, 2^{-n}i)$,

$$\lim_{n\to\infty} \sum_{i=0}^{\infty} \left(X_{t_{i+1}^{(n)}} - X_{t_i^{(n)}} \right)^2 = \sigma^2 \int_0^T X_t\,dt \qquad \text{a.s.}$$

and

$$\lim_{n\to\infty} \sum_{i=0}^{\infty} X_{t_i^{(n)}} \left(t_{i+1}^{(n)} - t_i^{(n)} \right) = \int_0^T X_t\,dt \qquad \text{a.s.}$$

For the model (8.10) the Radon-Nikodym derivative of the measure based on $(\alpha, \beta)'$ with respect to that based on $(\alpha_0, \beta_0)'$ is easily seen from (8.9) to be

$$\exp\left\{ \sigma^{-2} \int_0^T X_t^{-1} \left[\alpha(\beta - X_t) - a_0(\beta_0 - X_t) \right] dX_t \right.$$
$$\left. - \frac{1}{2}\sigma^{-2} \int_0^T X_t^{-1} \left[\alpha^2(\beta - X_t)^2 - a_0^2(\beta_0 - X_t)^2 \right] dt \right\}$$

and differentiating the logarithm of this likelihood ratio with respect to $\boldsymbol{\theta} = (\alpha, \beta)'$ gives the likelihood score

$$\boldsymbol{U}_T = \sigma^{-2} \int_0^T \left(\begin{array}{c} \beta - X_t \\ \alpha \end{array} \right) X_t^{-\frac{1}{2}} \, dW_t, \tag{8.11}$$

which is also the quasi-score given by (8.7).

If we modify the model to the form

$$dX_t = \alpha(\beta - X_t)\, dt + \sigma(\alpha, \gamma) \sqrt{X_t}\, dW_t$$

where $\sigma(\alpha, \gamma)$ reflects a possibly rate dependent noise, then the likelihood ratio does not exist in general. Indeed, when $\sigma(\alpha_1, \gamma) \neq \sigma(\alpha_2, \gamma)$ the supports of the distributions of the two processes are disjoint. The quasi-likelihood methodology, however, is unaffected by this change. The quasi-score estimating function continues to be given by (8.11) and the asymptotic properties of the estimators are also unaffected.

From (8.11) the maximum likelihood/quasi-likelihood estimators $\hat{\alpha}_T$, $\hat{\beta}_T$ are given by

$$\int_0^T \left(\hat{\beta}_T - X_t\right) X_t^{-1} \left[dX_t - \hat{\alpha}_T \left(\hat{\beta}_T - X_t\right) dt\right] = 0,$$

$$\int_0^T X_t^{-1} \left[dX_t - \hat{\alpha}_T \left(\hat{\beta}_T - X_t\right) dt\right] = 0,$$

and putting

$$I_T = \int_0^T X_t^{-1}\, dX_t, \qquad J_T = \int_0^T X_t^{-1}\, dt, \qquad K_T = \int_0^T X_t\, dt,$$

we find that

$$\begin{aligned} \hat{\alpha}_T &= (I_T T - J_T(X_T - X_0))/(J_T K_T - T^2), \\ \hat{\beta}_T &= (I_T K_T - T(X_T - X_0))/(I_T T - J_T(X_T - X_0)). \end{aligned}$$

For the model with $2\alpha\beta \geq \sigma^2$ there is a strictly positive ergodic solution to (8.10) as $T \to \infty$ whose distribution has gamma density $\Gamma(2\alpha\beta/\sigma^2,\, 2\alpha/\sigma^2)$ (see, e.g., Kloeden and Platen (1992, p. 38)). Suppose X_∞ has this density; then

$$E X_\infty = \beta, \qquad E X_\infty^{-1} = 2\alpha/(2\alpha\beta - \sigma^2). \tag{8.12}$$

Using the ergodic theorem we obtain

$$T^{-1} J_T = T^{-1} \int_0^T X_t^{-1}\, dt \xrightarrow{\text{a.s.}} E X_\infty^{-1}, \tag{8.13}$$

$$T^{-1} K_T = T^{-1} \int_0^T X_t\, dt \xrightarrow{\text{a.s.}} E X_\infty, \tag{8.14}$$

and, since

$$I_T = \int_0^T X_t^{-1}\, dX_t = \log X_0^{-1} X_T + \frac{1}{2}\,\sigma^2 J_T$$

using Ito's formula,

$$T^{-1} I_T \xrightarrow{\text{a.s.}} \frac{1}{2}\,\sigma^2 E X_\infty^{-1}$$

as $T \to \infty$. These results readily give the strong consistency of the estimators $\hat{\alpha}_T$, $\hat{\beta}_T$. Asymptotic normality of $T^{\frac{1}{2}}\,(\hat{\alpha}_T - \alpha,\, \hat{\beta}_T - \beta)'$ can be obtained by applying Theorem 2.1, p. 405 of Basawa and Prakasa Rao (1980).

All these results continue to hold under substantially weakened conditions on the noise component in the model. For example, they hold if $\{W_t\}$ is replaced by a square integrable martingale with stationary independent increments $\{Z_t\}$ for which $\langle Z\rangle_t \equiv t$. The details involve straightforward applications of the martingale strong law and central limit theorem and are omitted.

It should be remarked that there is a considerable recent literature concerning estimation of parameters in diffusion type models, partly motivated by burgeoning applications in mathematical finance. See, for example, Bibby and Sørensen (1995), Pedersen (1995), Kessler and Sørensen (1995), Kloeden, Platen, Schurz and Sørensen (1996) and references therein. In most realistic situations the diffusion cannot be observed continuously, so discrete time approximations to stochastic integrals or a direct approach using the discrete time observations is required.

The formulation via (8.6) and (8.7) has to be used with care for problems with multiple sources of variation. Suppose, for example, that the Langevin stochastic differential equation (e.g., Kloeden and Platen (1992, pp. 104–105)) is augmented with jumps coming from a Poisson process and becomes

$$dX_t = \theta X_{t-}\, dt + dW_t + dN_t, \tag{8.15}$$

N_t being a Poisson process with intensity λ. Then, the process may be written in semimartingale form as

$$M_t = W_t + N_t - \lambda t. \tag{8.16}$$

Using (8.15) and (8.16) in (8.7), the quasi-score estimating function based on noise $\{M_t\}$ is

$$\int_0^T (X_{t-}\ 1)'\, dM_t,$$

leading to the estimating equations

$$\begin{aligned} \int_0^T X_{t-}\, dX_t &= \hat{\theta}_T \int_0^T X_t^2\, dt + \hat{\lambda}_T \int_0^T X_t\, dt, \\ X_T &= \hat{\theta}_T \int_0^T X_t^2\, dt + \hat{\lambda}_T T. \end{aligned}$$

These, however, are the maximum likelihood estimating equations for the model

$$dX_t = (\theta X_t + \lambda)\, dt + dW_t,$$

i.e., a version of (8.15) in which N_t has been replaced by its compensator λt. In this model the entire stochastic fluctuation is described by the Brownian process and this is only realistic if $\lambda \ll 1$.

This problem, first noted by Sørensen (1990), can be circumvented and the true maximum likelihood estimators for $\theta,\ \lambda$ obtained if we treat the sources of variation separately. Details are provided in Chapter 2, Section 5.

Models of the above kind are quite common and the general message is to identify relevant (local) martingales that focus on the individual sources of variation. It is then possible to obtain quasi-score estimating functions for each, and to combine them, provided the appropriate sample information is available.

8.4 Estimation in Hidden Markov Random Fields

Considerable recent interest has been shown in hidden Markov models or more generally, partially observed stochastic dynamical system models. The area embodies a wide class of problems in which a random process (or field) is not observed directly but instead is observed subject to noise through a second process (or field).

The techniques that have been developed for this class of problem typically involve full distributional assumptions. Examples are the use of the E-M algorithm (e.g., Qian and Titterington (1990)) or the use of a measure transformation which changes all the signal and noise variables into independent and identically distributed random variables. The monograph Elliott et al. (1994) is devoted to this latter topic and includes a wide range of references to the area.

In this section we shall feature a quite different technique, namely, that of optimal combination of estimating functions. This can be used to provide quick derivations of optimal estimating equations for problems of this type. A representative example taken from Heyde (1996) is given to illustrate the approach in the case of hidden Markov models. It is chosen for its clarity rather than its generality.

Suppose that we observe continuous variables $\{y_i\}$ on a lattice of sites indexed by i. The dimension of the lattice is $d \geq 1$ and the observation at site i depends on the value of its neighbors.

In order to specify the dependence on neighbors we introduce a matrix $\boldsymbol{W} = (w_{ij})$ for which $w_{ii} = 0$, while $w_{ij} = 1$ if i and j are neighbors and is zero otherwise. The matrix $\boldsymbol{W}$ is assumed to be symmetric and $\boldsymbol{w}_i$ denotes the ith column of $\boldsymbol{W}$. We shall write $[i]$ to denote the neighbors of i.

We shall consider the field $\{y_i\}$ that satisfies the spatial autoregression specification

$$y_i = \beta \boldsymbol{w}_i' \boldsymbol{y} + \epsilon_i, \tag{8.17}$$

where

$$E\left(\epsilon_i \,\middle|\, y_{[i]}\right) = 0, \qquad \text{var}\left(y_i \,\middle|\, y_{[i]}\right) = \tau^2$$

and β is scalar. These processes are also known as conditional autoregressions (CAR). See Ripley (1988, p. 11) for a discussion.

Now we are concerned with the context in which $\{y_i\}$ is not observed directly but is hidden in a second field $\{z_i\}$ that is observed on the same lattice. We suppose that the field $\{z_i\}$ satisfies the dynamics

$$z_i = \sum_{k \in \{i \cup [i]\}} b_k y_k + \eta_i, \tag{8.18}$$

where the $\{\eta_i\}$ are uncorrelated with zero mean and finite variance σ^2 and are uncorrelated with the $\{\epsilon_i\}$, while the $\{b_k\}$ are known. We now seek to estimate $\{y_i\}$, assuming β known and subsequently β, on the basis of the observations $\{z_i\}$. Optimal estimation in general is achieved by identifying the sources of variation and optimally combining estimating functions from each of these. Here we have two sources of noise, namely $\{\epsilon_i\}$ and $\{\eta_i\}$ to work with.

As a prelude to obtaining the quasi-score estimating functions corresponding to (8.17) and (8.18), we rewrite both in matrix form as

$$\boldsymbol{y} = \beta \boldsymbol{W} \boldsymbol{y} + \boldsymbol{\epsilon} \tag{8.19}$$

and

$$\boldsymbol{z} = \boldsymbol{B} \boldsymbol{y} + \boldsymbol{\eta}, \tag{8.20}$$

respectively, where $\boldsymbol{y} = (y_i)$, $\boldsymbol{\epsilon} = (\epsilon_i)$, $\boldsymbol{z} = (z_i)$ and $\boldsymbol{B} = (b_{ij})$, where $b_{ij} = b_j$ if $j \in \{i \cup [i]\}$, 0 otherwise.

Now, we find quasi-score estimating functions from the families $\{\boldsymbol{G} = \boldsymbol{A}\boldsymbol{\epsilon}, \ \text{nonrandom matrix } \boldsymbol{A}\}$ and $\{\boldsymbol{H} = \boldsymbol{C}\boldsymbol{\eta}, \ \text{nonrandom matrix } \boldsymbol{C}\}$ of estimating functions. Suppose that $\boldsymbol{G}^* = \boldsymbol{A}^*\boldsymbol{\epsilon}$, $\boldsymbol{H}^* = \boldsymbol{C}^*\boldsymbol{\eta}$ are the required forms.

Applying Theorem 2.1, we find that

$$E\boldsymbol{G}\boldsymbol{G}^{*'} = \boldsymbol{A}E\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}'\boldsymbol{A}^{*'}$$

and

$$E\dot{\boldsymbol{G}} = \left(\boldsymbol{A}E\,\frac{\partial}{\partial \boldsymbol{y}}\,(\boldsymbol{I} - \beta\boldsymbol{W})\boldsymbol{y}\right)' = \boldsymbol{A}(\boldsymbol{I} - \beta\boldsymbol{W})$$

so that

$$(E\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}')\,\boldsymbol{A}^{*'} = (\boldsymbol{I} - \beta\boldsymbol{W}),$$

while

$$E\boldsymbol{H}\boldsymbol{H}^{*'} = \boldsymbol{C}E\boldsymbol{\eta}\,\boldsymbol{\eta}'\boldsymbol{C}^{*'} = \sigma^2\boldsymbol{C}\boldsymbol{C}^{*'}$$

and

$$E\dot{\boldsymbol{H}} = \left(\boldsymbol{C} E \frac{\partial}{\partial \boldsymbol{y}} (\boldsymbol{z} - \boldsymbol{B}\boldsymbol{y})\right)' = -\boldsymbol{C}\boldsymbol{B}$$

so that $\boldsymbol{C}^* = -\boldsymbol{B}'/\sigma^2$.

Since we have taken $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$ as orthogonal we then have the combined estimating function from sources (8.17) and (8.18) as the sum

$$\boldsymbol{A}^*\boldsymbol{\epsilon} + \boldsymbol{C}^*\boldsymbol{\eta} = (\boldsymbol{I} - \beta\boldsymbol{W})'(E\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}')^{-1}(\boldsymbol{I} - \beta\boldsymbol{W})\,\boldsymbol{y} - \boldsymbol{B}'(\boldsymbol{z} - \boldsymbol{B}\boldsymbol{y})/\sigma^2 \tag{8.21}$$

(see Section 6.3).

Now the specification (8.17) gives

$$E\boldsymbol{y}\,\boldsymbol{y}' = \tau^2\,\boldsymbol{I} + \beta\boldsymbol{W}\,E\boldsymbol{y}\,\boldsymbol{y}',$$

i.e.,

$$(\boldsymbol{I} - \beta\boldsymbol{W})\,E\boldsymbol{y}\,\boldsymbol{y}' = \tau^2\boldsymbol{I}$$

and

$$E\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}' = \tau^2(\boldsymbol{I} - \beta\boldsymbol{W})'$$

so that, assuming that $(\boldsymbol{I} - \beta\boldsymbol{W})$ is positive definite,

$$(\boldsymbol{I} - \beta\boldsymbol{W})'(E\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}')^{-1}(\boldsymbol{I} - \beta\boldsymbol{W}) = (\boldsymbol{I} - \beta\boldsymbol{W})/\tau^2$$

and (8.21) becomes

$$((\boldsymbol{I} - \beta\boldsymbol{W})\,\boldsymbol{y}/\tau^2) - (\boldsymbol{B}\,(\boldsymbol{z} - \boldsymbol{B}\boldsymbol{y})/\sigma^2).$$

The quasi-score estimating equation for estimation of $\boldsymbol{y}$ is then

$$(\boldsymbol{I} - \beta\boldsymbol{W})\,\boldsymbol{y}/\tau^2 = \boldsymbol{B}'\,(\boldsymbol{z} - \boldsymbol{B}\boldsymbol{y})/\sigma^2,$$

i.e.,

$$\boldsymbol{y} = \left[((\boldsymbol{I} - \beta\boldsymbol{W})/\tau^2) + (\boldsymbol{B}\,\boldsymbol{B}'/\sigma^2)\right]^{-1}(\boldsymbol{B}'\boldsymbol{z}/\sigma^2)\,. \tag{8.22}$$

The estimating equation (8.22) is of the same as the one derived by Elliott, Aggoun and Moore (1994, Section 9.3) using assumptions of Gaussianity and Girsanov transformation type arguments to change measures and obtain conditional probability densities. The difference in formulation is just that they have begun from a model in Gibbs field form. The quasi-likelihood approach has enabled the specific distributional assumptions and change of measure arguments to be avoided.

Now we suppose that, rather than being known, β is to be estimated. Without specific distributional assumptions ordinary likelihood based methods are not available. However, we can find a quasi-score estimating function from the family

$$\{\boldsymbol{G}(\boldsymbol{S}) = \boldsymbol{\epsilon}'\boldsymbol{S}\,\boldsymbol{\epsilon} - E(\boldsymbol{\epsilon}'\boldsymbol{S}\,\boldsymbol{\epsilon}), \quad \boldsymbol{S} \text{ symmetric matrix}\}$$

to be

$$\boldsymbol{G}(\boldsymbol{S}^*) = \boldsymbol{y}'\boldsymbol{W}\boldsymbol{y} - \operatorname{tr}\boldsymbol{W}(\boldsymbol{I} - \beta\boldsymbol{W})^{-1} \tag{8.23}$$

following the argument of Section 8.2 and under the supposition that each y_i has kurtosis 3. The estimating function (8.23) is, in fact, the score-function under the assumption that $\boldsymbol{y}$ is normally distributed (see, e.g., Lindsay (1988, p. 226)).

Estimation of $\boldsymbol{y}$ and β can then, in principle, be achieved by simultaneous numerical solution (8.22) and (8.23).

Other simpler, albeit less efficient, procedures can be also be found. For example, it is possible to replace (8.23) by the estimating function

$$\boldsymbol{y}'\boldsymbol{W}\boldsymbol{y} - \beta\, \boldsymbol{y}'\boldsymbol{W}^2\boldsymbol{y}; \tag{8.24}$$

as has been mentioned in Section 6.1.

A similar approach to that used can be employed to deal with a wide variety of models, including those for which the different sources of variability are not uncorrelated. In each case the combined quasi-score estimating function can be found via Theorem 2.1. For an example of optimal combination of correlated estimating functions see Chapter 6, Section 3.

It is also worth remarking that the hidden Markov model problem is structurally similar to problems in statistics involving measurement errors where surrogate predictors are used. For a discussion of the use of quasi-likelihood based methods in this context see Carroll and Stefanski (1990).

8.5 Exercise

Suppose that the parameter $\boldsymbol{\theta} = (\boldsymbol{\phi}', \boldsymbol{\psi}')'$, where $\boldsymbol{\phi}$ is of interest and $\boldsymbol{\psi}$ is a nuisance component. For estimating functions $\boldsymbol{G}(\boldsymbol{\theta}) \in \mathcal{H}$ we partition into

$$\boldsymbol{G} = \left(\begin{array}{c} \boldsymbol{G}_\phi \\ \boldsymbol{G}_\psi \end{array} \right) \quad \in \quad \mathcal{H} = \left\{ \begin{array}{c} \mathcal{H}_\phi \\ \mathcal{H}_\psi \end{array} \right\}.$$

Consider the case where the likelihood score $\boldsymbol{U} = (\boldsymbol{U}'_\phi, \boldsymbol{U}'_\psi)'$ is available and write $\boldsymbol{I}_{\alpha\beta} = E(\boldsymbol{U}_\alpha \boldsymbol{U}'_\beta)$ where α, β are ϕ or ψ. Show that the likelihood score based estimating function $\boldsymbol{U}_\phi - \boldsymbol{I}_{\phi\psi}\boldsymbol{I}^{-1}_{\psi\psi}\boldsymbol{U}_\psi$ for $\boldsymbol{\phi}$ is a quasi-score estimating function in $\mathcal{H}_\phi$ if it belongs to $\mathcal{H}_\phi$. (Hint. Use the results of Section 7.3.)

Chapter 9

Hypothesis Testing

9.1 Introduction

In this chapter we discuss hypothesis testing based on the quasi-score estimating function. This is done by developing analogues of standard procedures of classical statistics.

The classical statistical setting for hypothesis testing involves a sequence of T, say, independent random variables whose distribution depends on a p-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)'$ belonging to a sample space Θ, an open subset of p-dimensional Euclidean space $\Re^p$. A null hypothesis H_0 usually involves specification of the $\theta_i,\ i = 1, 2, \dots, p$ to be functions $g_i(\theta_1, \theta_2, \dots, \theta_k)$ of $\boldsymbol{\theta} \in \Re^k$, or specifying restrictions $R_j(\theta_i) = 0$ for $j = 1, 2, \dots, r,\ (r + k = p)$. Tests of H_0 against the full model have typically involved one of the three test statistics:

(a) Likelihood ratio statistic (Neyman-Pearson)

$$\lambda_T = 2\left(L_T(\hat{\boldsymbol{\theta}}_T) - L_T(\tilde{\boldsymbol{\theta}}_T)\right);$$

(b) Efficient score statistic (Rao)

$$\mu_T = S'_T(\tilde{\boldsymbol{\theta}}_T)(I_T(\tilde{\boldsymbol{\theta}}_T))^{-1}\, S_T(\tilde{\boldsymbol{\theta}}_T);$$

(c) Wald test statistic

$$\nu_T = T\, R'(\hat{\boldsymbol{\theta}}_T)\left(W'(\hat{\boldsymbol{\theta}}_T)(I_T(\hat{\boldsymbol{\theta}}_T))^{-1} W(\hat{\boldsymbol{\theta}}_T)\right)^{-1} R(\hat{\boldsymbol{\theta}}_T),$$

which are ordinarily asymptotically equivalent. Here L_T is the log likelihood, $\boldsymbol{R}$ is the vector of restrictions that define H_0, $\boldsymbol{W}(\boldsymbol{\theta})$ is the matrix with elements $\partial R_i/\partial\theta_j$, $\boldsymbol{S}_T(\boldsymbol{\theta})$ is the likelihood score function $\partial L_T(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, $\boldsymbol{I}_T(\boldsymbol{\theta})$ is the Fisher information matrix, $\hat{\boldsymbol{\theta}}_T$ is the MLE for the full model and $\tilde{\boldsymbol{\theta}}_T$ is the MLE under the restrictions imposed by the hypothesis H_0.

Note that the efficient score statistic depends only on the MLE for the restricted class of parameters under H_0, while Wald's statistic depends only on the MLE over the whole parameter space. For further details see Rao (1973, Section 6e).

In the case of hypothesis testing for stochastic processes there is now a reasonably well developed large sample theory based on the use of the likelihood score. See for example Basawa and Prakasa Rao (1980), particularly Chapter

7 and Basawa and Scott (1983, Chapter 3). The use of what are termed generalized M-estimators is discussed in Basawa (1985) in extension of the efficient score and Wald test statistics and these ideas are further developed in Basawa (1991). This path is also followed in the present chapter. However, much of the discussion in the existing literature has involved testing relative to sequences of alternative hypotheses. We shall here avoid this formulation to focus on the special features of the quasi-likelihood setting. The discussion here is based on Thavaneswaran (1991) together with various extensions.

9.2 The Details

Neither the efficient score statistic nor the Wald statistic involve the existence of a scalar objective function (the likelihood) from which an optimal estimating function is derived by differentiation. Each can therefore be generalized straightforwardly to an estimating function setting and this we shall do below. We shall not consider restricted settings, such as that of the conservative quasi-score (Li and McCullagh (1994)) in which a likelihood ratio test generalization is available. The asymptotic properties of such a likelihood ratio test will be the same as those of the efficient score and Wald test generalizations (see the exercise of Section 9.3).

Suppose that $\boldsymbol{Q}(\boldsymbol{\theta})$ is a quasi-score estimating function, for $\boldsymbol{\theta}$, in standardized form, chosen from an appropriate family $\mathcal{H}$ of estimating functions. Here and below we drop the suffix T for convenience. As usual we take

$$\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta}) = \mathcal{E}(\boldsymbol{Q}(\boldsymbol{\theta})) = E(\boldsymbol{Q}(\boldsymbol{\theta})\, \boldsymbol{Q}'(\boldsymbol{\theta}))$$

for the information matrix.

In the following $\boldsymbol{\theta}^*$ denotes the unrestricted quasi-likelihood estimator whereas $\tilde{\boldsymbol{\theta}}$ is the quasi-likelihood estimator calculated under hypothesis H_0.

Most testing problems for stochastic models concern linear hypotheses on the unknown parameter. We shall follow the ideas of Section 7.2 on constrained parameter estimation, and consider testing $H_0 : \boldsymbol{F}'\,\boldsymbol{\theta} = \boldsymbol{d}$ against $H_1 : \boldsymbol{F}'\,\boldsymbol{\theta} \neq \boldsymbol{d}$ where $\boldsymbol{F}$ is a $p \times q$ matrix not depending on the data or $\boldsymbol{\theta}$ with full rank $q \leq p$. Note that this framework includes the problem of testing a subvector of length $q \ (\leq p)$, say $\boldsymbol{\theta}_2$, of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ with $H_0 : \boldsymbol{\theta}_2 = \boldsymbol{\theta}_{02}$ to be tested against $H_1 : \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_{02}$. Just take the block matrix form

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{O}_{(p-q)\times(p-q)} & \boldsymbol{O}_{(p-q),q} \\ \boldsymbol{O}_{q,(p-q)} & \boldsymbol{I}_q \end{bmatrix},$$

where $\boldsymbol{O}_{r,s}$ is the $r \times s$ matrix all of whose elements are zero and, of course, $\boldsymbol{I}_q$ is the $q \times q$ identity matrix.

For the $H_0 : \boldsymbol{F}'\boldsymbol{\theta} = \boldsymbol{d}$ setting and for the ergodic case (see Section 4.3) we have that the efficient score statistic analogue is

$$\mu = \boldsymbol{Q}'(\tilde{\boldsymbol{\theta}})(\mathcal{E}(\boldsymbol{Q}(\tilde{\boldsymbol{\theta}})))^{-}\boldsymbol{Q}(\tilde{\boldsymbol{\theta}}) \tag{9.1}$$

and the Wald test statistic analogue is

$$\nu = (\boldsymbol{F}'\boldsymbol{\theta}^* - \boldsymbol{d})'(\boldsymbol{F}'(\mathcal{E}(\boldsymbol{Q}(\boldsymbol{\theta}^*)))^- \boldsymbol{F})^- (\boldsymbol{F}'\boldsymbol{\theta}^* - \boldsymbol{d}). \tag{9.2}$$

These test statistics are, of course, the classical ones in the case where the quasi-score estimating function is the likelihood score.

We would envisage using the test statistics μ and ν in circumstances under which

$$\boldsymbol{V}^{-\frac{1}{2}}(\boldsymbol{\theta})\,\boldsymbol{Q}(\boldsymbol{\theta}) \overset{\mathrm{d}}{\simeq} MVN(\boldsymbol{0}, \boldsymbol{I}_p) \tag{9.3}$$

and, using Taylor series expansion,

$$\boldsymbol{V}^{\frac{1}{2}}(\boldsymbol{\theta})\,(\boldsymbol{\theta}^* - \boldsymbol{\theta}) \overset{\mathrm{d}}{\simeq} MVN(\boldsymbol{0}, \boldsymbol{I}_p). \tag{9.4}$$

Then, both μ and ν are approximately distributed as χ^2_q under H_0.

From the proof of Theorem 7.2 we have that

$$\boldsymbol{Q}(\tilde{\boldsymbol{\theta}}) \overset{\mathrm{d}}{\simeq} \boldsymbol{P}\,\boldsymbol{Q}(\boldsymbol{\theta}),$$

where $\boldsymbol{P} = \boldsymbol{F}(\boldsymbol{F}'\boldsymbol{V}^{-1}\boldsymbol{F})^-\,\boldsymbol{F}'\,\boldsymbol{V}^{-1}$, so that using (9.3),

$$\boldsymbol{Q}(\tilde{\boldsymbol{\theta}}) \simeq MVN(\boldsymbol{0}, \boldsymbol{P}\,\boldsymbol{V}\,\boldsymbol{P}'),$$

while

$$\mathcal{E}(\boldsymbol{Q}(\tilde{\boldsymbol{\theta}})) \simeq \boldsymbol{P}\,\boldsymbol{V}\,\boldsymbol{P}'$$

so that

$$\begin{aligned} \mu &\overset{\mathrm{d}}{\simeq} (\boldsymbol{P}\,\boldsymbol{Q})'(\boldsymbol{P}\,\boldsymbol{V}\,\boldsymbol{P}')^- \boldsymbol{P}\,\boldsymbol{Q} \\ &= \left(\boldsymbol{V}^{-\frac{1}{2}}\,\boldsymbol{Q}\right)'\,\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{P}'\,(\boldsymbol{P}\,\boldsymbol{V}\,\boldsymbol{P}')^-\,\boldsymbol{P}\,\boldsymbol{V}^{\frac{1}{2}}\,(\boldsymbol{V}^{-\frac{1}{2}}\,\boldsymbol{Q}), \end{aligned}$$

which is approximately distributed as χ^2_q in view of (9.3) and since

$$\boldsymbol{V}^{\frac{1}{2}}\,\boldsymbol{P}'(\boldsymbol{P}\,\boldsymbol{V}\,\boldsymbol{P}')^-\boldsymbol{P}\,\boldsymbol{V}^{\frac{1}{2}}$$

is idempotent with rank q. On the other hand, from (7.2) we have that

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^* - \boldsymbol{V}^{-1}\,\boldsymbol{F}\,(\boldsymbol{F}'\,\boldsymbol{V}^{-1}\,\boldsymbol{F})^-(\boldsymbol{F}'\,\boldsymbol{\theta}^* - \boldsymbol{d}) \tag{9.5}$$

and, using (9.4) and H_0,

$$\boldsymbol{F}'\boldsymbol{\theta}^* - \boldsymbol{d} \overset{\mathrm{d}}{\simeq} MVN(\boldsymbol{0}, \boldsymbol{F}'\,\boldsymbol{V}^{-1}\,\boldsymbol{F}).$$

Then, since

$$\mathcal{E}(\boldsymbol{Q}(\boldsymbol{\theta}^*)) \simeq \mathcal{E}(\boldsymbol{Q}(\boldsymbol{\theta})) = \boldsymbol{V}$$

we have that ν is also approximately distributed as χ^2_q.

Under H_1 we generally have that (9.3) and (9.4) still hold and, for large T, upon expanding $\boldsymbol{Q}(\boldsymbol{\theta})$ about $\boldsymbol{\theta}^*$ to the first order, and assuming that $\dot{\boldsymbol{Q}}$ approximates $-\boldsymbol{V}$ (as in the discussion of the proof of Theorem 7.2),

$$\begin{aligned}\boldsymbol{Q}(\tilde{\boldsymbol{\theta}}) &\simeq -\boldsymbol{V}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\\ &= \boldsymbol{F}(\boldsymbol{F}'\,\boldsymbol{V}^{-1}\,\boldsymbol{F})^{-}(\boldsymbol{F}'\,\boldsymbol{\theta}^* - \boldsymbol{d})\end{aligned}$$

from (9.5). Thus, from (9.4),

$$\boldsymbol{F}'\,\boldsymbol{\theta}^* \overset{\mathrm{d}}{\simeq} MVN(\boldsymbol{F}'\boldsymbol{\theta}, \boldsymbol{F}'\,\boldsymbol{V}^{-1}\,\boldsymbol{F}) \tag{9.6}$$

and then

$$\boldsymbol{Q}(\tilde{\boldsymbol{\theta}}) \overset{\mathrm{d}}{\simeq} MVN(\boldsymbol{F}(\boldsymbol{F}'\,\boldsymbol{V}^{-1}\,\boldsymbol{F})^{-}(\boldsymbol{F}'\boldsymbol{\theta} - \boldsymbol{d}), \boldsymbol{F}(\boldsymbol{F}'\,\boldsymbol{V}^{-1}\,\boldsymbol{F})^{-}\boldsymbol{F}'). \tag{9.7}$$

We also have

$$\mathcal{E}(\boldsymbol{Q}(\boldsymbol{\theta}^*)) \simeq \mathcal{E}(\boldsymbol{Q}(\boldsymbol{\theta})) = \boldsymbol{V} \tag{9.8}$$

and

$$\mathcal{E}(\boldsymbol{Q}(\tilde{\boldsymbol{\theta}})) \simeq \boldsymbol{F}(\boldsymbol{F}'\,\boldsymbol{V}^{-1}\boldsymbol{F})^{-}\boldsymbol{F}'. \tag{9.9}$$

Then, from (9.7) and (9.9) in the case of μ and (9.6) and (9.8) in the case of ν we find that both μ and ν are approximately distributed as noncentral χ^2_q, with noncentrality parameter

$$C(\boldsymbol{V}) = (\boldsymbol{F}'\boldsymbol{\theta} - \boldsymbol{d})'(\boldsymbol{F}'\boldsymbol{V}^{-1}\boldsymbol{F})^{-}(\boldsymbol{F}'\boldsymbol{\theta} - \boldsymbol{d}). \tag{9.10}$$

The test based on μ or ν using the quasi-score $\boldsymbol{Q}$ is optimal in the sense of providing a maximum noncentrality parameter under H_1 for estimating functions within the family $\mathcal{H}$. If another estimating function were chosen from $\mathcal{H}$, then, under appropriate regularity conditions, we would obtain (9.10) with $\boldsymbol{V}$ replaced by $\boldsymbol{V}_1$ say, with $\boldsymbol{V}_1 \leq \boldsymbol{V}$ in the Loewner ordering. It is easily seen that $C(\boldsymbol{V}_1) \leq C(\boldsymbol{V})$. This justifies the choice of the quasi-score as a basis for hypothesis testing.

These results can readily be extended to the non-ergodic case when $\{\boldsymbol{Q}_T\}$ is a martingale. In this setting we replace (9.1) by

$$\mu = \boldsymbol{Q}_T'(\hat{\boldsymbol{\theta}})\,\langle \boldsymbol{Q}(\hat{\boldsymbol{\theta}})\rangle_T^{-}\,\boldsymbol{Q}_T(\hat{\boldsymbol{\theta}}), \tag{9.11}$$

where $\langle \boldsymbol{Q}(\boldsymbol{\theta})\rangle_T$ is the quadratic characteristic (see Section 4.3) and replace (9.2) by

$$\nu = \left(\boldsymbol{F}'\,\boldsymbol{\theta}^* - \boldsymbol{d}\right)'\left(\boldsymbol{F}'\,\langle \boldsymbol{Q}(\boldsymbol{\theta}^*)\rangle_T^{-}\,\boldsymbol{F}\right)^{-}\left(\boldsymbol{F}'\,\boldsymbol{\theta}^* - \boldsymbol{d}\right). \tag{9.12}$$

The role of $\boldsymbol{V}$ in the above discussion is now taken by the (random) quantity $\langle \boldsymbol{Q}\rangle$.

9.3 Exercise

Consider the setting of Section 9.2 in which $H_0 : \boldsymbol{F}'\boldsymbol{\theta} = \boldsymbol{d}$ is to be tested against $H_1 : \boldsymbol{F}'\boldsymbol{\theta} \neq \boldsymbol{d}$. Suppose that $\boldsymbol{Q}(\boldsymbol{\theta})$ is a quasi-score estimating function chosen from some family $\mathcal{H}$ of estimating functions and that there is a scalar function q such that $\boldsymbol{Q}(\boldsymbol{\theta}) = \partial q/\partial\boldsymbol{\theta}$. The natural analogue of the likelihood ratio statistic in this setting is

$$\lambda = 2(q(\boldsymbol{\theta}^*) - q(\boldsymbol{\theta})).$$

Using the first order approximations of this chapter show that

$$\lambda \overset{\mathrm{d}}{\simeq} (\theta^* - \tilde{\theta})' V (\theta^* - \tilde{\theta})$$

and that this is approximately distributed as χ_q^2 under H_0 or as noncentral χ_q^2 with noncentrality parameter (9.10) under H_1.

Chapter 10

Infinite Dimensional Problems

10.1 Introduction

Many nonparametric estimation problems can be regarded as involving estimation of infinite dimensional parameters. This is the situation, for example, in the neuronal membrane potential model

$$dV(t) = (-\rho\, V(t) + \theta(t))\, dt + dM(t)$$

discussed earlier in Section 2.6 with parameters held constant, if changes in $\theta(t)$ over the recording interval $0 \leq t \leq T$ are important and the function needs to be estimated. Both the situation where n replica trajectories of $\{V(t),\ 0 \leq t \leq T\}$ are observed and $n \to \infty$ or one trajectory is observed and $T \to \infty$ are of interest but most of the established results deal with the former case.

Various possible estimation procedures have been developed for use in such problems including kernel estimation and the method of sieves. Efforts have also been made to develop the quasi-likelihood methodology in a Banach space setting to deal directly with infinite dimensional problems (e.g., Thompson and Thavaneswaran (1990)) but the results are of limited scope and no asymptotic analysis has been provided. The topic is of sufficient intrinsic importance to be dignified with a chapter, albeit brief and more suggestive than definitive. Here the quasi-likelihood methodology described elsewhere in the book does not play a central role. We shall first sketch the approach via the method of sieves and then offer some heuristic discussion related to particular problems.

10.2 Sieves

The method of sieves was first developed by Grenander (1981) as a natural extension of the finite dimensional theory. It is ordinarily applied in the context where there is replication of some basic process. We use as a sieve a suitably chosen complete orthonormal sequence $\{\phi_i(t)\}$, say, and approximate the unknown function $\theta(t)$ by

$$\theta^{(n)}(t) = \sum_{k=1}^{n} \theta_i\, \phi_i(t),$$

where n increases to infinity as the sample size increases. Then, the θ_i are estimated, say by quasi-likelihood, to give an optimal estimate $\theta^{(n)*}(t)$ of $\theta^{(n)}(t)$. The remaining task is to show that as $n \to \infty$, $\theta^{(n)*}(t)$ is consistent for $\theta(t)$ and, hopefully, asymptotically normally distributed.

Considerable recent literature has been devoted to such studies. For example, Nguyen and Pham (1982) established consistency of the sieve estimator of the drift function in a linear diffusion model. McKeague (1986) dealt with a general semimartingale regression model and obtained a strong consistency result for certain sieve estimators. Sieve estimation problems for point processes, such as in the multiplicative intensity model, have been studied by Karr (1987), Leskow (1989) and Leskow and Rozanski (1989). Kallianpur and Selukar (1993) have provided a general discussion of sieve estimation where the focus is on maximum likelihood for the nested finite dimensional problems. Rate of convergence results for sieve estimators are discussed in Shen and Wong (1994). The method is a valuable one but it does suffer from the disadvantage of lack of exlicit expressions and the problem in practice of trade-offs between approximation error and estimation error. Much remains to be done and a general discussion, based on the use of quasi-likelihood for the nested finite dimensional problems, is awaited.

10.3 Semimartingale Models

Here we shall consider semimartingale models of the form

$$dX_t = dA_t + dM_t,$$

where $\{X_t\}$ is the observation process, $\{M_t\}$ is a square integrable martingale and the predictable bounded variation process $\{A_t\}$ is of linear form

$$dA_t = \theta_t \, dB_t$$

in the unknown function θ_t of interest, $\{B_t\}$ being an observable process, possibly a covariate or a function of the observation process, such as $dB_t = X_{t-}dt$.

Observations on the process typically involve a history over an extended time interval $0 \le t \le T$, say, or we have copies $X_{1t}, \dots, X_{nt}$, say, of the process over a fixed time interval, which we can take as $[0, 1]$. In any case, we seek estimators that, in a suitable sense, are consistent as $T \to \infty$ or $n \to \infty$.

For models of the type considered here, the semimartingale representation immediately suggests a simple estimation procedure. If we have n histories, write $\boldsymbol{X}_t = (X_{1t}, \dots, X_{nt})'$, $\boldsymbol{M}_t = (M_{1t}, \dots, M_{nt})'$ and $\boldsymbol{B}_t = (B_{1t}, \dots, B_{nt})'$. Then,

$$d\boldsymbol{X}_t = \theta_t \, d\boldsymbol{B}_t + d\boldsymbol{M}_t,$$

so that if $\mathbf{1}$ is the $n \times 1$ vector, all of whose elements are unity,

$$\int_0^T \theta_s \, ds = \int_0^T (\mathbf{1}' \, d\boldsymbol{B}_s)^{-1} \, ds \, (\mathbf{1}' \, d\boldsymbol{X}_s) - \int_0^T (\mathbf{1}' \, d\boldsymbol{B}_s)^{-1} \, ds \, (\mathbf{1}' \, d\boldsymbol{M}_s),$$

and

$$\beta_T = \int_0^T \theta_s \, ds \tag{10.1}$$

is unbiasedly estimated by

$$\hat{\beta}_T = \int_0^T (\mathbf{1}'\, d\boldsymbol{B}_s)^{-1}\, ds\, (\mathbf{1}'\, d\boldsymbol{X}_s). \tag{10.2}$$

In a certain formal sense, θ_t is estimated by $(\mathbf{1}'\, d\boldsymbol{B}_t)^{-1}\, (\mathbf{1}'\, d\boldsymbol{X}_t)$, the derivative of a generally (pointwise) non-differentiable process.

This approach is convenient but of course it does not readily lead to an estimator of θ_t, but rather of its indefinite integral β_t.

Since

$$\hat{\beta}_T - \beta_T = \int_0^T (\mathbf{1}'\, d\boldsymbol{B}_s)^{-1}\, ds\, (\mathbf{1}'\, d\boldsymbol{M}_s) \tag{10.3}$$

is a square integrable martingale, we obtain from the Burkholder, Davis, Gundy inequality (e.g., Theorem 4.2.1, p. 93 of Rogers and Williams (1987)) that for C a universal constant,

$$\begin{aligned} E\left[\sup_{T\in[0,1]} (\hat{\beta}_T - \beta_T)^2\right] &\le C\, E(\hat{\beta}_1 - \beta_1)^2 \\ &= C \left\langle \int_0^{\cdot} (\mathbf{1}'\, d\boldsymbol{B}_s)^{-1}\, ds\, (\mathbf{1}'\, d\boldsymbol{M}_s) \right\rangle_1 \\ &= C \int_0^1 (\mathbf{1}'\, d\boldsymbol{B}_s)^{-2}\, (ds)^2\, d\langle \mathbf{1}'\, \boldsymbol{M}\rangle_s \end{aligned}$$

so we have consistency if

$$\int_0^1 (\mathbf{1}'\, d\boldsymbol{B}_s)^{-2}\, (ds)^2\, d\langle \mathbf{1}'\, \boldsymbol{M}\rangle_s \to 0$$

as $n \to \infty$. Various central limit results can also be formulated on the basis of (10.3).

On the other hand, if we have a single history it is usually the case that

$$\int_0^T (dB_s)^{-1}\, ds\, dX_s \Big/ \int_0^T \theta_s\, ds \xrightarrow{\text{a.s.}} 1 \tag{10.4}$$

as $T \to \infty$. Note that

$$\frac{\int_0^T (dB_s)^{-1}\, ds\, dX_s}{\int_0^T \theta_s\, ds} = 1 + \frac{\int_0^T (dB_s)^{-1}\, ds\, dM_s}{\int_0^T \theta_s\, ds},$$

and for the martingale

$$N_t = \int_0^T (dB_s)^{-1}\, ds\, dM_s,$$

we have

$$d\langle N\rangle_t = (dB_t)^{-2}\,(dt)^2\,d\langle M\rangle_t$$

and, so long as $\langle N\rangle_T \to \infty$ a.s. as $T \to \infty$, the martingale strong law (e.g., Theorem 12.5) ensures that $N_T/\langle N\rangle_T \xrightarrow{\text{a.s.}} 0$, i.e. (10.4) holds if

$$\limsup_{T\to\infty} \int_0^T (dB_s)^{-2}\,(ds)^2\,d\langle M\rangle_s \Big/ \int_0^T \theta_s\,ds < \infty \quad \text{a.s.} \tag{10.5}$$

This is easily checked in practice for simple models. For example, if $dB_s = ds$ and $d\langle M\rangle_s = ds$ we require

$$\limsup_{T\to\infty} T \Big/ \int_0^T \theta_s\,ds < \infty \quad \text{a.s.}$$

Sometimes minor adjustments are necessary, as for example in the case of the Aalen (1978) model for nonparametric estimation of the cumulative hazard function using censored lifetime data. In this setup X_t is a counting process representing the number of deaths up to time t, $dB_t = \theta_t\,Y_t\,dt$, where Y_t is the number at risk at time t (possibly zero) and θ_t is the hazard function of the lifetime distribution. Then, $d\langle M\rangle_s = \theta_s\,Y_s\,ds$, and using I to denote the indicator function,

$$\frac{\int_0^T I(Y_s > 0)\,Y_s^{-1}\,dX_s}{\int_0^T I(Y_s > 0)\,\theta_s\,ds} = 1 + \frac{\int_0^T I(Y_s > 0)\,Y_s^{-1}\,dM_s}{\int_0^T I(Y_s > 0)\,\theta_s\,ds},$$

and provided that its denominator goes a.s. to infinity.

$$\frac{\int_0^T I(Y_s > 0)\,Y_s^{-1}\,dM_s}{\langle \int_0^{\cdot} I(Y_s > 0)\,Y_s^{-1}\,dM_s\rangle_T} = \frac{\int_0^T I(Y_s > 0)\,Y_s^{-1}\,dM_s}{\int_0^T I(Y_s > 0)\,Y_s^{-1}\,\theta_s\,ds} \xrightarrow{\text{a.s.}} 0$$

using the above mentioned martingale strong law. The consistency result

$$\int_0^T I(Y_s > 0)\,Y_s^{-1}\,dX_s \Big/ \int_0^T I(Y_s > 0)\,\theta_s\,ds \xrightarrow{\text{a.s.}} 1$$

as $T \to \infty$ then follows since $Y_s \ge 1$ on $I(Y_s > 0)$.

It is possible to use the quasi-likelihood theory of this book if, for example, θ_t is assumed to be a step function with jumps at $0 = t_0 < t_1 < \ldots < t_{k-1} < t_k = T$, namely, $\theta_t = \theta_j,\ t_{j-1} \le t < t_j,\ j = 1, 2, \ldots, k$. Then, postulating the model

$$d\tilde{X}_t = \left(\sum_{j=1}^{k} \theta_j\, I(t_{j-1} \le t < t_j) \right) d\tilde{B}_t + d\tilde{M}_t,$$

say, the Hutton-Nelson quasi-score for estimation of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ is

$$\int_0^T \text{diag}\,(I(t_0 \le t < t_1), \ldots, I(t_{k-1} \le t < t_k))\,\frac{d\tilde{B}_t}{d\langle\tilde{M}\rangle_t}\,d\tilde{M}_t$$

from which we see that

$$\hat{\theta}_j = \int_{t_{j-1}}^{t_j} \frac{d\tilde{B}_t}{d\langle \tilde{M} \rangle_t}\, d\tilde{X}_t \Big/ \int_{t_{j-1}}^{t_j} \frac{(d\tilde{B}_t)^2}{d\langle \tilde{M} \rangle_t}, \quad 1 \le j \le k, \tag{10.6}$$

are the corresponding quasi-likelihood estimators if $\langle \tilde{M} \rangle_t$ does not involve the θ's.

If, on the other hand,

$$d\langle \tilde{M} \rangle_t = \theta_t\, d\tilde{B}_t$$

as for counting process models, then we find that

$$\hat{\theta}_j = \frac{\tilde{X}_{t_j} - \tilde{X}_{t_{j-1}}}{\tilde{B}_{t_j} - \tilde{B}_{t_{j-1}}}, \quad 1 \le j \le k, \tag{10.7}$$

for the quasi-likelihood estimators. One can envisage approximating the model

$$dX_t = \theta_t\, dB_t + dM_t, \tag{10.8}$$

observed over $0 \le t \le T$, by a sequence of models such as

$$d\tilde{X}_t = \sum_{j=1}^{N} \theta_j\, I\left(\frac{j-1}{N} \le t < \frac{j}{N}\right) d\tilde{B}_t + d\tilde{M}_t$$

where $N \to \infty$. This idea is formalized in the theory of the histogram sieve.

An alternative approach would be to use a discrete model approximation to (10.8), such as provided by the Euler scheme, say

$$\tilde{X}_{j\Delta} - \tilde{X}_{(j-1)\Delta} = \theta_j \left(\tilde{B}_{j\Delta} - \tilde{B}_{(j-1)\Delta}\right) + \left(\tilde{M}_{j\Delta} - \tilde{M}_{(j-1)\Delta}\right), \tag{10.9}$$

$j = 1, 2, \ldots, [T/\Delta]$, $[x]$ denoting the integer part of x.

Here the Hutton-Nelson quasi-score leads to the estimators

$$\hat{\theta}_j = \frac{\tilde{X}_{j\Delta} - \tilde{X}_{(j-1)\Delta}}{\tilde{B}_{j\Delta} - \tilde{B}_{(j-1)\Delta}}, \qquad 1 \le j \le [T/\Delta], \tag{10.10}$$

regardless of the form of the quadratic characteristic of the martingale $\{\tilde{M}_t\}$. It should be noted that (10.7) and (10.10) accord with the interpretation mentioned above of the estimator of θ_t as dX_t/dB_t. They are straightforward to deal with numerically.

For asymptotic results we need, for example, replication as indicated above and it is again worth remarking on the model approximation error and the estimation error. It can ordinarily be expected that for Euler schemes the error in model approximation is $O(\Delta^{\frac{1}{2}})$ (see, e.g., Section 9.6 of Kloeden and Platen (1992) which readily extends beyond the diffusion context). On the other hand, if n replicates are being used, the estimation error is $O(n^{\frac{1}{2}})$ (central limit rate).

Chapter 11

Miscellaneous Applications

11.1 Estimating the Mean of a Stationary Process

A method of last resort, when more sophisticated methods are not tractable or not available, is the method of moments, which typically relies on an ergodic theorem and involves no significant structural assumptions. Nevertheless, in various circumstances, such as for estimating the mean of a stationary process, it turns out that the method of moments (which estimates $EX_i = \theta$ by $T^{-1}\sum_{i=1}^{T} X_i$) produces an estimator that has the same asymptotic variance as the best linear unbiased estimator (BLUE) under broad conditions on the spectral density or covariance function. This is of considerable practical significance. The discussion in this section follows the papers Heyde (1988b), (1992b).

Let $\{X_t,\ t = \dots, -1, 0, 1, \dots\}$ be a stationary ergodic process with $EX_t = \theta$, which is to be estimated, and $EX_t^2 < \infty$. Write

$$\gamma(t-s) = \operatorname{cov}(X_s - \theta, X_t - \theta) = \int_{-\pi}^{\pi} e^{i(s-t)\lambda} f(\lambda)\, d\lambda$$

for the covariance function, where

$$f(\lambda) = (2\pi)^{-1}\left\{\gamma(0) + 2\sum_{j=1}^{\infty}\gamma(j)\cos j\lambda\right\},\ \lambda \neq 0,\ |\lambda| \leq \pi,$$

is the spectral density, the spectral function being assumed to be absolutely continuous. This is the case for all purely nondeterministic (i.e., contain no component which is exactly predictable) processes (e.g., Hannan (1970, Theorem 3′, p. 154)).

Our concern here is to estimate θ on the basis of observations $(X_1, \dots, X_T)$ and we shall first restrict consideration to the class of estimating functions

$$\mathcal{H}_1 = \left\{\sum_{i=1}^{T} a_{i,T}(X_i - \theta), \quad a_{i,T} \text{ constants}, \sum_{i=1}^{T} a_{i,T} = 1\right\}.$$

Then, if $G_T = \sum_{i=1}^{T} a_{i,T}(X_i - \theta)$, $G_T^* = \sum_{i=1}^{T} a_{i,T}^*(X_i - \theta)$, we find that

$$E\dot{G}_T = -1, \qquad EG_T G_T^* = \sum_{i=1}^{T}\sum_{j=1}^{T} a_{i,T}^*\, a_{j,T}\, \gamma(j-i)$$

and, using Theorem 2.1, G_T^* is a quasi-score estimating function within $\mathcal{H}$, if

$$\sum_{i=1}^{T} \gamma(j-i)\, a_{i,T}^* = c_T \text{ (constant)}, \quad j = 1, 2, \ldots, T. \tag{11.1}$$

In this case the quasi-score estimating function leads to a quasi-likelihood estimator that is just the best linear unbiased estimator (BLUE) θ_{MV} possessing minimum asymptotic variance. To see this, note that

$$\begin{aligned} EG_T^2 - EG_T^{*2} &= E(G_T - G_T^*)^2 + 2EG_T^*(G_T - G_T^*) \\ &= E(G_T - G_T^*)^2 \geq 0. \end{aligned}$$

Of course the calculation of the BLUE involves a full knowledge of the covariance structure of the process so that it is ordinarily not feasible to use it in practice.

To obtain an expression for the minimum variance c_T, we write $\boldsymbol{\Gamma}(T)$ for the $T \times T$ matrix $(\gamma(i-j))$ and let $\boldsymbol{\alpha}^*(T)$ and $\boldsymbol{e}(T)$ denote the T-vectors $(\alpha_{1,T}^*, \ldots, \alpha_{T,T}^*)'$ and $(1, 1, \ldots, 1)'$, respectively, the prime denoting transpose. Then, (11.1) gives

$$\boldsymbol{\Gamma}(T)\, \boldsymbol{\alpha}^*(T) = c_T\, \boldsymbol{e}(T),$$

while

$$(\boldsymbol{e}(T))' \boldsymbol{\alpha}^*(T) = 1,$$

and hence

$$c_T = [(\boldsymbol{e}(T))'(\boldsymbol{\Gamma}(T))^{-1}\, \boldsymbol{e}(T)]^{-1}. \tag{11.2}$$

The explicit expressions for $\boldsymbol{\alpha}^*(T)$ and c_T are, however, of limited practical value as, even when the covariances $\gamma(i)$ are known, they involve calculation of the inverse of the Toeplitz matrix $\boldsymbol{\Gamma}(T)$, which will usually be of high order.

On the other hand, the *method of moments* estimator

$$\theta_M = T^{-1} \sum_{i=1}^{T} X_i$$

uses equally weighted observations and requires no direct knowledge of the underlying distribution. We shall examine conditions under which

$$\operatorname{var} \theta_{MV} = [(\boldsymbol{e}(T))'\, (\boldsymbol{\Gamma}(T))^{-1}\, \boldsymbol{e}(T)]^{-1} \tag{11.3}$$

and

$$\operatorname{var} \theta_M = T^{-2}\, (\boldsymbol{e}(T))'\, \boldsymbol{\Gamma}(T)\, \boldsymbol{e}(T) \tag{11.4}$$

have the same asymptotic behavior. That is, the estimator θ_M is an asymptotic quasi-score estimating function.

Theorem 11.1 Suppose that the density $f(\lambda)$ of $\{X_t\}$ is continuous and positive at $\lambda = 0$. Then

$$T \operatorname{var} \theta_M \to 2\pi f(0) \quad \text{as} \quad T \to \infty. \tag{11.5}$$

If, in addition, $\int_{-\pi}^{\pi} (|p(\lambda)|^2 / f(\lambda))\, d\lambda < \infty$ for some trigonometric polynomial $p(\lambda)$, then

$$T \operatorname{var} \theta_{MV} \to 2\pi f(0) \quad \text{as} \quad T \to \infty, \tag{11.6}$$

and

$$T^{\frac{1}{2}} (\theta_{MV} - \theta_M) \xrightarrow{\text{p}} 0 \quad \text{as} \quad T \to \infty, \tag{11.7}$$

p denoting convergence in probability.

Proof The result (11.5) is a well-known consequence of the representation

$$\operatorname{var} \theta_M = \frac{1}{T^2} \int_{-\pi}^{\pi} \left[\frac{\sin(T\lambda/2)}{\sin(\lambda/2)} \right]^2 f(\lambda)\, d\lambda;$$

see, for example, Ibragimov and Linnik (1971, p. 322). The result (11.6) follows from Theorem 3 of Adenstadt and Eisenberg (1974).

To obtain (11.7) we note that

$$\begin{aligned}
T\, E(\theta_{MV} - \theta_M)^2 &= T\, E \left[\sum_{i=1}^{T} (\alpha^*_{i,T} - T^{-1})\, X_i \right]^2 \\
&= T \sum_{i=1}^{T} (\alpha^*_{i,T} - T^{-1}) \sum_{j=1}^{T} (\alpha^*_{j,T} - T^{-1})\, \gamma(i-j) \\
&= T \sum_{i=1}^{T} (\alpha^*_{i,T} - T^{-1}) \left[c_T - T^{-1} \sum_{j=1}^{T} \gamma(i-j) \right] \\
&= - \sum_{i=1}^{T} (\alpha^*_{i,T} - T^{-1}) \sum_{j=1}^{T} \gamma(i-j) \\
&= T(\operatorname{var} \theta_M - \operatorname{var} \theta_{MV}) \to 0
\end{aligned}$$

as $T \to \infty$ using (11.5) and (11.6) and the result follows. This completes the proof.

It should be noted from (11.7) that if a central limit result holds for $T^{\frac{1}{2}} (\theta_M - \theta)$, then the corresponding one also holds for $T^{\frac{1}{2}} (\theta_{MV} - \theta)$. Various conditions of asymptotic independence lead to the result

$$T^{\frac{1}{2}} (\theta_M - \theta) = T^{-\frac{1}{2}} \sum_{i=1}^{T} (X_i - \theta) \xrightarrow{\text{d}} N(0, 2\pi\, f(0)),$$

namely, convergence in distribution to the normal law with zero mean and variance $2\pi f(0)$. For example, this holds if there is a σ-field $\mathcal{M}_0$ such that X_0 is $\mathcal{M}_0$-measurable and

$$\sum_{k=0}^{\infty} E|E(X_k \,|\, \mathcal{M}_0) - \theta| < \infty$$

(Hall and Heyde (1980, Theorem 5.4, p. 136)). This is a very general result from which many special cases, such as for mixing sequences, follow.

For the results of Theorem 11.1 to be useful it is essential that $f(0) > 0$. An example where this may not be the case is furnished by the moving average model

$$X_t = \theta + \epsilon_t - r\epsilon_{t-1}, \quad |r| \le 1,$$

where the ϵ_i are stationary and orthogonal with mean 0 and variance σ^2. Here

$$f(\lambda) = \sigma^2(1 + r^2 - 2r\cos\lambda)/2\pi,$$

so that $f(0) > 0$ requires $r < 1$. If $r = 1$ it turns out that (Grenander and Szegö (1958, p. 211)),

$$\operatorname{var}\theta_M \sim 2\sigma^2 T^{-2}, \quad \operatorname{var}\theta_{MV} \sim 12\sigma^2 T^{-3}$$

as $T \to \infty$, so that the method of moments estimator is far from efficient in this case. In fact, if $f(0) = 0$, it holds rather generally that $\operatorname{var}\theta_{MV} = o(\operatorname{var}\theta_M)$ as $T \to \infty$ (Adenstadt (1974), Vitale (1973)).

Now Theorem 11.1 can be used to show that θ_{MV} and θ_M have the same asymptotic variance if the spectral density $f(\lambda)$ is everywhere positive and continuous. Such information is ordinarily not directly available but a useful sufficient condition in terms of covariances is given in the following theorem.

Theorem 11.2 If the covariance $\gamma(T)$ decreases monotonically to zero as $T \to \infty$, then $\{X_t\}$ has a spectral function that is absolutely continuous. The spectral density $f(\lambda)$ is nonnegative and continuous, except possibly at zero, being continuous at zero when $\sum_{j=1}^{\infty} \gamma(j) < \infty$. If $\{\gamma(n)\}$ is convex and $\gamma(0) + \gamma(2) > 2\gamma(1)$, then the spectral density is strictly positive.

Proof The first two parts of the theorem are given in Theorem 3 of Daley (1969). For the last part it should be noted that when $\{\gamma(n)\}$ is convex, namely,

$$\Delta^2\gamma(n) = \gamma(n) + \gamma(n+2) - 2\gamma(n+1) \ge 0, \quad n \ge 0,$$

the spectral density $f(\lambda)$ can be expressed in the form

$$f(\lambda) = (2\pi)^{-1} \sum_{n=0}^{\infty} (n+1)\Delta^2\,\gamma(n)\,K_n(\lambda), \quad |\lambda| \le \pi,$$

where

$$K_n(\lambda) = \frac{2}{n+1}\left[\frac{\sin((n+1)\lambda/2)}{2\sin(\lambda/2)}\right]^2 \geq 0$$

is the Fejér kernel (Zygmund (1959, p. 183)). Then, since $K_0(\lambda) = \frac{1}{2}$ and $\Delta^2\gamma(n)\,K_n(\lambda) \geq 0$ for $n \geq 1$, $f(\lambda) > 0$ if $\Delta^2\gamma(0) > 0$ as required.

Corollary 11.1 If the covariance $\gamma(T)$ decreases monotonically to zero as $T \to \infty$, $\sum_{j=1}^{\infty}\gamma(j) < \infty$ and $\{\gamma(n)\}$ is convex with $\gamma(0) + \gamma(2) > 2\gamma(1)$, then

$$\operatorname{var}\theta_{MV} \sim \operatorname{var}\theta_M \sim T^{-1}\left[\gamma(0) + 2\sum_{j=1}^{\infty}\gamma(j)\right]$$

as $T \to \infty$.

This result follows immediately from Theorem 11.1 and 11.2.

The particular case of the result of Corollary 11.1 for which the covariance function $\gamma(n)$ is of the form

$$\gamma(n) = \sum_{i=1}^{K} a_i\, e^{-r_i|n|} \tag{11.8}$$

where $a_i > 0$, $i = 1, 2, \ldots, K$, $0 < r_1 < r_2 < \ldots, r_K$, has been obtained by Halfin (1982). Covariance functions of the type (11.8) are common for Markovian queueing models with limited queue space and various applications are given in Halfin's paper.

Corollary 11.1, however, is very much more widely applicable and many examples can be noted from the review article of Reynolds (1975) on the covariance structure of queues and related processes. As one example, we mention the queue length process in an $M/G/\infty$ queue (Reynolds (1975, p. 386)) where

$$\rho(n) = \gamma(n)/\gamma(0) = \beta\int_n^{\infty} G(x)\,dx$$

with $\beta^{-1} = ES$, the mean service time and

$$G(x) = P(S \geq x).$$

Then, it is easily checked that $\sum_{n=0}^{\infty}\gamma(n) < \infty$ is equivalent to $ES^2 < \infty$, while convexity of $\{\gamma(n)\}$ is automatically satisfied and $\gamma(0) + \gamma(2) > 2\gamma(1)$ requires

$$\int_0^1 G(x)\,dx > \int_1^2 G(x)\,dx.$$

Another example involves the waiting time $\{W_n\}$ in an $M/G/1$ queue where (Reynolds (1975, p. 401))

$$\rho(n) = 1 - \left[n\theta(1-\rho) - \left[\sum_{r=1}^{n} C_r\right]^2\right]\Big/\alpha\sigma^2, \quad n \geq 1,$$

for certain constants $C_r \uparrow$, with $\theta = EW_n$, $\sigma^2 = \operatorname{var} W_n$, ρ the traffic intensity and α the mean arrival rate. Here it is again easily checked that $\{\rho(n)\}$ is monotone and convex with

$$\rho(0) + \rho(2) > 2\rho(1)$$

while $\sum_{n=0}^{\infty} \rho(n) < \infty$ if and only if $ES^4 < \infty$.

Now if the spectral density has zeros, Theorem 11.1 allows for their removal by $|p(\lambda)|^2$. This can be done, for example, for the queue length process in an $M/D/\infty$ queue where the constant service time $T = 2$.

For the $M/D/\infty$ queue with constant service time S, $\rho(j) = \gamma(j)/\gamma(0)$ has the particularly simple form

$$\rho(j) = \begin{cases} 1 - |j|/S, & |j| \le S, \\ 0, & \text{otherwise} \end{cases}$$

(Reynolds (1975, p. 387)) and $\gamma(j) = \alpha T^2 \rho(j)$ where α denotes the mean arrival rate. When $S = 2$, the spectral density is given by

$$2\pi f(\lambda) = 4\alpha(1 + \cos\lambda)$$

and we find that

$$T \operatorname{var} \theta_{MV} \to 8\alpha, \quad T \operatorname{var} \theta_M \to 8\alpha \tag{11.9}$$

as $T \to \infty$ despite the fact that $f(-\pi) = f(\pi) = 0$.

Now a crucial property that underlies the parity of (11.3) and (11.4) is that of short-range dependence. It is said that short-range or long-range dependence holds according as $\sum_{j=1}^{\infty} \gamma(j)$ converges or diverges or, essentially equivalently, the spectral density $f(\lambda)$ converges or diverges at $\lambda = 0$.

For long-range dependence, $T \operatorname{var} \bar{X}_T$ ordinarily diverges as $T \to \infty$ in contrast to the short-range dependence case where convergence usually holds. Indeed, rather typical of the behavior in the long-range dependence case is the convergence in distribution of $T^{\alpha+1/2}(\bar{X}_T - \theta)$ to a proper limit law as $T \to \infty$ where $-\frac{1}{2} < \alpha < 0$. The limit law is not Gaussian in general, save when the stationary process $\{X_t\}$ is itself Gaussian. Little is known in general about the asymptotic behavior of θ_{MV}. This should be contrasted with the case of short-range dependence where $T^{1/2}(\bar{X}_T - \theta)$ converges to the normal law $N(0, 2\pi f(0))$ under very broad conditions and then a similar result holds for $T^{1/2}(\theta_{MV} - \theta)$ as readily follows from (11.7). These results are vitally important as they provide the basis for asymptotic confidence statements about θ.

Efficiency comparisons of $\bar{X}$ and θ_{MV} for the long-range dependence case appear in Samarov and Taqqu (1988), while Beran (1989) has provided confidence intervals (under Gaussian assumptions) that take the estimation of the parameter α into account. The relevance of all this, however, rests critically on the assumption of Gaussianity because efficiency calculations based on variances may be of little use in assessing sizes of confidence zones save when

Gaussian limits obtain. Gaussian limits would not be typical for queueing models under long-range dependence.

The parity of (11.5) and (11.6) for short-range dependence also breaks down in the case where the spectral density $f(\lambda)$ has a zero at $\lambda = 0$. Again efficiency comparisons can sometimes be made, for example in the moving-average model considered earlier, but the problem remains that asymptotic normality does not hold for the estimators in general.

11.2 Estimation for a Heteroscedastic Regression

The model we consider in this section is the general linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{u}, \tag{11.10}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of observations, $\boldsymbol{X}$ is an $n \times k$ matrix of known constants of rank $k < n$, $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters and $\boldsymbol{u}$ is an $n \times 1$ vector of independent residuals with zero mean and covariance matrix

$$\boldsymbol{\Omega} = \operatorname{diag}\left(g_1(\boldsymbol{\theta}), \dots, g_n(\boldsymbol{\theta})\right),$$

where $g_i(\boldsymbol{\theta}) = \sigma^2(\boldsymbol{X}_i\boldsymbol{\beta})^{2(1-\alpha)}$, $\boldsymbol{X}_t = (X_{t1}, \dots, X_{tk})$, $(t = 1, \dots, n)$, and $\boldsymbol{X} = (\boldsymbol{X}_1' \vdots \cdots \vdots \boldsymbol{X}_n')'$. Variances proportional to a power of expectations have been widely observed, for example in cross-sectional data on microeconomic units such as firms or households. The object is the efficient estimation of the $(k+2) \times 1$ vector $\boldsymbol{\theta} = (\boldsymbol{\beta}'\,\sigma^2\alpha)'$. The results here are from Heyde and Lin (1992).

The obvious martingales to use for the model (11.10) are $\{U_t\}$ and $\{V_t\}$ given by

$$U_t = \sum_{s=1}^{t} u_s, \qquad V_t = \sum_{s=1}^{t} (u_s^2 - g_s(\boldsymbol{\theta})).$$

Then, the quasi-score estimating function based on $\{U_t\}$ is

$$\sum_{t=1}^{n} (\boldsymbol{X}_t \ \ 0 \ \ 0)'\, g_t^{-1}(\boldsymbol{\theta})\, u_t = \left(\boldsymbol{u}'\boldsymbol{\Omega}^{-1}\boldsymbol{X} \ \ \vdots \ \ 0 \ \ 0 \right)', \tag{11.11}$$

which focuses on $\boldsymbol{\beta}$, provides no information on σ^2 and involves α as a nuisance parameter. On the other hand, the quasi-score estimating function based on $\{V_t\}$ is

$$\sum_{t=1}^{n} (\dot{g}_t(\boldsymbol{\theta}))'(\operatorname{var} u_t^2)^{-1}(u_t^2 - g_t(\boldsymbol{\theta})) = \left(\frac{\partial \operatorname{vec} \boldsymbol{\Omega}}{\partial \boldsymbol{\theta}} \right)' \operatorname{vec} \boldsymbol{F}^{-1}(\boldsymbol{\Omega} - \boldsymbol{u}\,\boldsymbol{u}'), \tag{11.12}$$

where

$$\boldsymbol{F} = \operatorname{diag}(\operatorname{var} u_1^2, \ldots, \operatorname{var} u_n^2),$$

which allows estimation of all components of $\boldsymbol{\theta}$ provided $\boldsymbol{F}$ is a known function of $\boldsymbol{\theta}$. However, efficiency in the estimation is enhanced if the martingales $\{U_t\}$ and $\{V_t\}$ are used in combination and, using results of Chapter 6 (e.g., (6.5)), the combined quasi-score estimating function is

$$\sum_{t=1}^{n} \frac{(X_t\ 0\ 0)'\,[u_t + R_t\,(u_t^2 - g_t(\boldsymbol{\theta}))]}{g_t(\boldsymbol{\theta})(1 - R_t S_t)} + \sum_{t=1}^{n} \frac{(\dot{g}_t(\boldsymbol{\theta}))'[S_t\, u_t + (u_t^2 - g_t(\boldsymbol{\theta}))]}{(\operatorname{var} u_t^2)(1 - R_t S_t)}, \tag{11.13}$$

where $R_t = E u_t^3 / \operatorname{var} u_t^2$, $S_t = E u_t^3 / g_t(\boldsymbol{\theta})$. This simplifies considerably when $E u_t^3 = 0$ for each t, for then the martingales $\{U_t\}$ and $\{V_t\}$ are orthogonal and (11.13) is a sum of the quasi-score estimating functions (11.11) and (11.12). In this case we may separate the corresponding estimating equation into the two components

$$\begin{aligned} \boldsymbol{X}'\boldsymbol{\Omega}'\boldsymbol{u} + \left(\frac{\partial \operatorname{vec} \boldsymbol{\Omega}}{\partial \boldsymbol{\beta}}\right)' \operatorname{vec} \boldsymbol{F}^{-1}(\boldsymbol{\Omega} - \boldsymbol{u}\,\boldsymbol{u}') &= \boldsymbol{0}, \\ \left(\frac{\partial \operatorname{vec} \boldsymbol{\Omega}}{\partial \boldsymbol{\gamma}}\right)' \operatorname{vec} \boldsymbol{F}^{-1}(\boldsymbol{\Omega} - \boldsymbol{u}\,\boldsymbol{u}') &= \boldsymbol{0}, \end{aligned} \tag{11.14}$$

where $\boldsymbol{\gamma} = (\alpha, \sigma^2)'$.

If the u_t are normally distributed, $\boldsymbol{F} = 2\boldsymbol{\Omega}$ and the quasi-likelihood estimator for $\boldsymbol{\theta}$ coincides with the maximum likelihood estimator. The equation (11.14) then corresponds to the true score equations (4.2), (4.3) of Anh (1988).

The martingale information (see Section 6.3) contained in (11.11) is

$$\boldsymbol{I}_{QS(U)} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X} & \boldsymbol{O}_{k\times 2} \\ \boldsymbol{O}_{2\times k} & \boldsymbol{O}_{2\times 2} \end{pmatrix},$$

which clearly contributes only to the estimation of $\boldsymbol{\beta}$ while that in (11.12) is

$$\boldsymbol{I}_{QS(V)} = \sum_{t=1}^{n} (\operatorname{var} u_t^2)^{-1} (\dot{g}_t(\boldsymbol{\theta}))'(\dot{g}_t(\boldsymbol{\theta})) = \boldsymbol{A}'\,\boldsymbol{F}^{-1}\,\boldsymbol{A},$$

where $\boldsymbol{A} = (\partial g_i(\boldsymbol{\theta})/\partial \theta_j)$, and the combined information in (11.13) is

$$\begin{aligned} I_{QS(U,V)} = {} & \left(\begin{array}{c c c} \boldsymbol{X}'\boldsymbol{\Omega}^{-1}(\boldsymbol{I} - \boldsymbol{R}\boldsymbol{S})^{-1}\boldsymbol{X} & \vdots & \boldsymbol{O}_{k\times 2} \\ & & \\ \boldsymbol{O}_{2\times k} & \vdots & \boldsymbol{O}_{2\times 2} \end{array}\right) + \boldsymbol{A}'\boldsymbol{F}^{-1}(\boldsymbol{I} - \boldsymbol{R}\boldsymbol{S})^{-1}\boldsymbol{A} \\ & + \begin{pmatrix} \boldsymbol{X}'\boldsymbol{\Omega}^{-1}(\boldsymbol{I} - \boldsymbol{R}\boldsymbol{S})^{-1}\boldsymbol{R}\boldsymbol{A} \\ \boldsymbol{O}_{2\times(k+2)} \end{pmatrix} \end{aligned} \tag{11.15}$$

$$+\left(\boldsymbol{A}'\boldsymbol{R}(\boldsymbol{I}-\boldsymbol{RS})^{-1}\boldsymbol{\Omega}^{-1}\boldsymbol{X} \vdots \boldsymbol{O}_{(k+2)\times 2}\right),$$

where $\boldsymbol{R} = \operatorname{diag}(R_1, \ldots, R_n)$, $\boldsymbol{S} = \operatorname{diag}(S_1, \ldots, S_n)$, which of course is $\boldsymbol{I}_{QS(U)} + \boldsymbol{I}_{QS(V)}$ when $Eu_t^3 = 0$ for each t.

In contrast to the above estimating functions, Anh (1988) used nonlinear least squares based on minimization of

$$\sum_{t=1}^{n}(e_t^2 - Ee_t^2)^2, \tag{11.16}$$

where the e_t are the (observable) elements of the vector

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\,\boldsymbol{u} = \boldsymbol{P}\boldsymbol{u} = \boldsymbol{P}\boldsymbol{y},$$

where $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Note that (11.16) amounts to using the estimating function

$$\sum_{t=1}^{n} \dot{f}_t(\boldsymbol{\theta})(e_t^2 - f_t(\boldsymbol{\theta})),$$

where $f_t(\boldsymbol{\theta}) = Ee_t^2$. Anh showed that this leads to a strongly consistent and asymptotically normal estimator $\hat{\boldsymbol{\theta}}_A$, say, satisfying

$$\hat{\boldsymbol{\theta}}_A - \boldsymbol{\theta} \overset{\mathrm{d}}{\simeq} MVN(\boldsymbol{0},\ (\boldsymbol{A}'\boldsymbol{A})^{-1}\,\boldsymbol{A}'\boldsymbol{F}\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1})$$

for large n under his Conditions 1–3. However, under the same conditions, minor modifications of the same method of proof can be used to establish that

$$\hat{\boldsymbol{\theta}}_{QS(V)} - \boldsymbol{\theta} \overset{\mathrm{d}}{\simeq} MVN(\boldsymbol{0},\ (\boldsymbol{A}'\boldsymbol{F}^{-1}\boldsymbol{A})^{-1})$$

and

$$\hat{\boldsymbol{\theta}}_{QS(U,V)} - \boldsymbol{\theta} \overset{\mathrm{d}}{\simeq} MVN(\boldsymbol{0},\ \boldsymbol{I}_{QS(U,V)}^{-1}),$$

where $\hat{\boldsymbol{\theta}}_{QS(V)}$ and $\hat{\boldsymbol{\theta}}_{QS(U,V)}$ are, respectively, the quasi-likelihood estimators based on (11.12) and (11.13). By construction we have that

$$\boldsymbol{I}_{QS(V)}^{-1} - \boldsymbol{I}_{QS(U,V)}^{-1}$$

in nonnegative definite, while

$$(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{F}\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1} - (\boldsymbol{A}'\,\boldsymbol{F}^{-1}\boldsymbol{A})^{-1}$$

is also nonnegative definite. This last result holds because, for a covariance matrix $\boldsymbol{F}$ and $n \times p$ matrix $\boldsymbol{A}$, the Gauss-Markov Theorem gives nonnegativity of

$$\boldsymbol{B}\boldsymbol{F}\boldsymbol{B}' - (\boldsymbol{A}'\boldsymbol{F}^{-1}\boldsymbol{A})^{-1}$$

for every $p \times n$ matrix $\boldsymbol{B}$ satisfying $\boldsymbol{B}\boldsymbol{A} = \boldsymbol{I}_p$ (e.g., Heyde (1989)), and the result holds in particular for $\boldsymbol{B} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$.

Thus, from the point of view of asymptotic variance, the order of preference is $\hat{\boldsymbol{\theta}}_{QS(U,V)}$, $\hat{\boldsymbol{\theta}}_{QS(V)}$, $\hat{\boldsymbol{\theta}}_A$.

11.3 Estimating the Infection Rate in an Epidemic

The discussion in this section is based on the results of Watson and Yip (1992). We shall begin by considering the simple stochastic epidemic model for a closed population of size N and where the infection rate is β. Let $S(t)$ denote the number of susceptibles and $I(t)$ the number of infectives as time t, so that $S(t)+I(t)=N$, and suppose that $I(0)=a$.

The simple epidemic process (e.g., Bailey (1975, p. 39)) is a Markov process for which, when $I(t)$ infectives are present at time t, the chance of an infective contact with a specified individual in the time interval $(t, t+\Delta t)$ is $\theta\, I(t)\, \Delta t + o(\Delta t)$, where θ is the infection rate, which is taken as constant. Then, since there are $N - N(t)$ susceptibles present at time t, the probability of an infection in time $(t, t+\Delta t)$ is $\theta\, I(t)(N - I(t))\, \Delta t + o(\Delta t)$. It is assumed that the probability of more than one infection in $(t, t+\Delta t)$ is in $o(\Delta t)$. Then, if $\mathcal{F}_t$ denotes the σ-field of history up to time t, we have in an obvious notation,

$$E\left(d\,I(t)\,\middle|\,\mathcal{F}_{t-}\right) = \theta\, I(t)(N - I(t))\, dt$$

and the basic zero mean martingale defined on the infectives process $\{I(t)\}$ is

$$\begin{aligned} M(t) &= I(t) - I(0) - \int_0^t E\left(d\,I(s)\,\middle|\,\mathcal{F}_{s-}\right) \\ &= I(t) - I(0) - \theta \int_0^t I(s)(N - I(s))\, ds. \end{aligned} \tag{11.17}$$

Then, the Hutton-Nelson family of estimating based on the martingale (11.17) is

$$\mathcal{M} = \left\{ \int_0^t \alpha(s)\, dI(s) - \theta \int_0^t \alpha(s)\, I(s)(N - I(s))\, ds, \qquad \alpha(s) \text{ predictable} \right\}$$

from which the quasi-score estimating function for estimation of θ is obtained by choosing $\alpha(s) \equiv 1$ for the predictable weight function. This follows (e.g., using results of Section 2.5) since it is easily seen that

$$E\left(d\dot{M}(t)\,\middle|\,\mathcal{F}_{t-}\right) = -I(t)(N - I(t))\, dt,$$

$$\begin{aligned} d\langle M\rangle_t = E\left((dM(t))^2\,\middle|\,\mathcal{F}_{t-}\right) &= E\left(dM(t)\,\middle|\,\mathcal{F}_{t-}\right) \\ &= \theta\, I(t)\,(N - I(t))\, dt. \end{aligned}$$

The quasi-likelihood estimator of θ based on complete observation of the process $\{I(t)\}$ over $[0, T]$ is then

$$\hat{\theta}_T = (I(T) - a) \Big/ \int_0^T I(s)(N - I(s))\, ds \tag{11.18}$$

and it should be noted that this coincides with the maximum likelihood estimator. It is not difficult to check that $\hat{\theta}_T$ is a strongly consistent estimator of θ and

$$(I(t) - a)^{-\frac{1}{2}} \, M(T) \xrightarrow{\text{d}} N(0,1)$$

as $T \to \infty$ so that

$$(I(t) - a)^{\frac{1}{2}} \, \hat{\theta}_T^{-1} \, (\hat{\theta}_T - \theta) \xrightarrow{\text{d}} N(0,1)$$

from which confidence intervals for θ could be constructed.

In practice, one would not expect complete observation of the process. The available data might, for example, be of the form $\{(t_k, i_k), \ k = 0, 1, \dots, m\}$, where $m \geq 1, \ 0 \leq t_0 < t_1 < \dots < T_m \leq T$ are fixed time points and i_k is the observed number of infectives at time t_k. Then, a reasonable approximation to the estimator (11.18) is provided by

$$(I(t_m) - a) \Big/ \left[\frac{1}{2} \left(I(t_0)(N - I(t_0)) + I(t_m)(N - I(t_m)) \right) + \sum_{r=1}^{m-1} I(t_r)(N - I(t_r)) \right]$$

which is obtained by using the simple trapezoidal rule to approximate the stochastic integral. This estimator is due to Choi and Severo (1988).

A rather more realistic model is provided by the general stochastic epidemic model that allows for removals from the population (e.g., Bailey (1975, pp. 88-89)). Here the probability of one infection in the time interval $(t, t + \Delta t)$ is $\theta \, I(t) \, S(t) \, \Delta t + o(\Delta t)$ and of one removal is, say, $\gamma \, I(t) \, \Delta(t) + o(\Delta t)$, while the probability of more than one change is $o(\Delta t)$. Suppose $S(0) = n$ and $I(0) = a$. Since the removals come only from the infectives, we have that

$$E\left(dS(t) \, \Big| \, \mathcal{F}_{t-} \right) = -\theta \, I(t) \, S(t) \, dt$$

and the basic zero-mean martingale defined on the process $\{S(t)\}$ is

$$\begin{aligned} N(t) &= n - S(t) + \int_0^t E\left(dS(s) \, \Big| \, \mathcal{F}_{s-} \right) \\ &= n - S(t) - \theta \int_0^t I(s) \, S(s) \, ds. \end{aligned}$$

Estimation of θ can then be achieved along similar lines to those described above for the case of the simple epidemic. The quasi-score estimating function for the Hutton-Nelson family based on the martingale $\{N(t)\}$ is

$$n - S(T) - \theta \int_0^T I(s) \, S(s) \, ds,$$

so that the quasi-likelihood estimator is

$$\hat{\theta}_T = (n - S(T)) \Big/ \int_0^T I(s) \, S(s) \, ds. \qquad (11.19)$$

Again $\hat{\theta}_T$ is strongly consistent for θ and asymptotically normally distributed, this time with

$$(n - S(T))^{\frac{1}{2}}\, \hat{\theta}_T^{-1}\, (\hat{\theta}_T - \theta) \xrightarrow{\mathrm{d}} N(0,1)$$

as $T \to \infty$.

Similar ideas can be applied for the estimation of the removal rate γ. If $R(t)$ denotes the number of removals up to time t, we have that

$$E\left(dR(t) \,\middle|\, \mathcal{F}_{t-}\right) = \gamma\, I(t)\, dt$$

and the basic zero-mean martingale defined on the process $\{R(t)\}$ is

$$\begin{aligned} K(t) &= R(t) - \int_0^t E\left(dR(s) \,\middle|\, \mathcal{F}_{s-}\right) \\ &= R(t) - \gamma \int_0^s I(s)\, ds. \end{aligned}$$

We have

$$E\left(d\dot{K}(t) \,\middle|\, \mathcal{F}_{t-}\right) = I(t)\, dt,$$

the dot now referring to differentiation with respect to γ, and

$$d\langle K\rangle_t = E\left((dR(t))^2 \,\middle|\, \mathcal{F}_{t-}\right) = \gamma\, I(t)\, dt,$$

so that, as in the previous cases, the basic zero-mean martingale is itself the quasi-score for the Hutton-Nelson family. The quasi-likelihood estimator of γ is then

$$\hat{\gamma}_T = R(T) \Big/ \int_0^T I(s)\, ds. \tag{11.20}$$

Again we would not expect complete observation of the epidemic process. The available information might be of the form $\{(t_k, s_k, i_k),\ k = 0, 1, \ldots, m\}$, which provides progressive information on both the numbers of susceptibles and infectives. This would allow for approximation to the quasi-likelihood estimator $\hat{\theta}_T$ given in (11.19), but the removal rate γ would not be able to be estimated via (11.20) without information on the number of removals.

11.4 Estimating Population Size from Multiple Recapture Experiments

Consider a population in which there are θ animals, where θ is unknown. A multiple recapture experiment is conducted in which animals are marked at the time of their first capture. At each subsequent capture the mark is correctly recorded. Apart from time allowed for marking or noting of marks, each captured animal is released immediately. The aim is to make inferences about θ from the set of observed captures of marked and unmarked animals.

The results in this section follow Becker and Heyde (1990) and are derived using a continuous-time formulation. Analogous results also hold for multiple recapture experiments in discrete time.

We label the animals by $1, 2, \ldots, \theta$ and let $N_i(t)$ denote the number of times animal i has been caught in $[0, t]$. Write $\boldsymbol{N}_t$ for $(N_1(t), \ldots, N_\theta(t))'$. Each $\{N_i(t);\ t \geq 0\}$ is a continuous-time counting process with right-continuous sample paths and common intensity function λ defined by

$$\lambda_t = \lim_{h \downarrow 0} \frac{1}{h} \Pr\left[N_i(t+h) = N_i(t) + 1 \,\middle|\, \mathcal{F}_t \right],$$

where $\mathcal{F}_t$ is the σ-field generated by $\{\boldsymbol{N}_x :\ 0 \leq x \leq t\}$. The dependence of λ on t can reflect the dependence of the animal's behavior on time as well as variations in the intensity of trapping.

Consider observations from a multiple capture experiment over $[0, \tau]$. The total number of captures by time t is given by $N_t = \sum_{i=1}^{\theta} N_i(t)$. Write M_t for the number of animals that are marked by time t. Then M_t is the number of captures of unmarked animals, while $K_t = N_t - M_t$ gives the number of captures of marked animals by time t.

Maximum likelihood estimation of θ based on observation of $\{M_t,\ N_t,\ 0 \leq t \leq \tau\}$ is complicated by the fact that the intensity function λ_x is generally unknown. We find that the log-likelihood function is

$$\begin{aligned} L &= \int_0^\tau \log(\lambda_x)\, dN_x - \theta \int_0^\tau \lambda_x\, dx + \int_0^\tau \log(\theta - M_{x-})\, dM_x \\ &= \int_0^\tau \log(\lambda_x^*)\, dN_x - \int_0^\tau \lambda_x^*\, dx - N_\tau \log\theta + \int_0^\tau \log(\theta - M_{x-})\, dM_x, \end{aligned}$$

where $\lambda_x^* = \theta\, \lambda_x$ is an arbitrary positive value if λ_x is arbitrary positive. The likelihood function is maximized with respect to θ by the solution of

$$-\frac{\partial L}{\partial \theta} = \theta^{-1} N_\tau - \int_0^\tau \frac{dM_x}{\theta - M_{x-}} = 0, \tag{11.21}$$

irrespective of the form of λ^*, which provides the maximum likelihood estimator for θ. See Samuel (1969), and references quoted by her, for more detailed consideration of related maximum likelihood estimation.

Estimation of θ via equation (11.21) requires the use of numerical computation because of the term $\theta - M_{x-}$ in the denominator. However, there are alternative estimators for θ that have explicit expressions, are easy to compute, and have high asymptotic efficiency relative to the benchmark provided by (11.21).

To facilitate avoidance of the computationally troublesome $(\theta - M_{x-})$ in the denominator of (11.21) it is sensible to work with the martingale H_t defined by

$$dH_t = M_{t-}\, dM_t - (\theta - M_{t-})\, dK_t, \tag{11.22}$$

which is suggested by the right-hand alternative representation for the martingale

$$N_{\mathcal{T}} - \int_0^{\mathcal{T}} \frac{\theta\, dM_x}{\theta - M_{x-}} = K_{\mathcal{T}} - \int_0^{\mathcal{T}} \frac{M_{x-}\, dM_x}{\theta - M_{x-}}, \tag{11.23}$$

and to seek effective estimating functions from within the class of martingale estimating functions

$$\left\{ \int_0^{\mathcal{T}} f_x\, dH_x, \qquad f_x \text{ predictable} \right\}. \tag{11.24}$$

The quasi-score estimating function from this class is

$$\int_0^{\mathcal{T}} (d\bar{H}_x)(d\langle H\rangle_x)^{-1}\, dH_x, \tag{11.25}$$

where

$$d\bar{H}_x = E\left(d(dH_x/d\theta) \,\middle|\, \mathcal{F}_{x-} \right) = -E\left(dK_x \,\middle|\, \mathcal{F}_{x-} \right) = -M_{x-}\, d\Lambda_x$$

and $\langle H\rangle_t$ is the quadratic characteristic of the martingale H_t, so that

$$\begin{aligned} d\langle H\rangle_x &= E\left((dH_x)^2 \,\middle|\, \mathcal{F}_{x-} \right) = E\left(M_{x-}^2\, dM_x + (\theta - M_{x-})^2\, dK_x \,\middle|\, \mathcal{F}_{x-} \right) \\ &= \theta\, M_{x-}(\theta - M_{x-})\, d\Lambda_x, \end{aligned}$$

and (11.25) corresponds to (11.23). Thus the quasi-likelihood estimator is the maximum likelihood estimator (MLE), $\hat{\theta}$ say.

Now the martingale information in (11.23) (see Chapter 6), whose reciprocal is essentially an asymptotic variance, is

$$I_{ML} = \theta^{-1} \int_0^{\mathcal{T}} \frac{M_x\, d\Lambda_x}{\theta - M_x}, \tag{11.26}$$

and the central limit result Theorem 12.6 or, for example, those of Aalen (1977) and Rebolledo (1980) can be used to show that

$$I_{ML}^{1/2}(\hat{\theta} - \theta)$$

tends in distribution to the standard normal law as $\theta \to \infty$.

Maximum likelihood provides a benchmark but other estimating functions from the family (11.24) of the type

$$\int_0^{\mathcal{T}} g_x\, dH_x, \tag{11.27}$$

where g_x is of the form $g(N_{x-})$, produce simple and easily computable estimators

$$\hat{\theta}_g = \int_0^{\mathcal{T}} g_x\, M_{x-}\, dN_x \bigg/ \int_0^{\mathcal{T}} g_x\, dK_x, \tag{11.28}$$

whose performance relative to the benchmark is well worth considering. We shall examine a number of choices for g_x and, in particular, determine whether $g_x \equiv 1$ is a good choice.

The martingale information in the estimating function (11.27) is

$$I_g = \left(\int_0^\tau g_x\, M_x\, d\Lambda_x\right)^2 \Big/ \left(\theta \int_0^\tau g_x^2\, M_x(\theta - M_x)\, d\Lambda_x\right) \tag{11.29}$$

and, subject to suitable restrictions on g, the above cited central limit results can be used to show that

$$I_g^{1/2}(\hat{\theta}_g - \theta)$$

tends in distribution to a standard normal law as $\theta \to \infty$. To investigate the relative behavior of the random quantities (11.26) and (11.29) as θ increases, information on the distributions of M_x and N_x, $x > 0$, is needed and one can show that the probability generating function of $(M_x,\, N_x)$ is

$$E(y^{M_x}\, z^{N_x}) = e^{-\theta\Lambda_x}(1 - y + y\, e^{x\Lambda_x})^\theta.$$

Thus, M_x is distributed as Binomial $(\theta, 1 - e^{-\Lambda_x})$ and N_x as Poisson $(\theta\,\Lambda_x)$. In particular, $\theta^{-1}\, M_x \xrightarrow{\text{a.s.}} 1 - e^{-\Lambda_x}$ and $\theta^{-1}\, N_x \xrightarrow{\text{a.s.}} \Lambda_x$ as $\theta \to \infty$.

Then, if the function g is well behaved, and, in particular, there exists a sequence of constants $\{c_\theta\}$ such that

$$c_\theta\, g_x \xrightarrow{\text{P}} g_d(x), \qquad x \in [0, \tau], \tag{11.30}$$

as $\theta \to \infty$, where $g_d(x)$ is a deterministic function, then it follows that

$$\theta\, I_{ML} \xrightarrow{\text{a.s.}} \int_0^\tau (e^{\Lambda_x} - 1)\, d\Lambda_x = e^{\Lambda_\tau} - 1 - \Lambda_\tau$$

and

$$\theta\, I_g \xrightarrow{\text{P}} \left(\int_0^\tau g_d(x)(1 - e^{-\Lambda_x})\, d\Lambda_x\right)^2 \Big/ \int_0^\tau g_d^2(x)\, e^{-\Lambda_x}\,(1 - e^{-\Lambda_x})\, d\Lambda_x.$$

The asymptotic relative efficiency of $\hat{\theta}_g$ to the benchmark of the MLE is the ratio

$$\text{ARE}(\hat{\theta}_g) = \frac{\left(\int_0^\tau g_d(x)(1 - e^{-\Lambda_x})\, d\Lambda_x\right)^2}{(e^{-\Lambda_\tau} - 1 - \Lambda_\tau)\int_0^\tau g_d^2(x)\, e^{-\Lambda_x}\,(1 - e^{-\Lambda_x})\, d\Lambda_x}. \tag{11.31}$$

That $\text{ARE}(\hat{\theta}_g)$ is less than or equal to unity follows from the Cauchy-Schwarz inequality and equality holds iff $g_d(x) = \exp \Lambda_x$.

There is no obvious choice for g which satisfies the requirement (11.30) for $g_d(x) = \exp \Lambda_x$ and does not involve θ. Using (11.28), we consider the convenient estimators

$$\begin{aligned}
\hat{\theta}_1 &= \int_0^\tau M_{x-}\, dN_x \Big/ \int_0^\tau dK_x, \\
\hat{\theta}_2 &= \int_0^\tau M_{x-}^2\, dN_x \Big/ \int_0^\tau M_{x-}\, dK_x,
\end{aligned}$$

and

$$\hat{\theta}_3 = \int_0^{\mathcal{T}} M_{x-}\, N_{x-}\, dN_x \bigg/ \int_0^{\mathcal{T}} N_{x-}\, dK_x$$

for which the form $g_d(x)$ in (11.30) is 1, $1 - e^{-\Lambda_x}$ and Λ_x, respectively. By direct integration,

$$\text{ARE}(\hat{\theta}_1) = \frac{2(\Lambda_{\mathcal{T}} + e^{-\Lambda_{\mathcal{T}}} - 1)^2}{(e^{\Lambda_{\mathcal{T}}} - 1 - \Lambda_{\mathcal{T}})(1 - e^{-\Lambda_{\mathcal{T}}})^2},$$

$$\text{ARE}(\hat{\theta}_2) = \frac{(2\Lambda_{\mathcal{T}} + 1 - (2 - e^{-\Lambda_{\mathcal{T}}})^2)^2}{(e^{\Lambda_{\mathcal{T}}} - 1 - \Lambda_{\mathcal{T}})(1 - e^{-\Lambda_{\mathcal{T}}})^4},$$

and

$$\text{ARE}(\hat{\theta}_3) = \frac{(\Lambda_{\mathcal{T}}^2 + 2\Lambda_{\mathcal{T}}\, e^{-\Lambda_{\mathcal{T}}} - 2 + 2e^{-\Lambda_{\mathcal{T}}})^2}{(e^{\Lambda_{\mathcal{T}}} - 1 - \Lambda_{\mathcal{T}})\{3 + 2e^{-\Lambda_{\mathcal{T}}}(\Lambda_{\mathcal{T}} + 1)(\Lambda_{\mathcal{T}}\, e^{-\Lambda_{\mathcal{T}}} - 2\Lambda_{\mathcal{T}} - 2) + (2 - e^{-\Lambda_{\mathcal{T}}})^2\}}.$$

It is convenient to compare these efficiencies against the scale $1 - e^{-\Lambda_{\mathcal{T}}}$ which represents the expected fraction of animals marked by time $\mathcal{T}$ and therefore has direct experimental relevance. Graphs are given in Becker and Heyde (1990).

All three estimators are very efficient over the range $0 < 1 - e^{-\Lambda_{\mathcal{T}}} < 0.5$, which should contain the range of most practical importance. The efficiency of $\hat{\theta}_1$ is highest over this range, which suggests that it is better to have a function g in (11.28) that is not zero initially. The efficiency of $\hat{\theta}_1$ drops significantly once the expected fraction of marked animals exceeds 0.5, as might be expected, since $g_x = 1$ is not an increasing function. These observations encourage us to consider the estimator

$$\hat{\theta}_4 = \int_0^{\mathcal{T}} \exp\left(\frac{N_{x-}}{M_{x-}}\, 1_{\{M_{x-} > 0\}}\right) M_{x-}\, dN_x \bigg/ \int_0^{\mathcal{T}} \exp\left(\frac{N_{x-}}{M_{x-}}\, 1_{\{M_{x-} > 0\}}\right) dK_x.$$

Corresponding to this estimator we have

$$g_d(x) = \exp\left(\frac{\Lambda_x}{1 - e^{-\Lambda_x}}\right),$$

which is an increasing function starting at the value e and is close to the optimal e^{Λ_x} for large Λ_x. The efficiency of this estimator, $\text{ARE}(\hat{\theta}_4)$ is ≥ 0.995 over the range $0 < 1 - e^{-\Lambda_{\mathcal{T}}} < 0.5$ and never falls below 0.9774.

It should finally be remarked that the argument leading to estimators given by (11.28) remains valid even when λ is random. For example, random fluctuations in environmental factors might affect the trapping rate.

If a single trap is used that captures just one animal at a time, then the estimators given by (11.28) remain valid even when immediate release of the animals is not possible. Occupancy of the trap simply has the effect of suspending the entire capture process temporarily.

11.5 Robust Estimation

The results in this section are from Kulkarni and Heyde (1987). Let $Y_1, \dots, Y_n$ be a sample of n observations from a discrete time stochastic process whose distribution depends on a real valued parameter belonging to an open interval of the real line. It is known that when there are outliers in the observations, the standard methods of estimation (maximum likelihood, method of moments etc.) may be seriously affected with adverse results. Robustified versions of some of these procedures are discussed in quite a number of sources, for example, Huber (1981), Gastwirth and Rubin (1975), Denby and Martin (1979), Martin (1980, 1982), Martin and Yohai (1985), Künsch (1984), Basawa, Huggins and Staudte (1985), Bustos (1982) and references therein. These authors, however, concentrate on particular problems and we shall here discuss a general procedure of broad applicability based on robustifying the quasi-likelihood framework. As usual this produces an estimating function, which has certain optimality properties, within a specified class of estimating functions.

First note, that it is generally possible to specify a set of functions $h_i = h_i(Y_1, \dots, Y_i, \theta)$, $i = 1, 2, \dots, n$, that are martingale differences, namely,

$$E\left(h_i \,\middle|\, \mathcal{F}_{i-1}\right) \quad \text{a.s.}, \quad 1 \le i \le n, \tag{11.32}$$

where $\mathcal{F}_k$ denotes the past history σ-field generated by Y_j, $1 \le j \le k$, $k \ge 1$, and $\mathcal{F}_0$ is the trivial σ-field. A natural choice for h_i in the case of integrable Y_i is just $Y_i - E(Y_i \,|\, \mathcal{F}_{i-1})$.

Now in this book we have given special consideration to the class $\mathcal{M}$ of square integrable martingale estimating functions

$$G_n = \sum_{i=1}^{n} a_{i-1}\, h_i, \tag{11.33}$$

where the h_i are specified and the coefficients a_{i-1} are functions of $Y_i, \dots, Y_{i-1}$ and θ, i.e., are $\mathcal{F}_{i-1}$-measurable. We have shown in Chapter 2 how to choose from $\mathcal{M}$ a quasi-score estimating function $G_n^* = \sum_{i=1}^{n} a_{i-1}^*\, h_i$ with certain optimality properties. However, although the h_i can be chosen to be robust functions of the observations, outliers can enter the estimating function through the weights a_{i-1} and hence adversely affect the properties of the estimates. Thus the class $\mathcal{M}$ of estimating functions is not resistant to outliers.

To overcome this difficulty we constrain the a_{i-1}'s in (11.33) to be of robust form by considering a subset $\mathcal{S}$ of $\mathcal{M}$ whose elements are of the form

$$S_n = \sum_{i=1}^{n} h_i \sum_{j=1}^{i} b_{j,i}\, k_{i-j}, \tag{11.34}$$

where the $b_{j,i}$, $1 \le j \le i$, are constants and the k_{i-j}, $1 \le j \le i$, $i \ge 1$, are specified functions of $Y_1, \dots, Y_{i-j}$ and θ such that $k_0 = 1$ and

$$E(k_{i-j} \,|\, \mathcal{F}_{i-j-1}) = 0, \qquad i > j. \tag{11.35}$$

That is, the k_{i-j}, $i > j$, are martingale differences. Note that, unlike the a_{i-1}'s in (11.33), the $b_{j,i}$'s in (11.34) are constants and free of the observations. Also, the k_{i-j}'s in (11.34) being specified functions, can be chosen to be robust, just as with the h_i's. We may take, for example $k_{i-j} = h_{i-j}$. Thus, the estimating functions in $\mathcal{S}$ can be chosen to be robust and we shall show how to select the optimum S^* within $\mathcal{S}$.

In these introductory remarks we have described the case of a scalar parameter for clarity. The results given below, however, deal with a vector parameter and vector valued estimating functions. In Section 11.5.1 we shall develop a general theory of robust quasi-score estimating functions and discuss the loss of efficiency due to robustification. The analysis is carried out formally without specific attention to whether the resultant quasi-score estimating functions are strictly allowable in the sense of involving only the data and the parameter to be estimated. Indeed, they often involve unobservable quantities but are nevertheless amenable to iterative solution. In Section 11.5.2 we shall illustrate by considering the example of a regression model with autoregressive errors.

11.5.1 Optimal Robust Estimating Functions

Suppose that $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_n$ is a vector process and that $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)'$ is a parameter taking values in an open subset of p dimensional Euclidean space. As usual we consider the class $\mathcal{G}$ of zero mean, square integrable p dimensional vector estimating functions $\boldsymbol{G}_n = \boldsymbol{G}_n(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n, \boldsymbol{\theta})$ that are a.s. differentiable with respect to the components of $\boldsymbol{\theta}$ and such that

$$E\dot{\boldsymbol{G}}_n = (E\,\partial \boldsymbol{G}_{n,i}/\partial\theta_j)$$

and $E\boldsymbol{G}_n\,\boldsymbol{G}_n'$ are nonsingular.

Now let $\boldsymbol{h}_i$ and $\boldsymbol{k}_{i-j}$ be specified vectors of dimension q satisfying vector forms of (11.32) and (11.35) for $1 \le j \le i-1$, $i \ge 1$ and $\boldsymbol{k}_0 = (1, 1, \dots, 1)'$. Usually $q \le p$. We consider the subset $\mathcal{S}$ of $\mathcal{G}$ having elements

$$\boldsymbol{S}_n = \sum_{i=1}^{n} \left(\sum_{j=1}^{i} \boldsymbol{b}_{j,i}\,\boldsymbol{k}_{i-j}' \right) \boldsymbol{h}_i \tag{11.36}$$

where the $\boldsymbol{b}_{j,i}$ are constant vectors of dimension p. Then, with the aid of Theorem 2.1 we shall obtain the following theorem.

Theorem 11.3 The quasi-score estimating function $\boldsymbol{S}_n^*$ within the class $\mathcal{S}$ defined by (11.36) is given by

$$\boldsymbol{S}_n^* = \sum_{i=1}^{n} \left(\sum_{j=1}^{i} \boldsymbol{b}_{j,i}^*\,\boldsymbol{k}_{i-j}' \right) \boldsymbol{h}_i,$$

where the $\boldsymbol{b}^*_{j,i}$, $1 \le i \le n$, satisfy

$$\begin{bmatrix} E(\boldsymbol{k}'_{i-1}\,\boldsymbol{h}_i\,\boldsymbol{h}'_i\,\boldsymbol{k}_{i-1}) & \cdots & E(\boldsymbol{k}'_{i-1}\,\boldsymbol{h}_i\,\boldsymbol{h}'_i\,\boldsymbol{k}_0) \\ E(\boldsymbol{k}'_{i-2}\,\boldsymbol{h}_i\,\boldsymbol{h}'_i\,\boldsymbol{k}_{i-1}) & \cdots & E(\boldsymbol{k}'_{i-2}\,\boldsymbol{h}_i\,\boldsymbol{h}'_i\,\boldsymbol{k}_0) \\ \vdots & & \vdots \\ E(\boldsymbol{k}'_0\,\boldsymbol{h}_i\,\boldsymbol{h}'_i\,\boldsymbol{k}_{i-1}) & \cdots & E(\boldsymbol{k}'_0\,\boldsymbol{h}_i\,\boldsymbol{h}'_i\,\boldsymbol{k}_0) \end{bmatrix} \begin{bmatrix} \boldsymbol{b}^{*\prime}_{1,i} \\ \boldsymbol{b}^{*\prime}_{2,i} \\ \vdots \\ \boldsymbol{b}^{*\prime}_{i,i} \end{bmatrix} = \begin{bmatrix} E(\boldsymbol{k}'_{i-1}\,\dot{\boldsymbol{h}}_i) \\ E(\boldsymbol{k}'_{i-2}\,\dot{\boldsymbol{h}}_i) \\ \vdots \\ E(\boldsymbol{k}'_0\,\dot{\boldsymbol{h}}_i) \end{bmatrix}. \tag{11.37}$$

For the special case when $E(\boldsymbol{k}'_i\,\boldsymbol{h}_m\,\boldsymbol{h}'_m\,\boldsymbol{k}_j) = 0$, $i \ne j$ and all m (which holds, for example, if $E(\boldsymbol{h}_m\,\boldsymbol{h}'_m \,|\, \mathcal{F}_{m-1}) = \boldsymbol{c}_m$, a constant, for each m),

$$\boldsymbol{b}^{*\prime}_{j,i} = (E(\boldsymbol{k}'_{i-j}\,\boldsymbol{h}_i\,\boldsymbol{h}'_i\,\boldsymbol{k}_{i-j}))^{-1}\,E(\boldsymbol{k}'_{i-j}\,\dot{\boldsymbol{h}}_i), \quad 1 \le j \le i,\ 1 \le i \le n. \tag{11.38}$$

Proof We have, using the fact that the $\boldsymbol{h}_i$'s are martingale differences,

$$E\dot{\boldsymbol{S}}_n = E\sum_{i=1}^n \left(\sum_{j=1}^i \boldsymbol{b}_{j,i}\,\boldsymbol{k}'_{i-j}\right)\dot{\boldsymbol{h}}_i,$$

while

$$E\boldsymbol{S}_n\,\boldsymbol{S}^{*\prime}_n = E\sum_{i=1}^n \left\{\sum_{j=1}^i \boldsymbol{b}_{j,i}\,\boldsymbol{k}'_{i-j}\right\}\boldsymbol{h}_i\,\boldsymbol{h}'_i\left\{\sum_{l=1}^i \boldsymbol{k}_{i-l}\,\boldsymbol{b}^{*\prime}_{l,i}\right\}.$$

Then, the result (11.37) follows from Theorem 2.1 since

$$E\dot{\boldsymbol{S}}_n = E\boldsymbol{S}_n\,\boldsymbol{S}^{*\prime}_n$$

for all $\boldsymbol{S}_n \in \mathcal{S}$ when

$$E\left[\boldsymbol{k}'_{i-j}\,\boldsymbol{h}_i\,\boldsymbol{h}'_i\left(\sum_{l=1}^i \boldsymbol{k}_{i-l}\,\boldsymbol{b}^{*\prime}_{l,i}\right)\right] = E(\boldsymbol{k}'_{i-j}\,\dot{\boldsymbol{h}}_i),$$

$1 \le j \le i$, and this gives (11.37). The important special case (11.38) follows immediately from (11.37).

Now it is important to be able to assess the effect on efficiency of choosing a quasi-score estimating function from $\mathcal{S}$ rather than the broader class $\mathcal{M}$ of martingales of the form

$$\boldsymbol{G}_n = \sum_{i=1}^n \boldsymbol{a}_{i-1}\,\boldsymbol{h}_i$$

with the $\boldsymbol{a}_{i-1}$ being matrices of dimension $p \times q$ depending on $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{i-1}$ and $\boldsymbol{\theta}$, $1 \le i \le n$.

Efficiency may conveniently be assessed by comparing the martingale information in the quasi-score estimating functions within $\mathcal{G}$ and $\mathcal{S}$. In particular, it determines the size of the confidence zone in asymptotic confidence statements about the unknown parameter $\boldsymbol{\theta}$.

Let $\boldsymbol{G}_n^*$ be a quasi-score estimating function within $\mathcal{G}$. Then,

$$\boldsymbol{G}_n^* = \sum_{i=1}^n \boldsymbol{a}_{i-1}^* \, \boldsymbol{h}_i \tag{11.39}$$

with

$$\boldsymbol{a}_{i-1}^* = \left(E\left(\dot{\boldsymbol{h}}_i \,\middle|\, \mathcal{F}_{i-1}\right)\right)' \left(E\left(\boldsymbol{h}_i \, \boldsymbol{h}_i' \,\middle|\, \mathcal{F}_{i-1}\right)\right)^{-1},$$

the inverse being assumed to exist a.s. and the martingale information in $\boldsymbol{G}_n^*$ is

$$\boldsymbol{I}_{G_n^*} = \sum_{i=1}^n \left(E\left(\dot{\boldsymbol{h}}_i \,\middle|\, \mathcal{F}_{i-1}\right)\right)' \left(E\left(\boldsymbol{h}_i \, \boldsymbol{h}_i' \,\middle|\, \mathcal{F}_{i-1}\right)\right)^{-1} \left(E\left(\dot{\boldsymbol{h}}_i \,\middle|\, \mathcal{F}_{i-1}\right)\right) \tag{11.40}$$

(see Section 6.3). On the other hand, the corresponding result for the quasi-score estimating function from within $\mathcal{S}$ is

$$\begin{aligned} \boldsymbol{I}_{S_n^*} &= \left[\sum_{i=1}^n \boldsymbol{B}_i E\left(\dot{\boldsymbol{h}}_i \,\middle|\, \mathcal{F}_{i-1}\right)\right]' \left[\sum_{i=1}^n \boldsymbol{B}_i E\left(\boldsymbol{h}_i \, \boldsymbol{h}_i' \,\middle|\, \mathcal{F}_{i-1}\right) \boldsymbol{B}_i'\right]^{-1} \\ &\quad \cdot \left[\sum_{i=1}^n \boldsymbol{B}_i E\left(\dot{\boldsymbol{h}}_i \,\middle|\, \mathcal{F}_{i-1}\right)\right], \end{aligned} \tag{11.41}$$

where

$$\boldsymbol{B}_i = \sum_{j=1}^i \boldsymbol{b}_{j,i}^* \, \boldsymbol{k}_{i-j}'$$

and the $\boldsymbol{b}_{j,i}^*$ are given by (11.37).

It should be noted that if $\hat{\boldsymbol{\theta}}_G$ and $\hat{\boldsymbol{\theta}}_S$ are the estimators obtained from the estimating equations $\boldsymbol{G}_n^* = \boldsymbol{0}$ and $\boldsymbol{S}_n^* = \boldsymbol{0}$, respectively, then, under regularity conditions that are usually satisfied in cases which are of practical relevance,

$$(\hat{\boldsymbol{\theta}}_G - \boldsymbol{\theta})' \, \boldsymbol{I}_{G^*} \, (\hat{\boldsymbol{\theta}}_G - \boldsymbol{\theta}) \xrightarrow{\mathrm{d}} \chi_p^2, \tag{11.42}$$

$$(\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta})' \, \boldsymbol{I}_{S^*} \, (\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}) \xrightarrow{\mathrm{d}} \chi_p^2, \tag{11.43}$$

as $n \to \infty$ (Godambe and Heyde (1987, Section 4)). Here χ_p^2 is the chi-squared distribution with p degrees of freedom.

Of course the matrices $\boldsymbol{I}_{G^*}$ and $\boldsymbol{I}_{S^*}$ are of dimension $p \times p$ and their comparison is not straightforward in general. However, comparison can often be confined to scalar criteria; see Section 2.3. Furthermore, in special cases, such as that treated in the next section, useful results can be obtained from examination of the diagonal elements of interest.

It should be remarked that statistics based on optimal robust estimating functions usually involve quantities which cannot be calculated explicitely in terms of the data provided. However, in specific cases iterative computations can be carried out beginning from preliminary estimates of the unknown parameters. For information on similar procedures see Martin and Yohai (1985).

11.5.2 Example

A *regression model with autoregressive errors.* Suppose that $(Y_1, \dots, Y_n)$ comes from a model of the form

$$Y_i = \boldsymbol{\beta}' \boldsymbol{C}_i + X_i, \qquad X_i = \alpha\, X_{i-1} + \epsilon_i,$$

$i = 1, 2, \dots, n$, where $\boldsymbol{\beta}' = (\beta_1, \dots, \beta_r)$, $\boldsymbol{C}_i' = (C_{1i}, \dots, C_{ri})$, $|\alpha| < 1$, $X_0 = 0$ and the ϵ_i are independent and identically distributed random variables (i.i.d. r.v.'s) having zero mean and unit variance. Here the X_i are not directly observed, the $\boldsymbol{C}_i$ are fixed regressors and $\boldsymbol{\theta}' = (\alpha, \boldsymbol{\beta}')$ are unknown parameters.

Let $h_i = \psi(X_i - \alpha\, X_i) = \psi(\epsilon_i)$ and $k_{i-j} = h_{i-j}$ where $\psi(\cdot)$ is a bounded function chosen so that $E\psi(\epsilon_i) = 0$. If the ϵ_i have a symmetric distribution, then a convenient choice for ψ is Huber's function

$$\psi(u) = \begin{cases} u & \text{if } |u| < m, \\ m \operatorname{sgn} u & \text{if } |u| \ge m, \end{cases}$$

m being any specified constant.

Here $E(h_i^2 \,|\, \mathcal{F}_{i-1}) = E\psi^2(\epsilon)$ is constant and from Theorem 11.3 we readily find (see (11.38)) that

$$b_{j,i}^* = \begin{cases} K_1(\alpha^{j-1}, \mathbf{0}')', & j < i, \\ K_2(0, \boldsymbol{C}_i' - \alpha\, \boldsymbol{C}_{i-1}')', & j = i, \end{cases}$$

where $\mathbf{0}' = (0, \dots, 0)'$ (of dimension r) while K_1 and K_2 are (scalar) constants and the quasi-score robust estimating function for $\boldsymbol{\theta}$ within $\mathcal{S}$ based on $\{h_i\}$ is given by

$$\boldsymbol{S}_n^{*\prime} = (S_{n1}^*, \boldsymbol{S}_{n2}^{*\prime})'$$

with

$$\begin{aligned} S_{n1}^* &= K_1 \sum_{i=1}^{N} \psi(\epsilon_i) \sum_{j=1}^{i-1} \alpha^{j-1} \psi(\epsilon_{i-j}) = K_1 \sum_{i=1}^{n} \psi(\epsilon_i) \tilde{X}_{i-1}, \\ \boldsymbol{S}_{n2}^* &= K_2 \sum_{i=1}^{n} \psi(\epsilon_i)(\boldsymbol{C}_i - \alpha\, \boldsymbol{C}_{i-1}), \end{aligned}$$

where $\tilde{X}_t$ is given by

$$\tilde{X}_t = \alpha\, \tilde{X}_{t-1} + \psi(\epsilon_t) = \sum_{j=1}^{t} \alpha^{j-1} \psi(\epsilon_{t-j+1}).$$

This shows that the estimating function approach discussed by Basawa et al. (1985, Example 1), under the assumption of normally distributed ϵ_i, gives a quasi-score estimating function within $\mathcal{S}$ whatever the distribution of the ϵ_i.

Now using the same set of h_i, the estimating function which is optimal within $\mathcal{G}$ is, from (11.39),

$$\boldsymbol{G}_n^{*\prime} = (G_{n1}^*, \boldsymbol{G}_{n2}^{*\prime})'$$

with

$$G^*_{n1} \quad = \quad K_3 \sum_{i=1}^n \psi(\epsilon_i)\, X_{i-1},$$

$$\boldsymbol{G}^*_{n2} \quad = \quad K_3 \sum_{i=1}^n \psi(\epsilon_i)\, (\boldsymbol{C}_i - \alpha\, \boldsymbol{C}_{i-1}),$$

K_3 being a scalar constant. Clearly the $\psi(\epsilon_i)$ will not be directly observable in practice and iterative calculations will be necessary to deal with all these estimating functions. For example, the basic strategy for α is as follows. Start from $\tilde{X}_0(\hat{\alpha}_0, \boldsymbol{\beta}'_0)$ where $(\hat{\alpha}_0, \hat{\boldsymbol{\beta}}'_0)$ are robust preliminary estimates. Then calculate

$$\tilde{\epsilon}_{0,i} = \psi(Y_i - \hat{\boldsymbol{\beta}}'_0\, \boldsymbol{C}_i - \hat{\alpha}_0 Y_{i-1} + \hat{\boldsymbol{\beta}}'_0 \boldsymbol{C}_{i-1}), \quad i = 1, 2, \dots$$

and obtain $\{\tilde{X}_{0,i},\ i = 1, 2, \dots\}$ recursively from

$$\tilde{X}_{0,i} = \hat{\alpha}_0 \tilde{X}_{0,i-1} + \tilde{\epsilon}_{0,i}$$

taking $\tilde{X}_{0,0} = 0$. Next compute

$$\hat{\alpha}_1 = \sum_{i=1}^n \tilde{X}_{0,i} \tilde{X}_{0,i-1} \Big/ \sum_{i=1}^n \tilde{X}^2_{0,i-1}$$

and repeat the procedure. Continue the iterations until a stable value for $\hat{\alpha}$ is obtained.

We note that $\boldsymbol{S}^*_{n2}$ and $\boldsymbol{G}^*_{n2}$ are the same, except for the constant multiplier, both being robust, and it is of special interest to compare the performance of the robust estimating function S^*_{n1} with that of the nonrobust G^*_{n1}.

For the purpose of this comparison the regression part of the model is irrelevant. Consequently, we now focus attention on the autoregression and suppose that the data is $(X_1, \dots, X_n)$.

From (11.40) and (11.41) we obtain

$$I_{G^*_1} \quad = \quad d \sum_{i=1}^n X^2_{i-1}, \tag{11.44}$$

$$I_{S^*_1} \quad = \quad d \left(\sum_{i=1}^n \tilde{X}_{i-1}\, X_{i-1} \right)^2 \left(\sum_{i=1}^n \tilde{X}^2_{i-1} \right)^{-1}, \tag{11.45}$$

where $d = (E\dot{\psi}(\epsilon))^2 / E\psi^2(\epsilon)$.

Now suppose that $|\alpha| < 1$, i.e., the autoregression is (asymptotically) stationary. Then, from the martingale strong law of large numbers applied in turn to the martingales

$$\sum_{i=1}^n \left(X_i^2 - E\left(X_i^2 \,\middle|\, \mathcal{F}_{i-1} \right) \right),$$

$$\sum_{i=1}^{n} \left(X_i \tilde{X}_i - E\left(X_i \tilde{X}_i \,\middle|\, \mathcal{F}_{i-1} \right) \right)$$

and

$$\sum_{i=1}^{n} \left(\tilde{X}_i^2 - E\left(\tilde{X}_i^2 \,\middle|\, \mathcal{F}_{i-1} \right) \right),$$

we obtain after some straightforward analysis that

$$n^{-1} \sum_{i=1}^{n} X_{i-1}^2 \xrightarrow{\text{a.s.}} (1-\alpha^2)^{-1}, \tag{11.46}$$

$$n^{-1} \sum_{i=1}^{n} \tilde{X}_{i-1} X_{i-1} \xrightarrow{\text{a.s.}} E(\epsilon\psi(\epsilon))(1-\alpha^2)^{-1}, \tag{11.47}$$

$$n^{-1} \sum_{i=1}^{n} \tilde{X}_{i-1}^2 \xrightarrow{\text{a.s.}} E\psi^2(\epsilon)(1-\alpha^2)^{-1} \tag{11.48}$$

as $n \to \infty$. Consequently,

$$I_{G_1^*} \sim nd\,(1-\alpha^2)^{-1} \quad \text{a.s.},$$

$$I_{S_1^*} \sim nd\,(E(\epsilon\,\psi(\epsilon)))^2\,(E\psi^2(\epsilon))^{-1}(1-\alpha^2)^{-1},$$

which gives

$$I_{S_2^*}\, I_{G_1^*}^{-1} \xrightarrow{\text{a.s.}} \operatorname{corr}^2(\epsilon, \psi(\epsilon)) \le 1$$

where corr denotes correlation.

The quantity $\operatorname{corr}^2(\epsilon, \psi(\epsilon))$ is the asymptotic relative efficiency of the estimator $\hat{\alpha}_S$ obtained from the estimating equation $S_{n1}^* = 0$ compared with the estimator $\hat{\alpha}_G$ obtained from the estimating equation $G_{n1}^* = 0$. This follows along the lines of (11.42), (11.43), which are readily formalized in the present context.

We also note that the standard (nonrobust) estimator for α is

$$\hat{\alpha}_{QL} = \sum_{i=1}^{n} X_i X_{i-1} \Big/ \sum_{i=1}^{n} X_{i-1}^2,$$

which is a quasi-score estimating function within the class

$$\sum_{i=1}^{n} a_{i-1}(X_i - \alpha\, X_{i-1}),$$

a_{i-1} being $\mathcal{F}_{i-1}$-measurable. It is easily shown that

$$n^{1/2}(\hat{\alpha}_{QL} - \alpha) \xrightarrow{\text{d}} N(0, 1-\alpha^2),$$

so that the asymptotic efficiency of $\hat{\alpha}_S$ relative to $\hat{\alpha}_{QL}$ is $d\,\operatorname{corr}^2(\epsilon, \psi(\epsilon))$.

11.6 Recursive Estimation

There are important applications in which data arrive sequentially in time and parameter estimates need to be updated expeditiously, for example, for on-line signal processing. We shall illustrate the general principles in this section.

To focus discussion we shall, following Thavaneswaran and Abraham (1988), consider a time series model of the form

$$X_t = \theta\, f_{t-1} + \epsilon_t, \tag{11.49}$$

where θ is to be estimated, f_{t-1} is measurable w.r.t. $\mathcal{F}_{t-1}$, the σ-field generated by $X_{t-1}, \dots, X_1$ and the $\{\epsilon_t, \mathcal{F}_t\}$ are martingale differences. This framework covers many non-linear time series models such as random coefficient autoregressive and threshold autoregressive models. However, we shall just treat the scalar case here for simplicity.

For the family of estimating functions

$$\mathcal{H}_T = \left\{ \sum_{t=1}^{T} a_t (X_t - \theta\, f_{t-1}), \quad a_t \;\; \mathcal{F}_{t-1} \,\text{measurable} \right\}$$

we have that the quasi-score estimating function is

$$G_T^* = \sum_{t=1}^{T} a_t^* \, (X_t - \theta\, f_{t-1}),$$

where

$$a_t^* = E\left(\dot{h}_t \,|\, \mathcal{F}_{t-1}\right) \Big/ E\left(h_t^2 \,|\, \mathcal{F}_{t-1}\right) = f_{t-1} \Big/ E\left(\epsilon_t^2 \,|\, \mathcal{F}_{t-1}\right).$$

The optimal estimator for θ based on $\mathcal{H}_T$ and the first T observations is then given by

$$\hat{\theta}_T = \sum_{t=1}^{T} a_t^* \, X_t \Big/ \sum_{t=1}^{T} a_t^* \, f_{t-1}. \tag{11.50}$$

When the $(T+1)$st observation becomes available, we now carry out estimation within $\mathcal{H}_{T+1}$. The estimator for θ based on the first $(T+1)$ observations, $\hat{\theta}_{T+1}$, is given by (11.50) with T replaced by $(T+1)$ and

$$\hat{\theta}_{T+1} - \hat{\theta}_T = K_{T+1} \left(\sum_{t=1}^{T+1} a_t^* \, X_t - \hat{\theta}_T \, K_{T+1}^{-1} \right),$$

where

$$K_{T+1}^{-1} = \sum_{t=1}^{T+1} a_t^* \, f_{t-1}.$$

After a little more algebra we find that

$$K_{T+1} = \frac{K_T}{1 + f_T \, a_{T+1}^* \, K_T} \tag{11.51}$$

and

$$\hat{\theta}_{T+1} = \hat{\theta}_T + \frac{K_T\, a^*_{T+1}}{1 + f_T\, a^*_{T+1}\, K_T} \left(X_{T+1} - \hat{\theta}_T\, f_T \right). \tag{11.52}$$

The algorithm given by (11.52) provides the new estimate at time $(T+1)$ in terms of the old estimate at T plus an adjustment. This is based on the prediction error since the forecast of X_{T+1} at time T is $\hat{\theta}_T f_T$. From starting values θ_0 and K_0 the estimator can be calculated recursively from (11.51) and (11.52). In practice θ_0 and K_0 are often chosen on the basis of an additional data run. Various extensions are possible. A similar recursive procedure may be developed for a rather more complicated model that (11.49) proposed by Aase (1983). For details see Thavaneswaran and Abraham (1988).

Chapter 12

Consistency and Asymptotic Normality for Estimating Functions

12.1 Introduction

Throughout this book it has at least been implicit that consistency and asymptotic normality will ordinarily hold, under appropriate regularity conditions, just as for the case of ordinary likelihood. Sometimes we have been quite explicit about this expectation, such as with the meta theorem enunciated in Chapter 4. We have chosen not to attempt a substantiation of these principles because of the elusiveness of a satisfying general statement and the delicacy of the results as exercises in mathematics as distinct from the reality of practically relevant examples. Also, we have made it clear already that it is generally preferable to check directly, in any particular example, for consistency and asymptotic normality, rather than trying to check the conditions of a special purpose theorem.

When the quasi-score (or more generally, the estimating function) is a martingale, consistency and asymptotic nomality are usually straightforward to check directly using the strong law of large numbers (SLLN) and central limit theorem (CLT), respectively, for martingales. In this chapter we shall give some general consistency results and versions of the SLLN and CLT for martingales which enable most typical examples to be treated.

There are, of course, applications for which the martingale paradigm is not a satisfactory general tool. The most notable context in which this is the case is that of random fields. Sometimes causal representations of fields are possible for which martingale asymptotics works well but other limit theorems are necessary in general. Consistency results are usually straightforward to obtain, especially in the context of stationary fields where ergodic theorems can be used, but asymptotic normality poses more problems. A standard approach to the central limit theorem here has been to develop results based on the use of mixing conditions to provide the necessary asymptotic independence (e.g., Rosenblatt (1985), Doukhan (1994), Guyon (1995, Chapter 3)). However, mixing conditions are difficult, if not impossible, to check, so this approach is not entirely satisfactory. Fortunately, recent central limit results for random fields based on conditional centering (e.g., Comets and Janžura (1996)) seem very promising.

Non-random field examples of estimating functions that are not martingales can often be replaced, for asymptotic purposes, by martingales. Thus, for example, if

$$G_T(\theta) = \sum_{t=1}^{T} (X_t - \theta)$$

where X_t is the moving average $X_t = \theta + \epsilon_t - r\epsilon_{t-1}, \quad |r| < 1$, the ϵ's being i.i.d., then

$$G_T(\theta) = (1 - r) \sum_{t=1}^{T} \epsilon_t + r(\epsilon_T - \epsilon_0),$$

which is well approximated by the martingale $\{(1-r)\sum_{t=1}^{T}\epsilon_t\}$ for the purposes of studying the asymptotics of $\hat{\theta}_T = T^{-1}\sum_{t=1}^{T} X_t$.

12.2 Consistency

A range of consistency results for quasi-likelihood estimators, and more generally for the roots of estimating functions, can be developed to parallel corresponding ones for the maximum likelihood estimator (MLE). We therefore begin with a brief discussion of methods that have been established for this context.

There are two classical approaches for the MLE. These are the one due to Wald (1949), which yields consistency of global maximizers of the likelihood, and the one due to Cramér (1946), which yields a consistent sequence of local maximizers.

The Wald approach is not available in general in a quasi-likelihood setting. Certainly its adoption requires the quasi-score to be the derivative with respect to the parameter of a scalar objective function. However, there are certainly cases in which the principles can be used. For a discussion of the approach see, for example, Cox and Hinkley (1974, pp. 288–289). Similar comments on applicability also apply to some recent necessary and sufficient results of Vajda (1995). There is, however, a newly developed minimax approach to consistency of quasi-likelihood estimators (Li (1996a)) that has some of the advantages of the Wald approach.

The Cramér approach is based on local Taylor series expansion of the likelihood function and an analogous route can easily be followed for estimating functions. It may be remarked that the law of large numbers plays a central role in both approaches.

We shall use a stochastic process setting to describe the ideas behind the Cramér approach. Let $L_T(\{X_t\};\, \theta)$ denote the likelihood that we suppose is differentiable with respect to θ, here assumed to be scalar for clarity, and for which the derivative of the log likelihood is square integrable. Then, under modest regularity conditions which involve differentiability under an integral sign, $\{d \log L_T(\{X_t\};\, \theta)/d\theta,\ \mathcal{F}_T,\ T \geq 1\}$ is a martingale, $\mathcal{F}_t$ denoting the σ-field generated by $X_1, \ldots, X_t,\ t \geq 1$, e.g., Hall and Heyde (1980, Chapter 6, p. 157).

We write

$$\frac{d}{d\theta}\log L_T(\{X_t\};\,\theta)=\sum_{t=1}^{T}u_t(\theta),$$

where the $u_t(\theta)$ are martingale differences, and take the local Taylor series expansion for $\theta' \in \Theta$. This gives

$$\begin{aligned}\frac{d}{d\theta}\log L_T(\{X_t\};\,\theta') &= \sum_{t=1}^{T}u_t(\theta')\\ &= \sum_{t=1}^{T}u_t(\theta)-(\theta'-\theta)\,I_T(\theta)\\ &\quad +(\theta'-\theta)(J_T(\theta_T^*)-I_T(\theta)),\end{aligned} \tag{12.1}$$

where

$$I_T(\theta)=\sum_{t=1}^{T}E\left(u_t^2(\theta)\,\middle|\,\mathcal{F}_{t-1}\right),\quad J_T(\theta)=\sum_{t=1}^{T}\frac{d}{d\theta}\,u_i(\theta),$$

and $\theta_T^*=\theta+\gamma(\theta'-\theta)$ with $\gamma=\gamma(\{X_t\};\,\theta)$ satisfying $|\gamma|<1$.

Now the martingale strong law of large numbers (e.g., using Theorem 2.18 of Hall and Heyde (1980); see pp. 158, 159, or Theorem 12.4 herein) gives

$$(I_T(\theta))^{-1}\sum_{i=1}^{T}u_i(\theta)\xrightarrow{\text{a.s.}}0 \tag{12.2}$$

provided that $I_T(\theta)\xrightarrow{\text{a.s.}}\infty$ as $T\to\infty$. Thus, using (12.1) and (12.2), the score function has a root $\hat{\boldsymbol{\theta}}_T$, which is strongly consistent for the true parameter if $I_T(\theta)\xrightarrow{\text{a.s.}}\infty$ and

$$\limsup_{T\to\infty}\left(\left|I_T(\theta)+J_T(\theta_T^*)\right|\,/I_T(\theta)\right)<1\quad\text{a.s.} \tag{12.3}$$

Various sufficient conditions can be imposed to ensure that (12.3) holds but these are not illuminating.

The approach discussed above can be paralleled in many quasi-likelihood contexts. If convenient families of martingale estimating functions are available for a particular context, they would ordinarily be used. Then the quasi-score estimating function will be a martingale, local Taylor series expansion can be used just as above, and the martingale strong law of large numbers retains its role. The case of vector valued $\boldsymbol{\theta}$ can be dealt with in the same way.

The different approaches to establishing consistency results depend fundamentally on whether the estimating function is the derivative with respect to $\boldsymbol{\theta}$ of a scalar objective function. If such a representation is possible, as it is in the case of the score function (which is the derivative of the log likelihood), then the consistency question can be translated into the problem of checking for local maxima. However, the representation is not possible in general for quasi-score estimating functions; a scalar objective function (quasi-likelihood) may not exist.

If there is a scalar objective function and the parameter space Θ is a compact subset of p-dimensional Euclidean space, then an estimator obtained by maximization always exists. The situation, however, is more complicated if a scalar objective function does not exist or is not specified.

Example (Chen (1991)). Consider the quadratic estimating function

$$g = \sum_{i=1}^{T} \left[a(X_i - \theta) + b\left((X_i - \theta)^2 - \sigma^2\right)\right],$$

where the X_i's are i.i.d. with $EX_i = \theta$, $\operatorname{var} X_i = \sigma^2$, a, b and σ^2 being known and θ is the parameter to be estimated. This is a quasi-score estimating function if $\mu_3\, a + (\mu_4 - \sigma^4)\, b = 0$, where $\mu_k = E(X_1 - \theta)^k$.

It is easily seen that there are two roots for $g(\theta) = 0$, namely,

$$\hat{\theta}_{1,T} = \frac{a}{2b} + \frac{1}{T}\sum_{i=1}^{T} X_i - \Delta_T^{1/2}, \quad \hat{\theta}_{2,T} = \frac{a}{2b} + \frac{1}{T}\sum_{i=1}^{T} X_i + \Delta_T^{1/2},$$

where

$$\Delta_T = \left(\frac{a}{2b} + \frac{1}{T}\sum_{i=1}^{T} X_i\right)^2 - \left(\frac{a}{Tb}\sum_{i=1}^{T} X_i + \frac{1}{T}\sum_{i=1}^{T} X_i^2 - \sigma^2\right).$$

The roots only exist on the sequence of events $\{\Delta_T \geq 0\}$, but it follows from the strong law of large numbers that

$$\Delta_T \xrightarrow{\text{a.s.}} (a/2b)^2$$

as $T \to \infty$, so that they exist almost surely. Note that $\hat{\theta}_{1,T} \xrightarrow{\text{a.s.}} \theta$, but $\hat{\theta}_{2,T} \xrightarrow{\text{a.s.}} \theta + (a/b)$. The point here is that both the existence and the consistency of estimators can at most be expected in the almost sure sense.

For various problems we have to deal with situations in which the asymptotic results on which we rely hold only on a set E, say, with $P(E) < 1$. This is the case, for example, for estimating the mean of the offspring distribution for a supercritical branching process; the asymptotic results hold on the nonextinction set whose probability is less than one in general.

Thus, in view of the above considerations, and given a sequence of estimating functions $\{\boldsymbol{G}_T(\boldsymbol{\theta})\}$, we shall say that existence and consistency of the estimator $\hat{\boldsymbol{\theta}}$ that solves $\boldsymbol{G}_T(\boldsymbol{\theta}) = \mathbf{0}$ holds if: for all sufficiently small $\delta > 0$, a.s. on E, and for T sufficiently large, $\boldsymbol{G}_T(\boldsymbol{\theta}) = \mathbf{0}$ has a root in the sphere $\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta\}$, $\boldsymbol{\theta}_0$ denoting the "true value".

We shall now give a consistency criterion that does not require the estimating function to be the derivative with respect to $\boldsymbol{\theta}$ of a scalar function. This is an adaption of a result originally due to Aitchison and Silvey (1958), which uses the Brouwer fixed-point theorem.

Theorem 12.1 (Consistency Criterion) Let $\{\boldsymbol{G}_T(\boldsymbol{\theta})\}$ be a sequence of estimating functions that are continuous in $\boldsymbol{\theta}$ a.e. on E for $T \geq 1$, and for all small $\delta > 0$ a.e. on E, there an $\epsilon > 0$ so that

$$\limsup_{T\to\infty} \left(\sup_{\|\theta-\theta_0\|=\delta} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \boldsymbol{G}_T(\boldsymbol{\theta}) \right) < -\epsilon, \tag{12.4}$$

then there exists a sequence of estimators $\hat{\boldsymbol{\theta}}_T$ such that for any $\omega \in E$, $\hat{\boldsymbol{\theta}}_T(\omega) \to \boldsymbol{\theta}_0$ and $\boldsymbol{G}_T(\hat{\boldsymbol{\theta}}_T(\omega)) = 0$ when $T > T_\omega$, a constant depending on ω.

On the other hand, if instead we have an in probability version of (12.4), namely,

$$\lim_{T\to\infty} P\left(\sup_{\|\theta-\theta_0\|=\delta} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \boldsymbol{G}_T(\boldsymbol{\theta}) < -\epsilon \,\middle|\, E \right) = 1 \tag{12.5}$$

for any $\epsilon > 0$, then there is a sequence of estimators $\hat{\boldsymbol{\theta}}_T$ such that $\hat{\boldsymbol{\theta}}_T$ converges to $\boldsymbol{\theta}_0$ in probability on E and

$$P\left(\boldsymbol{G}_T(\hat{\boldsymbol{\theta}}_T) = \mathbf{0} \,\middle|\, E\right) \longrightarrow 1$$

as $T \to \infty$.

Proof For any suitable small positive ϵ and δ let

$$E_T = \left\{ \omega : \sup_{\|\theta-\theta_0\|=\delta} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \boldsymbol{G}_T(\boldsymbol{\theta}) \leq -\epsilon \right\}.$$

For $\omega \in E_T$, define $\hat{\boldsymbol{\theta}}(\omega)$ to be a root of $\boldsymbol{G}_T(\boldsymbol{\theta}) = \mathbf{0}$ with $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \leq \delta$. The existence of such a root is ensured by Brouwer's fixed point theorem. Next, define $\hat{\boldsymbol{\theta}}_T(\omega)$ for $\omega \in E_T^c$, the complementary event, to be an arbitrary point in the parameter space.

Then, from the definition of $\hat{\boldsymbol{\theta}}_T$, we have

$$\{\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| < \delta\} \supset E_T$$

so that

$$\{\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| \geq \delta\} \subset E_T^c$$

and hence

$$P(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| \geq \delta \text{ i.o.} \,|\, E) \leq P(E_T^c \text{ i.o.} \,|\, E) = 0,$$

i.o. denoting infinitely often and the right-hand equality following from condition (12.4). This gives a.s. convergence of $\hat{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}_0$ on E. Furthermore, since $P(E_T^c \text{ i.o.} \,|\, E) = 0$, we have $\boldsymbol{G}_T(\hat{\boldsymbol{\theta}}_T(\omega)) = \mathbf{0}$ on E when $T > T_\omega$.

In the case where (12.5) holds, the results follow from

$$P(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| < \delta \,|\, E) \geq P(E_T \,|\, E) \to 1$$

as $T \to \infty$.

Theorem 12.1 is not usually easy to use in practice. The principal difficulty is in showing that (12.4) holds uniformly in $\boldsymbol{\theta}$. Some sufficient conditions for ensuring applicability are given in Theorem 12.1 of Hutton, Ogunyemi and Nelson (1991). However, direct checking is usually preferable, albeit tedious. See Exercises 12.1, 12.2 for example.

Next we shall consider the case where $\boldsymbol{G}_T(\boldsymbol{\theta})$ is the gradient of a scalar potential function $q(\boldsymbol{\theta})$, $\boldsymbol{G}_T(\boldsymbol{\theta}) = \partial q(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. For this to be the case the matrix $\partial\boldsymbol{G}_T(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ must be symmetrical (e.g., McCullagh and Nelder (1989)). The results that can then be obtained parallel these which have been developed for maximum likelihood estimators. The scalar potential function takes the place of the likelihood.

In the following we will use $\|\boldsymbol{x}\|$ for vector $\boldsymbol{x}$ to denote the Euclidean norm $(\boldsymbol{x}'\,\boldsymbol{x})^{1/2}$ and $\|\boldsymbol{A}\|$ for matrix $\boldsymbol{A}$ to denote the corresponding matrix norm $(\lambda_{\max}\boldsymbol{A}'\,\boldsymbol{A})^{1/2}$, $\lambda_{\max}(\text{resp.}\,\lambda_{\min})$ being the maximum (resp. minimum) eigenvalue.

For a sequence $\{\boldsymbol{A}_T\}$ of $p \times p$ matrices, possibly random or depending on the true parameter value $\boldsymbol{\theta}_0$, we shall define the sets

$$M_T(\delta) = \left\{\boldsymbol{\theta} \in \Re^p : \|\boldsymbol{A}_T(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\| \leq \delta\,\|(\boldsymbol{A}_T')^-\,\boldsymbol{G}_T(\boldsymbol{\theta}_0)\|\right\},$$

$T = 1, 2, \ldots,\ 0 < \delta < \infty$, the minus denoting generalized inverse. Then, we shall make use of the following consistency conditions, W (for weak) and S (for strong). These conditions refer to the true probability measure and hence they must usually be checked for all possible $\boldsymbol{\theta}_0$, with c perhaps depending on $\boldsymbol{\theta}_0$.

Condition W (i) For all $0 < \delta < \infty$,

$$\sup_{\theta \in M_T(\delta)} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \xrightarrow{\text{p}} 0.$$

(ii) For some random variable c, $0 < c < \infty$, and $\delta = 2/c$,

$$P(-\dot{\boldsymbol{G}}_T(\boldsymbol{\theta}) - c\boldsymbol{A}_T'\,\boldsymbol{A}_T \geq \mathbf{0} \ \text{ for all } \boldsymbol{\theta} \in M_T(\delta) \cap \Theta) \to 1,$$

Θ being the parameter space, an open subset of $\Re^p$.

Condition S (i) For all $0 < \delta < \infty$,

$$\sup_{\theta \in M_T(\delta)} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \xrightarrow{\text{a.s.}} 0.$$

(ii) For some random variables T_1, c, $0 < c < \infty$, and $\delta = 2/c$, $-\dot{\boldsymbol{G}}_T(\boldsymbol{\theta}) - c\boldsymbol{A}_T'\,\boldsymbol{A}_T \geq \mathbf{0}$ for all $\boldsymbol{\theta} \in M_T(\delta) \cap \Theta$ and all $T \geq T_0$, $\boldsymbol{A}_T$ is nonsingular, $T \geq T_0$, for some (random) T_0 a.s., and

$$\|(\boldsymbol{A}_T')^-\,\boldsymbol{G}_T(\boldsymbol{\theta}_0)\|^2 = o\left(\lambda_{\min}(\boldsymbol{A}_T'\,\boldsymbol{A}_T)\right) \quad \text{a.s.}$$

Some explanations are in order. The Condition W(i) can be thought of as $M_T(\delta)$ shrinking to $\boldsymbol{\theta}_0$ in probability and it is equivalent to

$$P(\boldsymbol{A}_T \,\text{nonsingular}) \to 1,$$

$$\|(\boldsymbol{A}_T')^{-}\, \boldsymbol{G}_T(\boldsymbol{\theta}_0)\|^2 = o_p\,(\lambda_{\min}(\boldsymbol{A}_T'\, \boldsymbol{A}_T)),$$

o_p meaning smaller order in probability, from which the parallel with Condition S is evident.

Theorem 12.2 Under Condition W, there exists a weakly consistent sequence $\{\hat{\boldsymbol{\theta}}_T\}$ of estimators. Actually, with δ from W(ii), we have

$$P(\hat{\boldsymbol{\theta}}_T \in M_T(\delta)) \to 1$$

and $\hat{\boldsymbol{\theta}}_T \xrightarrow{\text{p}} \boldsymbol{\theta}_0$.

To facilitate comparison with published results, assume that

$$\|(A_T')^{-}\, G_T(\theta)\| = O_p(1),$$

order one in probability having the obvious meaning. Then, it is natural to consider the simpler sets

$$N_T(\delta) = \{\boldsymbol{\theta} \in \Re^p :\ \|\boldsymbol{A}_T(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\| \le \delta\},\, T = 1, 2, \ldots$$

$0 < \delta < \infty$ and it is not difficult to see that the Condition W(i) is implied by the analogous assumption with $N_T(\delta)$ replacing $M_T(\delta)$. However, no such simple substitution is possible for W(ii). The usual Cramér type results on weak consistency in the literature (e.g., Basawa and Scott (1983)) are essentially covered by this formulation. A plausible choice for the norming matrix $\boldsymbol{A}_T$ is usually $(-E_G\, \dot{\boldsymbol{G}}_T(\boldsymbol{\theta}))^{1/2}$.

For the case of strong consistency we have the following result.

Theorem 12.3 Under Condition S, there exists a strongly consistent sequence of estimators $\{\boldsymbol{\theta}_T\}$. We have

$$\hat{\boldsymbol{\theta}}_T \in M_T(\delta) \quad \text{a.s.} \quad \text{for all } T \ge T_2$$

with some random T_2 and $\hat{\boldsymbol{\theta}}_T \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_0$.

Proof of Theorem 12.2 If Condition W(i) holds for all constants δ, $0 < \delta < \infty$, it also holds if δ is a random variable with $0 < \delta < \infty$. Let c, $\delta = 2/c$ be the random variables of W(ii) and denote by E_T the combined event:

$\boldsymbol{A}_T$ is nonsingular,
$M_T(\delta) \subset \Theta$,

$$-\dot{\boldsymbol{G}}_T(\boldsymbol{\theta}) - c\,\boldsymbol{A}_T'\,\boldsymbol{A}_T \ge \boldsymbol{0} \quad \text{for all} \quad \boldsymbol{\theta} \in M_T(\delta) \cap \Theta.$$

Conditions W(i),(ii) and Θ open ensure $P(E_T) \to 1$.

On the subset of E_T where $\boldsymbol{G}_T(\boldsymbol{\theta}_0) = \mathbf{0}$, $M_T(\delta)$ is degenerate to $\{\boldsymbol{\theta}_0\}$. Taking $\hat{\boldsymbol{\theta}}_T = \boldsymbol{\theta}_0$, we have $\boldsymbol{G}_T(\hat{\boldsymbol{\theta}}_T) = \mathbf{0}$, $-\dot{\boldsymbol{G}}_T(\hat{\boldsymbol{\theta}}_T) > 0$ and $\hat{\boldsymbol{\theta}}_T$ is a (locally unique) maximizer of q with $\hat{\boldsymbol{\theta}}_T \in M_T(\delta)$.

On the subset of E_T where $\boldsymbol{G}_T(\boldsymbol{\theta}_0) \neq \mathbf{0}$, $M_T(\delta) \subset \Theta$ is a compact ellipsoidal neighborhood of $\boldsymbol{\theta}_0$. For $\boldsymbol{\theta}$ on the boundary of $M_T(\delta)$, set $\boldsymbol{v} = \boldsymbol{A}_T(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. Then, taking the Taylor expansion

$$q_T(\boldsymbol{\theta}) - q_T(\boldsymbol{\theta}_0) = \boldsymbol{v}'(\boldsymbol{A}_T')^{-1}\,\boldsymbol{G}_T(\boldsymbol{\theta}) - \frac{1}{2}\,\boldsymbol{v}'\,(\boldsymbol{A}_T')^{-1}\,\dot{\boldsymbol{G}}_T(\tilde{\boldsymbol{\theta}})\,\boldsymbol{A}_T^{-1}\,\boldsymbol{v}$$

with $\tilde{\boldsymbol{\theta}} \in M_T(\delta)$, we obtain

$$\begin{aligned} q_T(\boldsymbol{\theta}) - q_T(\boldsymbol{\theta}_0) &\leq \|\boldsymbol{v}\|\,\|(\boldsymbol{A}_T')^{-1}\,\boldsymbol{G}_T(\boldsymbol{\theta})\| - \frac{1}{2}\,\|\boldsymbol{v}\|^2\,\lambda_{\min}\boldsymbol{A}_T^{-1}\,\dot{\boldsymbol{G}}_T(\tilde{\boldsymbol{\theta}})(\boldsymbol{A}_T')^{-1} \\ &\leq \|\boldsymbol{v}\|\left(\|(\boldsymbol{A}_T')^{-1}\,\boldsymbol{G}_T(\boldsymbol{\theta})\| - \frac{1}{2}\,\delta\,\|(\boldsymbol{A}_T')^{-1}\,\boldsymbol{G}_T(\boldsymbol{\theta})\|\,c\right) = 0. \end{aligned}$$

Let M denote the subset of $M_T(\delta)$ where $\sup_{\theta \in M_T(\delta)} \delta_T(\boldsymbol{\theta})$ is attained. Since $M_T(\delta)$ is compact, M is nonempty. Also, since $q_T(\boldsymbol{\theta}) \leq q_T(\boldsymbol{\theta}_0)$ for all $\boldsymbol{\theta}$ on the boundary of $M_T(\delta)$, if M were to contain the whole line segment between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$, it would also contain the whole line segment between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. But $-\dot{\boldsymbol{G}}_T(\tilde{\boldsymbol{\theta}}) > 0$, $\tilde{\boldsymbol{\theta}} \in M_T(\delta)$ makes this impossible and we conclude that M consists of a unique point in the interior of $M_T(\delta)$. This point defines a (measurable) local maximizer of $q(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ on the set $E_T \cap \{\boldsymbol{G}_T(\boldsymbol{\theta}_0) \neq \mathbf{0}\}$.

Proof of Theorem 12.3 Let $\{E_T\}$ be defined as in the proof of Theorem 12.2. Using the fact that $\liminf_{T\to\infty} E_T$ requires E_T to occur for all $T \geq T_2$, say, it is seen that Condition S implies $\liminf_{T\to\infty} E_T = 1$ with probability one. Arguing as above, $\hat{\boldsymbol{\theta}}_T \in M_T(\delta)$, $T \geq T_2$, and $\hat{\boldsymbol{\theta}}_T \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_0$ follow.

The general results given above obscure the simplicity of many applications. Very often the estimating functions $\{\boldsymbol{G}_T(\boldsymbol{\theta})\}$ of interest form a square integrable martingale. Then, direct application of an appropriate strong law of large numbers (SLLN) for martingales will usually give strong consistency of the corresponding estimator without requiring the checking of uniwieldy conditions. Useful general purpose SLLN's for martingales are given in Section 12.3 and a corresponding central limit result is given in Section 12.4.

12.3 The SLLN for Martingales

The first result given here is a multivariate version of the standard SLLN for martingales. It is from Lin (1994b, Theorems 2 and 3) and it extends work of Kaufmann (1987) to allow for random norming.

The martingales here will be assumed to be vectors of dimension $p \geq 1$. As usual we shall use the norm $\|\boldsymbol{x}\| = (\boldsymbol{x}'\,\boldsymbol{x})^{1/2}$ for vector $\boldsymbol{x}$ and, for matrix $\boldsymbol{A}$ we shall write $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ for the maximum and minimum eigenvalues, $\operatorname{tr}\boldsymbol{A}$ for the trace and $\|\boldsymbol{A}\| = (\lambda_{\max}(\boldsymbol{A}'\,\boldsymbol{A}))^{1/2}$ for the norm. A sequence of matrices $\{\boldsymbol{A}_n\}$ will be called monotonically increasing if $\boldsymbol{A}'_{n+1}\,\boldsymbol{A}_{n+1} - \boldsymbol{A}'_n\,\boldsymbol{A}_n$ is nonnegative definite for all n.

Theorem 12.4 Let $\{\boldsymbol{S}_n = \sum_{i=1}^n \boldsymbol{X}_i,\ \mathcal{F}_n,\ n \geq 1\}$ be a p-dimensional square integrable martingale and $\{\boldsymbol{A}_n\}$ a monotone increasing sequence of nonsingular symmetric matrices such that $\boldsymbol{A}_i$ is $\mathcal{F}_{i-1}$-measurable for each i, and $\lambda_{\min}(\boldsymbol{A}_n) \to \infty$ a.s. as $n \to \infty$. If

$$\sum_{n=1}^{\infty} E\|\boldsymbol{A}_n^{-1}\,\boldsymbol{X}_n\|^2 < \infty \tag{12.6}$$

or

$$\sum_{n=1}^{\infty} E\left(\|\boldsymbol{A}_n^{-1}\,\boldsymbol{X}_n\|^2 \,\middle|\, \mathcal{F}_{n-1}\right) < \infty \quad \text{a.s.}, \tag{12.7}$$

then $\boldsymbol{A}_n^{-1}\,\boldsymbol{S}_n \xrightarrow{\text{a.s.}} \boldsymbol{0}$ as $n \to \infty$.

Remarks

(1) Since

$$\|\boldsymbol{A}_n^{-1}\,\boldsymbol{X}_n\|^2 = \operatorname{tr}\,\boldsymbol{A}_n^{-1}\,\boldsymbol{X}_n\,\boldsymbol{X}'_n\,\boldsymbol{A}_n^{-1},$$

we have

$$\begin{aligned} E\,\|\boldsymbol{A}_n^{-1}\,\boldsymbol{X}_n\|^2 &= \operatorname{tr}\,\boldsymbol{A}_n^{-1}\,E(\boldsymbol{X}_n\,\boldsymbol{X}'_n)\,\boldsymbol{A}_n^{-1}, \\ E\left(\|\boldsymbol{A}_n^{-1}\,\boldsymbol{X}_n\|^2 \,\middle|\, \mathcal{F}_{n-1}\right) &= \operatorname{tr}\,\boldsymbol{A}_n^{-1}\,E\left(\boldsymbol{X}_n\,\boldsymbol{X}'_n \,\middle|\, \mathcal{F}_{n-1}\right)\,\boldsymbol{A}_n^{-1}. \end{aligned}$$

(2) Theorem 12.4 is for the case of discrete time but corresponding results are available for continuous time with the conditions (12.6) and (12.7) replaced by

$$\operatorname{tr}\left(\int_0^{\infty} \boldsymbol{A}_t^{-1}\,d\,(E\langle \boldsymbol{S}\rangle_t)\,\boldsymbol{A}_t^{-1}\right) < \infty \tag{12.8}$$

and

$$\operatorname{tr}\left(\int_0^{\infty} \boldsymbol{A}_t^{-1}\,d\langle \boldsymbol{S}\rangle_t\,\boldsymbol{A}_t^{-1}\right) < \infty \quad \text{a.s.}, \tag{12.9}$$

respectively, $\{\langle \boldsymbol{S}\rangle_t\}$ being the quadratic characteristic process for $\{\boldsymbol{S}_t\}$.

In some applications it may be convenient to use one dimensional martingale results on each component in the vector rather than the multivariate results quoted above. A comprehensive collection of one dimensional stong law results for martingales in discrete time are given in Chapter 2 of Hall and Heyde (1980). Again these results have natural analogues in continuous time. However, an additional one dimensional result for continuous time which is particularly useful is given in the following theorem due to Lépingle (1977).

Theorem 12.5 Let $\{S_t,\, \mathcal{F}_t,\, t \geq 0\}$ be a locally square integrable martingale that is right continuous with limits from the left (cadlag) and $h(x)$ be an increasing nonnegative function such that

$$\int_0^\infty (h(x))^{-2}\, dx < \infty.$$

Then, $S_t/h(\langle S\rangle_t) \to 0$ a.s. on the set $\{\langle S\rangle_t \to \infty\}$.

Proof of Theorem 12.4 Suppose $\sup_n E(\sum_{k=1}^n \|\boldsymbol{A}_k^{-1}\, \boldsymbol{X}_k\|^2) < \infty$. Then, since

$$\|\boldsymbol{x}+\boldsymbol{y}\|^2 - \|\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + 2\boldsymbol{x}'\boldsymbol{y},$$

we have, upon setting $\boldsymbol{x} = \boldsymbol{A}_k^{-1}\, \boldsymbol{X}_k$ and $\boldsymbol{y} = \boldsymbol{A}_k^{-1}\, \boldsymbol{S}_{k-1}$ and using the martingale property,

$$E\|\boldsymbol{A}_k^{-1}\, \boldsymbol{S}_k\|^2 - E\|\boldsymbol{A}_k^{-1}\, \boldsymbol{S}_{k-1}\|^2 = E\|\boldsymbol{A}_k^{-1}\, \boldsymbol{X}_k\|^2.$$

Write

$$z_n = \sum_{k=1}^{n} \left(\|\boldsymbol{A}_k^{-1}\, \boldsymbol{S}_k\|^2 - \|\boldsymbol{A}_k^{-1}\, \boldsymbol{S}_{k-1}\|^2\right).$$

Now $\sup_n\, Ez_n < \infty$ and since $\{z_n,\, \mathcal{F}_n,\, n \geq 1\}$ is a nonnegative submartingale we must have the a.s. convergence of

$$z_n = \|\boldsymbol{A}_n^{-1}\, \boldsymbol{S}_n\|^2 + \sum_{k=1}^{n-1} \left(\|\boldsymbol{A}_k^{-1}\, \boldsymbol{S}_k\|^2 - \|\boldsymbol{A}_{k+1}^{-1}\, \boldsymbol{S}_k\|^2\right).$$

But, the monotonicity condition for the $\boldsymbol{A}$'s ensures that

$$\sum_{k=1}^{n-1} \left(\|\boldsymbol{A}_k^{-1}\, \boldsymbol{S}_k\|^2 - \|\boldsymbol{A}_{k+1}^{-1}\, \boldsymbol{S}_k\|^2\right) = z_n - \|\boldsymbol{A}_n^{-1}\, \boldsymbol{S}_n\|^2$$

is nondecreasing and it is a.s. bounded by $\sup_n\, z_n$, so that it must converge a.s. This forces $\|\boldsymbol{A}_n^{-1}\, \boldsymbol{S}_n\|^2$ to converge a.s. Hence, it suffices to check that $\|\boldsymbol{A}_n^{-1}\, \boldsymbol{S}_n\|^2 \xrightarrow{\text{P}} 0$ to ensure that $\|\boldsymbol{A}_n^{-1}\, \boldsymbol{S}_n\|^2 \xrightarrow{\text{a.s.}} 0$ (and, consequently, $\boldsymbol{A}_n^{-1}\, \boldsymbol{S}_n \xrightarrow{\text{a.s.}} \boldsymbol{0}$).

Now, for any $\epsilon > 0$ and $N < n$,

$$\begin{aligned} P\left(\|\boldsymbol{A}_n^{-1}\, \boldsymbol{S}_n\|^2 > \epsilon\right) &\leq P\left(\|\boldsymbol{A}_n^{-1}\, \boldsymbol{S}_N\|^2 > \frac{1}{2}\,\epsilon\right) \\ &\quad + P\left(\|\boldsymbol{A}_n\,(\boldsymbol{S}_n - \boldsymbol{S}_N)\|^2 > \frac{1}{2}\,\epsilon\right) \\ &= I_1 + I_2, \end{aligned} \tag{12.10}$$

say. But, using Chebyshev's inequality and the monotonicity of the $\boldsymbol{A}$'s

$$\begin{aligned}
\frac{1}{2}\,\epsilon\, I_2 &\le E\left\| \boldsymbol{A}_n^{-1} \sum_{k=N+1}^{n} \boldsymbol{X}_k \right\|^2 \\
&\le E\left(\sum_{k=N+1}^{n} \boldsymbol{X}_k'\, \boldsymbol{A}_k^{-2}\, \boldsymbol{X}_k + \sum_{k=N+2}^{n} \boldsymbol{X}_k'\, \boldsymbol{A}_k^{-2} \sum_{j=N+1}^{k-1} \boldsymbol{X}_j \right. \\
&\qquad \left. + \sum_{k=N+2}^{n} \boldsymbol{X}_k'\, \boldsymbol{A}_k^{-2} \sum_{j=k+1}^{n} \boldsymbol{X}_j \right) \\
&= E\left(\sum_{k=N+1}^{n} \boldsymbol{X}_k'\, \boldsymbol{A}_k^{-2}\, \boldsymbol{X}_k \right) = E \sum_{k=N+1}^{n} \|\boldsymbol{A}_k^{-1}\, \boldsymbol{X}_k\|^2,
\end{aligned}$$

since $\boldsymbol{A}_k$ is $\mathcal{F}_{k-1}$-measurable. Also, since $E\sum_{k=1}^{\infty} \|\boldsymbol{A}_k^{-1}\, \boldsymbol{X}_k\|^2 < \infty$, for given $\epsilon > 0$ there is an $N_0 > 0$ such that

$$I_2 \le E\left(\sum_{k=N_0+1}^{n} \|\boldsymbol{A}_k^{-1}\, \boldsymbol{X}_k\|^2 \right) \Big/ (\epsilon/2) < (\epsilon/2) \tag{12.11}$$

for all $n > N_0$. Now fix N_0. In view of the condition that $\lambda_{\min}(\boldsymbol{A}_n) \to \infty$ a.s. as $n \to \infty$ there is an $N > N_0$ such that if $n > N$,

$$I_1 \le P\left(\|\boldsymbol{A}_n^{-1}\|^2\, \|\boldsymbol{S}_{N_0}\|^2 > \epsilon/2 \right) < \epsilon/2 \tag{12.12}$$

and hence, from (12.10), (12.11) and (12.12),

$$P\left(\|\boldsymbol{A}_n^{-1}\, \boldsymbol{S}_n\|^2 > \epsilon \right) < \epsilon$$

for $n > N$. This gives $\|\boldsymbol{A}_n^{-1}\, \boldsymbol{S}_n\|^2 \xrightarrow{\text{p}} 0$ and hence $\boldsymbol{A}_n^{-1}\, \boldsymbol{S}_n \xrightarrow{\text{a.s.}} \mathbf{0}$ and completes the proof of the first part of the theorem.

Now suppose that (12.7) holds. Given a real $C > 0$, let

$$T_C = \max\left[n : \sum_{k=1}^{n} \left(E\left(\|\boldsymbol{A}_k^{-1}\, \boldsymbol{X}_k\|^2 \,\middle|\, \mathcal{F}_{k-1} \right) \right) \le C \right].$$

Note that T_C is a stopping time since, for any n_0,

$$\{T_C < n_0\} = \left\{ \sum_{k=1}^{n_0} E\left(\|\boldsymbol{A}_k^{-1}\, \boldsymbol{X}_k\|^2 \,\middle|\, \mathcal{F}_{k-1} \right) > C \right\} \in \mathcal{F}_{n_0-1}.$$

Therefore, for each C we can define a new random norming, such as $I(n \le T_C)\, \boldsymbol{A}_n^{-1}$, I denoting the indicator function, which is still $\mathcal{F}_{n-1}$-measurable. Clearly

$$\sum_{k=1}^{n} E\left(\|I(k \le T_C)\, \boldsymbol{A}_k^{-1}\, \boldsymbol{X}_k\|^2 \,\middle|\, \mathcal{F}_{k-1} \right) \le C$$

for all n and hence

$$\sum_{k=1}^{n} E\left(\|I(k \leq T_C)\, \boldsymbol{A}_k^{-1}\, \boldsymbol{X}_k\|^2\right) \leq C.$$

Also, $\{I(n \leq T_C)\, \boldsymbol{A}_n^{-1}\}$ is monotonically decreasing, so applying the result of the first part of the theorem we obtain

$$I(n \leq T_C)\, \boldsymbol{A}_n^{-1}\, \boldsymbol{S}_n \xrightarrow{\text{a.s.}} \boldsymbol{0}$$

and hence $\boldsymbol{A}_n^{-1}\, \boldsymbol{S}_n \xrightarrow{\text{a.s.}} \boldsymbol{0}$ and $\{T_C = \infty\}$. The required result follows since C can be any positive number.

Proof of Theorem 12.5 The result basically follows from the martingale convergence theorem and a Kronecker lemma argument.

Put $A_t = \langle S \rangle_t$, $c_u = \inf\{t :\, A_t > u\}$ and define the martingale

$$Z_t = \int_0^t (h(A_u))^{-1}\, dS_u.$$

As $A_{c_u} \geq u$ for all $u < A_\infty$,

$$\langle Z \rangle_\infty = \int_0^\infty \frac{dA_u}{(h(A_u))^2} = \int_0^{A_\infty} \frac{du}{(h(A_{c_u}))^2} \leq \int_0^\infty \frac{dt}{(h(t))^2} < \infty.$$

It follows that Z is square integrable and converges a.s.

Now, using integration by parts,

$$\begin{aligned} S_t = \int_0^t h(A_u)\, dZ_u &= Z_t\, h(A_t) - \int_0^t Z_{u-}\, dh(A_u) \\ &= \int_0^t (Z_t - Z_{u-})\, dh(A_u). \end{aligned}$$

Furthermore, if $0 < v < t < \infty$,

$$\begin{aligned} \left|\frac{S_t}{h(A_t)}\right| &\leq \frac{1}{h(A_t)} \left|\int_0^v (Z_t - Z_{u-})\, dh(A_u)\right| \\ &\quad + \frac{1}{h(A_t)} (h(A_t) - h(A_v)) \sup_{v < u \leq t} |Z_t - Z_{u-}|. \end{aligned}$$

The required result then follows by letting $t \to \infty$ and $v \to \infty$.

12.4 Confidence Intervals and the CLT for Martingales

Consistency or strong consistency of estimators obtained with the aid of the LLN for martingales is usually able to be complemented with a corresponding asymptotic normality result which follows from the central limit theorem (CLT) for martingales. This provides asymptotic confidence zones for the parameters of interest.

Suppose that $\{\boldsymbol{G}_T(\boldsymbol{\theta})\}$ is a martingale sequence of estimating functions and that the estimator $\hat{\boldsymbol{\theta}}_T \to \boldsymbol{\theta}_0$ and $\boldsymbol{G}_T(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ a.s. on an event E for $T \geq 1$. Then, consider the expansion

$$\mathbf{0} = \boldsymbol{G}_T(\hat{\boldsymbol{\theta}}_T) = \boldsymbol{G}_T(\boldsymbol{\theta}_0) + \dot{\boldsymbol{G}}_T(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) + \boldsymbol{r}_T(\hat{\boldsymbol{\theta}}_T)$$

on E. If, on E, for a sequence $\{\boldsymbol{H}_T(\boldsymbol{\theta})\}$ of matrices, possibly random,

$$\boldsymbol{H}_T^{-1}(\boldsymbol{\theta}_0)\, \boldsymbol{G}_T(\boldsymbol{\theta}_0) \xrightarrow{\mathrm{d}} \boldsymbol{Z}, \tag{12.13}$$

a proper law, and

$$\sup_{\{\boldsymbol{\theta}:\, \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq \delta\}} \|\boldsymbol{H}_T^{-1}(\boldsymbol{\theta}_0)\, \boldsymbol{r}_T(\boldsymbol{\theta})\| \xrightarrow{\mathrm{p}} 0 \tag{12.14}$$

for some $\delta > 0$, then

$$\boldsymbol{H}_T^{-1}(\boldsymbol{\theta}_0)(-\dot{\boldsymbol{G}}_T(\boldsymbol{\theta}_0))(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{\mathrm{d}} \boldsymbol{Z}.$$

This last result is the basis for confidence zones for $\boldsymbol{\theta}_0$ as discussed in Chapter 4 and, as indicated in Section 4.2, $-\dot{\boldsymbol{G}}_T(\boldsymbol{\theta}_0)$ may often be replaced asymptotically by equivalent normalizers $[\boldsymbol{G}(\boldsymbol{\theta}_0)]_T$ or $\langle \boldsymbol{G}(\boldsymbol{\theta}_0)\rangle_T$. The usual practice is for a martingale CLT to be used to provide (12.13), while (12.14) is obtained from ad hoc use of law of large number results and inequalities.

The multivariate central limit result given here is an adaption of a result of Hutton and Nelson (1984) to deal with more general norming sequences. A similar, but slightly weaker, result appears in Sørensen (1991). Our proof follows that of Hutton and Nelson (1984). The result is phrased to deal with the continuous time case but it also covers the discrete time one. This is dealt with by replacing a discrete time process $\{\boldsymbol{X}_n,\ n \geq 0\}$ by a continuous version $\{\boldsymbol{X}_t^c,\ t \geq 0\}$ defined through $\boldsymbol{X}_t^c = \boldsymbol{X}_n,\ n \leq t \leq n+1$. Here all the processes that are considered are assumed to be right continuous with limits from the left (cadlag) and defined on a complete filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ satisfying the standard conditions. As usual we denote by I_p the $p \times p$ identity matrix.

We shall make use of the concepts of stable and mixing convergence in distribution and these are defined below. Let

$$\boldsymbol{Y}_n = (Y_{1n}, \ldots, Y_{pn})' \xrightarrow{\mathrm{d}} \boldsymbol{Y} = (Y_1, \ldots, Y_p)'.$$

Suppose that $\boldsymbol{y} = (y_1, \ldots, y_p)'$ is a continuity point for the distribution function of Y and $E \in \mathcal{F}$. Then, we say that the convergence holds *stably* if

$$\lim_{n\to\infty} P(Y_{1n} \leq y_1, \ldots, Y_{pn} \leq y_p, E) = Q_y(E)$$

exists and

$$Q_y(E) \to P(E)$$

as $y_i \to \infty$, $i = 1, 2, \ldots, p$. On the other hand, we say that the convergence holds in the *mixing* sense if

$$\lim_{n\to\infty} P(Y_{1n} \le y_1, \ldots, Y_{pn} \le y_p, E) = P(Y_1 \le y_1, \ldots, Y_p \le y_p)\, P(E).$$

For details see Hall and Heyde (1980, pp. 56–57) and references therein.

Theorem 12.6 Let $\{\boldsymbol{S}_t = (S_{1t}, \ldots, S_{pt})', \mathcal{F}_t, t \ge 0\}$ be a p-dimensional martingale with quadratic variation matrix $[\boldsymbol{S}]_t$. Suppose that there exists a non-random vector function $\boldsymbol{k}_t = (k_{1t}, \ldots, k_{pt})$ with $k_{it} > 0$ increasing to infinity as $t \to \infty$ for $i = 1, 2, \ldots, p$ such that as $t \to \infty$:

(i) $k_{it}^{-1} \sup_{s\le t} |\Delta S_{is}| \xrightarrow{\text{p}} 0,\ i = 1, 2, \ldots, p,$
where $\Delta S_{is} = S_{is} - S_{is-}$;

(ii) $\boldsymbol{K}_t^{-1}\, [\boldsymbol{S}]_t\, \boldsymbol{K}_t^{-1} \xrightarrow{\text{p}} \boldsymbol{\eta}^2$
where $\boldsymbol{K}_t = \operatorname{diag}(k_{1t}, \ldots, k_{pt})$ and $\boldsymbol{\eta}^2$ is a random nonnegative definite matrix;

(iii) $\boldsymbol{K}_t^{-1} \left(E(\boldsymbol{S}_t\, \boldsymbol{S}_t')\right) \boldsymbol{K}_t^{-1} \to \boldsymbol{\Sigma}$
where $\boldsymbol{\Sigma}$ is a positive definite matrix. Then,

$$\boldsymbol{K}_t^{-1}\, \boldsymbol{S}_t \xrightarrow{\text{d}} \boldsymbol{Z} \quad \text{(stably)} \tag{12.15}$$

where the distribution of $\boldsymbol{Z}$ is the normal variance mixture with characteristic function $E(\exp(-\frac{1}{2}\, \boldsymbol{u}'\, \boldsymbol{\eta}^2\, \boldsymbol{u}))$, $\boldsymbol{u} = (u_1, \ldots, u_p)'$,

$$\left(\boldsymbol{K}_t\, [\boldsymbol{S}]_t^{-1}\, \boldsymbol{K}_t\right)^{1/2} \boldsymbol{K}_t^{-1} \left(\boldsymbol{S}_t \,\middle|\, \det\boldsymbol{\eta}^2 > 0\right) \xrightarrow{\text{d}} N(0, \boldsymbol{I}_p) \quad \text{(mixing)} \tag{12.16}$$

det denoting determinant, and

$$\left(\boldsymbol{S}_t'\, [\boldsymbol{S}]_t^{-1}\, \boldsymbol{S}_t \,\middle|\, \det\boldsymbol{\eta}^2 > 0\right) \xrightarrow{\text{d}} \chi_p^2 \quad \text{(mixing)} \tag{12.17}$$

as $t \to \infty$.

Remark It should be noted that $\{[\boldsymbol{S}]_t - \langle \boldsymbol{S}\rangle_t, \mathcal{F}_t,\ t \ge 0\}$ is a martingale, $\langle \boldsymbol{S}\rangle_t$ being the quadratic characteristic matrix. Each of the quantities $[\boldsymbol{S}]_t$, $\langle \boldsymbol{S}\rangle_t$ goes a.s. to infinity as $t \to \infty$ and ordinarily $[\boldsymbol{S}]_t$, $\langle \boldsymbol{S}\rangle_t$ are asymptotically equivalent.

Proof of Theorem 12.6 First we deal with the case $p = 1$ and for this purpose we shall use an adaption of Theorem 3.2 of Hall and Heyde (1980). As it stands, this theorem deals with finite martingale arrays $(S_{ni}, \mathcal{F}_{ni},\ 1 \le i \le k_n,\ n \ge 1)$ with differences X_{ni} and it gives sufficient conditions under which

$$Y_n \equiv \sum_{i=1}^{k_n} X_{ni} \xrightarrow{\text{d}} Z \quad \text{(stably)}$$

where Z is as defined in (12.15). But, if $\{k_n\}$ increases monotonically to infinity in such a way that

$$E\left(\sum_{i>k_n} X_{ni}^2\right) \longrightarrow 0$$

as $n \to \infty$, the theorem extends to infinite sums. That

$$\sum_{i=1}^{\infty} X_{ni} = \sum_{i=1}^{k_n} X_{ni} + \sum_{i>k_n} X_{ni} \xrightarrow{\text{d}} Z \quad \text{(stably)}$$

under the condition $\sum_{i>k_n} X_{ni} \xrightarrow{\text{L}_1} 0$ follows from the definition of stable convergence.

To prove (12.15) in the $p = 1$ case it suffixes to show that for any sequence $\{T_n\}$ diverging monotonically to infinity sufficiently fast that $K_{T_{n+1}}/K_{T_n} \geq 4$, we have

$$K_{T_n}^{-1}\, S_{T_n} \xrightarrow{\text{d}} Z \quad \text{(stably).} \tag{12.18}$$

Indeed, if $K_t^{-1}\, S_t$ does not converge stably we can choose a subsequence along which the convergence is not stable and then a sufficiently sparse subsequence $\{T_n\}$ for which $K_{T_{n+1}}/K_{T_n} \geq 4$ and $K_{T_n}^{-1}\, S_{T_n}$ does not converge stably thus giving a contradiction.

To show (12.18), define for each n the stopping times:

$$\begin{aligned} T_n^0 &= 0, \\ T_n^{k+1} &= \inf\left(t > T_n^k : \ K_{T_n}^{-1}|S_t - S_{T_n^k}| \geq 2^{-n}\right) \\ &= T_n \quad \text{if no such } t \leq T_n \text{ exists.} \end{aligned}$$

We shall consider the martingale differences

$$\left\{X_{ik} = K_{T_n}^{-1}\left(S_{T_n^k} - S_{T_n^{k-1}}\right),\ \mathcal{F}_{nk} = \mathcal{F}_{T_n^k}\right\}.$$

Note that $T_n^k = T_n$ ultimately in k for $n = 1, 2, \ldots$ so that $\sum_k X_{nk} = K_{T_n}^{-1}\, S_{T_n}$.

Now, to establish (12.15) we need to check

(a) $\sup_n E\left(\sum_k X_{nk}\right)^2 < \infty$,

(b) $\sup_k |X_{nk}| \xrightarrow{\text{p}} 0$,

(c) $\sum_k X_{nk}^2 \xrightarrow{\text{p}} \eta^2$,

(d) $E\left(\sup_k X_{nk}^2\right)$ is bounded in n,

(e) for all n and k,

$$\mathcal{F}_{nk} = \sigma(X_{n1}, \ldots, X_{nk}) \subset \mathcal{F}_{n+1,k} = \sigma(X_{n1}, \ldots, X_{(n+1)k}).$$

Now (a) is satisfied because $K_{T_n}^{-2}\,E\,S_{T_n}^2$ is bounded, (b) is satisfied since

$$S_{T_n^k} - S_{T_n^{k-1}} \le 2^{-n} + \sup_{t\le T_n} \Delta\, S_t,$$

(d) is satisfied since

$$\begin{aligned} K_{T_n}^{-2}\,E\left(\sup_k \left|S_{T_n^k} - S_{T_n^{k-1}}\right|^2\right) &\le K_{T_n}^{-2}\,E\,\sup_k\left(2S_{T_n^k}^2 + 2S_{T_n^{k-1}}^2\right)\\ &\le 4K_{T_n}^{-2}\,E\,S_{T_n}^2 \end{aligned}$$

and (c) holds, while (e) is satisfied since by induction on k, we can show that $T_n^k \le T_{n+1}^k$ a.s.

To check (c), it suffices to show that

$$\sum_k X_{nk}^2 - K_{T_n}^{-2}\,[S]_{T_n} \xrightarrow{\text{p}} 0.$$

Now, by Ito's formula,

$$[S]_{T_n} = S_{T_n}^2 - S_0^2 - 2\int_0^{T_n} S_{u-}\,dS_u,$$

and similarly,

$$\sum_k X_{nk}^2 = K_{T_n}^{-2}\left(S_{T_n}^2 - S_0^2 - 2\sum_{k=0}^{\infty} S_{T_n^k}\,(S_{T_n^{k+1}} - S_{T_n^k})\right).$$

Then, setting

$$\phi_n(u) = \sum_{k=0}^{\infty} S_{T_n^k}\,I(T_n^k < u \le T_n^{k+1}),$$

we have

$$\sum_k X_{nk}^2 - K_{T_n}^{-2}\,S_{T_n}^2 = 2\,K_{T_n}^{-2}\int_0^{T_n}(S_{u-} - \phi_n(u))\,dS_u = I_n,$$

say.

Now, from the definitions of T_n^k and ϕ_n, we have

$$K_{T_n}^{-1}\,(S_u - \phi_u(u)) \le 2^{-n} \quad \text{a.s.},$$

so that for $\epsilon > 0$,

$$\begin{aligned} P(|I_n| > \epsilon) &\le \epsilon^{-2}\,EI_n^2\\ &= 4\,\epsilon^{-2}\,K_{T_n}^{-4}\,E\int_0^{T_n}(S_{u-} - \phi_n(u))^2\,d[S]_u\\ &\le 4\,\epsilon^{-2}\,K_{T_n}^{-2}\,2^{-2n}\,E[S]_{T_n} \longrightarrow 0 \end{aligned}$$

as $n \to \infty$. The completes the proof in the case $p = 1$.

To deal with the general case we use the so-called Cramér-Wold device. Let $\boldsymbol{c} = (c_1, \dots, c_p)'$ be a nonzero vector. By applying the theorem to the martingale $\{R_t = \boldsymbol{c}'\, \boldsymbol{S}_t\}$, we have

$$(\boldsymbol{c}'\, E(\boldsymbol{S}_t\, \boldsymbol{S}_t')\, \boldsymbol{c})^{-1/2}\, (\boldsymbol{c}'\, \boldsymbol{S}_t) \xrightarrow{\text{d}} Z_c \quad \text{(stably)}$$

where

$$E(\exp(i\, u\, Z_c)) = E\left(\exp\left(-\frac{1}{2}\, u^2\, \eta_c^2\right)\right)$$

with η_c given by

$$(\boldsymbol{c}'\, E(\boldsymbol{S}_t\, \boldsymbol{S}_t')\, \boldsymbol{c})^{-1} [\boldsymbol{c}'\, \boldsymbol{S}]_t \xrightarrow{\text{p}} \eta_c^2.$$

But, using Conditions (ii) and (iii) of the theorem,

$$\eta_c^2 = (\boldsymbol{c}'\, \boldsymbol{\eta}^2\, \boldsymbol{c}) / (\boldsymbol{c}'\, \boldsymbol{\Sigma}\, \boldsymbol{c}).$$

It then follows that

$$\boldsymbol{c}'\, \boldsymbol{K}_t^{-1}\, \boldsymbol{S}_t \xrightarrow{\text{d}} (\boldsymbol{c}'\, \boldsymbol{\Sigma}\, \boldsymbol{c})^{1/2}\, Z_c \quad \text{(stably)}$$

and since the characteristic function of $(\boldsymbol{c}'\, \boldsymbol{\Sigma}\, \boldsymbol{c})^{1/2}\, Z_c$ is $E[\exp(-\frac{1}{2}\, \boldsymbol{c}'\, \boldsymbol{\eta}^2\, c\, u^2)]$, the result (12.15) follows. The remaining results (12.16) and (12.17) are easy consequences of the definition of stable convergence together with (12.15). This completes the proof.

A broad range of one-dimensional CLT results may be found in Chapter 3 of Hall and Heyde (1980). These may be turned into multivariate results using the Cramér-Wold device as in the proof of Theorem 12.6 above.

12.5 Exercises

1. Suppose that $\{Z_i\}$ is a supercritical, finite variance Galton-Watson branching process with $\theta = E(Z_1 \,|\, Z_0 = 1)$, the mean of the offspring distribution to be estimated from a sample $\{Z_1, \dots, Z_T\}$. Show that

$$Q_T(\theta) = \sum_{i=1}^{T} (Z_i - \theta\, Z_{i-1})$$

may be obtained as a quasi-score estimating function for θ from an appropriate family of estimating functions. Establish the strong consistency of the corresponding quasi-likelihood estimator on the non-extinction set
(i) using Theorem 12.1, and
(ii) using the martingale SLLN, Theorem 12.4.
(Hint. Note that $\theta^{-n}\, Z_n \xrightarrow{\text{a.s.}} W$ as $n \to \infty$ where $W > 0$ a.s. on the non-extinction set.)

2. Let $\{M_t, \mathcal{F}_t\}$ be a square integrable martingale with quadratic characteristic $\langle M \rangle_T = T$ a.s. Thus, $\{M_t\}$ could be a process with stationary independent increments and need not be Brownian motion. Consider the model

$$dX_t = \left(\theta_1\, \theta_2 + \theta_1 (c\, t^{c-1})^{1/2}\right)\, d\theta + dM_t,$$

$\theta_1 > 0,\ \theta_2 > 0,\ \frac{1}{2} < c < 1$, and obtain the Hutton-Nelson solution (Section 2.3) for the quasi-score estimating function for $\boldsymbol{\theta} = (\theta_1, \theta_2)'$. Show that the quasi-likelihood estimator is strongly consistent for $\boldsymbol{\theta}$ using Theorem 12.1. (Hint. Use Taylor expansion and Lépingle's version of the SLLN for martingales (Theorem 12.5).) (Hutton and Nelson (1986).)

3. Suppose that $\{X_k\}$ is a process and $\{\mathcal{F}_k\}$ an increasing sequence of σ-fields such that

$$E(X_k \,|\, \mathcal{F}_{k-1}) = \theta, \quad E((X_k - \theta)^2 \,|\, \mathcal{F}_{k-1}) = 1 + \theta\, u_{k-1},$$

$k \geq 1$, with $\{u_k\}$ a family of non-negative random variables adapted to $\{\mathcal{F}_k\}$. For a sample $\{X_1, \ldots, X_T\}$, obtain the estimating function

$$Q_T(\theta) = \sum_{k=1}^{T} \left(\frac{X_k - \theta}{1 + \theta\, u_{k-1}} \right) \tag{12.19}$$

as a quasi-score.

Noting the difficulty of working with this quasi-score, it is interesting to make a comparison with the simple, non-optimal, estimating function

$$H_T(\theta) = \sum_{k=1}^{T} (X_k - \theta)\, u_{k-1}^{-1}\, I\ (u_{k-1} > 0). \tag{12.20}$$

In the particular case where $(u_k^2, \mathcal{F}_k)$ is a Poisson process with parameter 1 sampled at integer times, and $X_k = \theta + w_k\, (1 + \theta\, u_{k-1})^{1/2},\ k \leq 1$, the $\{w_k\}$ being i.i.d. with finite fourth moment and $E(w_k \,|\, \mathcal{F}_{k-1}) = 0,\ E(w_k^2 \,|\, \mathcal{F}_{k-1}) = 1$, $k \geq 1$, investigate the consistency and asymptotic normality of the estimators obtained from (12.19) and (12.20) (Hutton, Ogunyemi and Nelson (1991)).

4. Consider the stochastic differential equation

$$dX_t = (\theta_1 + \theta_2\, X_t)\, dt + \sigma\, dW_t + dZ_t$$

where W_t denotes standard Brownian motion and Z_t is a compound Poisson process

$$Z_t = \sum_{i=1}^{N_t} \epsilon_i,$$

N_t being a Poisson process with parameter θ_3 and the ϵ_i's being i.i.d random variables with mean θ_4, which are independent of N_t. This s.d.e has been used to model the dynamics of soil moisture; see Mtundu and Koch (1987). For a sample $\{X_t,\, 0 \leq t \leq T\}$ investigate the consistency and asymptotic normality, of quasi-likelihood estimators for $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$. Take σ as known (Sørensen (1991)).

Chapter 13

Complements and Strategies for Application

13.1 Some Useful Families of Estimating Functions

13.1.1 Introduction

A broad array of families of estimating functions has been presented in this monograph but many others are available and may be of particular use in special circumstances. The focus of the theory has usually been on polynomial functions of the data and these will not always provide ideal families of estimating functions. It must be remembered that the general theory requires estimating functions which are at least square integrable and if the underlying variables of interest do not have finite second moments then transformations are required. A number of useful approaches are indicated below.

13.1.2 Transform Martingale Families

Suppose we have real valued observations $X_1, \dots, X_n$ whose distribution involves a parameter θ. Let $\mathcal{F}_j$ be the σ-field generated by $X_1, \dots, X_j$, $j \geq 1$. Then, following Merkouris (1992), write

$$F_j(x \,|\, \mathcal{F}_{j-1}) = P(X_j \leq x \,|\, \mathcal{F}_{j-1}), \qquad \hat{F}_j(x) = I(X_j \leq x),$$

I being the indicator function. Now introduce a suitably chosen index set of $\{g_t(x),\, t \in T \subseteq \Re\}$, which is real or complex valued and such that the integral transform

$$\phi_j(t) = E(g_t(X_j) \,|\, \mathcal{F}_{j-1}) = \int_{-\infty}^{\infty} g_t(x)\, dF_j(x \,|\, \mathcal{F}_{j-1})$$

exists and is finite for all $\theta \in \Theta$ and all $t \in T$. Transforms such as the characteristic function and moment generating function, for which $g_t(x)$ is e^{itx} and e^{tx}, respectively, are clearly included.

Now let

$$\hat{\phi}_j(t) = g_t(X_j) = \int_{-\infty}^{\infty} g_t(x)\, d\hat{F}_j(x)$$

and write

$$\begin{aligned} h_j(t) &= \hat{\phi}_j(t) - \phi_j(t) \\ &= g_t(X_j) - E(g_t(X_j) \,|\, \mathcal{F}_{j-1}). \end{aligned}$$

For fixed $t \in T$ the $\{h_j(t), \mathcal{F}_j\}$ are martingale differences and the family of martingale estimating functions

$$\mathcal{H} = \left\{ \sum_{j=1}^{n} a_j \, h_j(t), \qquad a_j\text{'s } \mathcal{F}_{j-1}\text{-measurable} \right\}$$

will be useful in various settings. For an application to estimation of the scale parameter in a Cauchy distribution using $g_t(x) = \cos tx$, see Section 2.8, Exercise 5.

Estimation of parameters of progression time distributions in multi-stage models is treated by this methodology, modulo minor adaption to the context, in Schuh and Tweedie (1970) and Feigin, Tweedie and Belyea (1983). It is shown that Laplace transform based estimators may be very efficient when maximum likelihood estimators are available.

13.1.3 Use of the Infinitesimal Generator of a Markov Process

A number of useful approaches to estimation are based on the use of the infinitesimal generator of a continuous time Markov process $\{X_t, \, t > 0\}$ with states in a sample space χ, say a locally compact metric space.

Suppose that X has distribution P_θ, $\theta \in \Theta$. Let $L = L_\theta$ be the infinitesimal generator of the Markov process X. That is, for suitable functions f, Lf is given by

$$\lim_{\delta \downarrow 0} \delta^{-1} [E(f(X_\delta) \,|\, X_0 = x) - f(x)]$$

provided that the limit exists (see, e.g., Karlin and Taylor (1981, p. 294)). Often a bounded limit is prescribed. Now let $\mathcal{D}$ be the domain of L. Then, for an eigenfunction $\phi \in \mathcal{D}$ with eigenvalue λ,

$$L\phi = -\lambda\phi$$

and it is not difficult to prove, using the Markov property, that

$$L \, E(\phi(X_\delta) \,|\, X_0 = x) = -\lambda \, E(\phi(X_\delta) \,|\, X_0 = x)$$

and hence that

$$E(\phi(X_\delta) \,|\, X_0 = x) = e^{-\lambda\delta} \phi(x).$$

This last result is a basis for showing that $\{Y_t = e^{-\lambda t} \, \phi(X_t), \, t > 0\}$ is a martingale and hence for the defining of families of estimating functions.

Kessler and Sørensen (1995) have used these ideas for the estimation of parameters in a discretely observed diffusion process with observations $(X_{t_0}, \ldots, X_{t_n})$ and $t_i - t_{i-1} = \Delta$ for each i. In this context the family of martingale estimating functions

$$\left\{ \sum_{i=1}^{n} \alpha(X_{t_{i-1}}; \theta) \left[\phi(X_{t_i}; \theta) - e^{-\lambda(\theta)\Delta} \, \phi(X_{t_{i-1}}; \theta) \right] \right\}$$

with α's arbitrary functions, is considered. Also, more generally, when k eigenfunctions $\phi_1(x;\theta),\ldots,\phi_k(x;\theta)$ with distinct eigenvalues $\lambda_1(\theta),\ldots,\lambda_k(\theta)$ are available, the family of estimating functions

$$\left\{\sum_{i=1}^{n}\sum_{j=1}^{k}\alpha_j(X_{t_{i-1}};\theta)\,\left[\phi_j(X_{t_i};\theta)-e^{-\lambda_j(\theta)\Delta}\,\phi_j(X_{t_{i-1}};\theta)\right]\right\},$$

with α_j's arbitrary functions, can be used. Kessler and Sørensen (1995) have discussed quasi-likelihood estimators from these families and their consistency and asymptotic normality.

Another approach to estimation which makes use of the infinitesimal generator has been suggested by Baddeley (1995). Here a parametric model given by a family of distributions $\{P_\theta,\,\theta\in\Theta\}$ is assumed and the aim is to estimate θ on the basis of x drawn from P_θ. The idea is to find a continuous time Markov process $\{X_t,\,t>0\}$ for which P_θ is an equilibrium distribution for each θ.

Now let L_θ be the infinitesimal generator of $\{Y_t\}$ and choose a family of statistics $\{S(x)\}$ belonging to the domain $\mathcal{D}$ of L_θ. Each of these produces an estimating function $(L_\theta S)(x)$ and it is easily checked that these have zero mean. Baddeley calls the corresponding estimating equations *time invariant.*

As an example we consider the discrete random fields $X=(X_i,\,i\in G)$ in which the set of "sites" G is an arbitrary finite set and the site "labels" X_i take values in an arbitrary finite set. Let

$$P(X=x)=\pi_\theta(x),\;x\in\chi$$

and assume that $\pi_\theta(x)>0$ for all $x\in\chi$, $\theta\in\Theta$. If $\{Y_t,\,t>0\}$ is the Gibbs sampler for π_θ (e.g., Guyon (1995, p. 211)), then the infinitesimal generator of $\{Y_t\}$ operating on the statistic S turns out to be

$$(L_\theta\,S)(x)=\sum_{i\in G}\{E_\theta\,[S(X)\,|\,X_{G\backslash i}=x_{G\backslash i}]-S(x)\},$$

where $X_B=(X_i,\,i\in B)$ denotes the restriction of X to $B\subset G$. The corresponding time invariant estimator is thus the solution of

$$\frac{1}{|G|}\sum_{i\in G}E_\theta\,[S(X)\,|\,X_{G\backslash i}=x_{G\backslash i}]=S(x)$$

where $|G|$ is the number of "sites" in G. For more details see Baddeley (1995). Questions of choice of family and of optimality are largely open.

13.2 Solution of Estimating Equations

Iterative numerical methods are very often necessary for obtaining estimators from estimating equations. Detailed discussion of computational procedures is outside the scope of this monograph. However, there are various points which can usefully be noted.

Preliminary transformation of the parameter so that the covariance matrix is approximately diagonal can be very helpful. Also, if the dimension of the problem can be reduced via expressing some of the components explicitly in terms of others, then this is usually advantageous.

For details on computational methods the reader is referred to Thisted (1988). In particular, Chapter 3, Section 10 of this reference deals with GLIM and generalized least squares methods and Chapter 4 with the Newton-Raphson method and Fisher's method of scoring. These are perhaps the most widely used methods. For going beyond the primary tool of GLIM for possibly nonlinear models see Gay and Welsh (1988) and Nelder and Pregibon (1987).

Let $\boldsymbol{G}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ be an estimating equation. Then, the Newton-Raphson and scoring iterative schemes can be defined by

$$\hat{\boldsymbol{\theta}}^{(r+1)} = \hat{\boldsymbol{\theta}}^{(r)} - (\dot{\boldsymbol{G}}(\hat{\boldsymbol{\theta}}^{(r)}))^{-1}\,\boldsymbol{G}(\hat{\boldsymbol{\theta}}^{(r)})$$

and

$$\hat{\boldsymbol{\theta}}^{(r+1)} = \hat{\boldsymbol{\theta}}^{(r)} - (E\dot{\boldsymbol{G}}(\hat{\boldsymbol{\theta}}^{(r)}))^{-1}\,\boldsymbol{G}(\hat{\boldsymbol{\theta}}^{(r)}),$$

respectively. The former uses the observed $-\dot{\boldsymbol{G}}(\boldsymbol{\theta})$ and the latter the expected $-E\dot{\boldsymbol{G}}(\boldsymbol{\theta})$ which, when $\boldsymbol{G}(\boldsymbol{\theta})$ is a standardized estimating function, equals the covariance matrix $E\boldsymbol{G}(\boldsymbol{\theta})\,\boldsymbol{G}'(\boldsymbol{\theta})$

These schemes often work well, with rapid convergence, especially if good starting values are chosen. A simple method of estimation, such as the method of moments, if available, can provide suitable starting values.

An alternative approach which has the attraction of not involving matrices of partial derivatives of $\boldsymbol{G}$ has been suggested by Mak (1993). The idea here is as follows. For any $\boldsymbol{\theta}$, $\boldsymbol{\phi} \in \Theta$ let

$$\boldsymbol{H}(\boldsymbol{\phi}, \boldsymbol{\theta}) = E_{\phi}\,\boldsymbol{G}(\boldsymbol{\theta}).$$

Then, if $\hat{\boldsymbol{\theta}}^{(0)}$ is a given starting value, the sequence $\{\hat{\boldsymbol{\theta}}^{(r)}, r = 0, 1, \dots,\}$ is defined via

$$\boldsymbol{G}(\hat{\boldsymbol{\theta}}^{(r)}) = \boldsymbol{H}(\hat{\boldsymbol{\theta}}^{(r+1)}, \hat{\boldsymbol{\theta}}^{(r)}).$$

The properties of this scheme, which is equivalent to the method of scoring when $\boldsymbol{H}$ is linear in $\boldsymbol{\phi}$, have been investigated in Mak (1993). In particular, it is shown, under the usual kinds of regularity conditions, that if $\hat{\boldsymbol{\theta}}$ is the required estimator, then $\|\hat{\boldsymbol{\theta}}^{(r)} - \hat{\boldsymbol{\theta}}\| = O_p(n^{-r/2})$, n being the sample size and O_p denoting order in probability.

13.3 Multiple Roots

13.3.1 Introduction

For maximum likelihood estimation the accepted practice is to use the likelihood itself to discriminate between multiple roots. However, in cases where the estimating equation $\boldsymbol{G}(\boldsymbol{\theta}) = \mathbf{0}$, say, has been derived from other principles, there has been uncertainty about how to choose the appropriate root (e.g., Stefanski and Carroll (1987, Section 2.3), McCullagh (1991)). This has provided motivation for the search for scalar potential functions from which the estimating equation can be derived and, in particular, the development of a conservative quasi-score (Li and McCullagh (1994)), which, albeit with some loss in efficiency in general, is always the derivative of a scalar function.

Recently Li (1993) and Hanfelt and Liang (1995) have developed approximate likelihood ratio methods for quasi-likelihood settings. In each of these papers it is shown that, under various regularity conditions the approximate likelihood ratio can be used to discriminate asymptotically, with probability one, to allow the choice of the correct root. To use these methods it is necessary that the estimating functions, $G^{(s)}(\theta)$, say, are standardized to have the likelihood score property

$$E_\theta \left\{ -\dot{\boldsymbol{G}}^{(s)}(\boldsymbol{\theta}) \right\} = E_\theta \left\{ \boldsymbol{G}^{(s)}(\boldsymbol{\theta})\, \boldsymbol{G}^{(s)'}(\boldsymbol{\theta}) \right\}.$$

However, unless $\boldsymbol{\theta}$ has small dimension, the process of standardization may be tedious if explicit analytical expressions are desired. For example, the method of moments, or the eliminiation of incidential parameters, typically lead to equations which are not standardized. Furthermore, a wide range of estimating functions are used in practice, and many of these have no physically convincing interpretation. For example, if θ is 1-dimensional and G has an even number of roots, then its integral must increase steadily at one or other end of the range of θ. In many examples the integral is unbounded above, implying that the appropriate root is at a local rather than a global maximum. For these reasons we propose three simple direct methods which do not involve the construction of a potential function. This is clearly advantageous, for example, when

(i) the estimating function is not optimal (for example being derived by the method of moments), or

(ii) many incidental parameters are to be eliminated (e.g. for functional relationships).

Also, conceptually unsatisfying issues such as the arbitrariness of paths for integrals defining approximate likelihood ratios in non-conservative cases are then avoided. The results of this section are from Heyde and Morton (1997).

The methods which we advocate for choosing the correct root of $\boldsymbol{G}(\boldsymbol{\theta}) = \mathbf{0}$ are:

(1) examining the asymptotics to see which root provides a consistent result;

(2) picking the root for which $\dot{\boldsymbol{G}}(\boldsymbol{\theta})$ behaves asymptotically as its expected value $E_\theta\{\dot{\boldsymbol{G}}(\boldsymbol{\theta})\}$; and

(3) using a least squares or goodness of fit criterion to select the best root.

The motivation for Method 2 comes from generalization of the usual technique for identifying the maximum of a quasi-likelihood where $-\dot{\boldsymbol{G}}(\boldsymbol{\theta})$ and $-E_\theta \dot{\boldsymbol{G}}(\boldsymbol{\theta})$ have interpretations as empirical and expected information respectively.

The principles which underly the advocated methods are simple and transparent and we give three examples to illustrate their use. This is followed by a discussion of the underlying theoretical considerations.

13.3.2 Examples

We begin with some simple examples, the first of which concerns the *estimation of the mean of a non-normal distribution using the first two moments.*

Let $X_1, \ldots, X_n$ be i.i.d. with mean θ $(-\infty < \theta < \infty)$ and known variance σ^2 and consider the estimating function

$$G(\theta) = n^{-1} \sum_{i=1}^{n} \{a(X_i - \theta) + (X_i - \theta)^2 - \sigma^2\},$$

where a is a constant. Certainly we can calculate, say,

$$\phi(\theta) = \int_0^\theta G(u)\, du,$$

but to regard this cubic in θ as an appropriate likelihood surrogate requires a leap of faith. Nevertheless, the multiple root issue is simply resolved by each of the approaches mentioned above.

Let

$$\Delta = \frac{1}{4}\, a^2 + \bar{X}^2 - n^{-1} \sum_{i=1}^{n} X_i^2 + \sigma^2,$$

$\bar{X}$ being the sample mean. Then, on the set $\{\Delta > 0\}$ the roots of G are

$$\hat{\theta}_1 = \frac{1}{2}\, a + \bar{X} - \Delta^{1/2}, \qquad \hat{\theta}_2 = \frac{1}{2}\, a + \bar{X} + \Delta^{1/2},$$

and the strong law of large numbers gives $\hat{\theta}_1 \xrightarrow{\text{a.s.}} \theta_0$, $\hat{\theta}_2 \xrightarrow{\text{a.s.}} \theta_0 + a$, θ_0 being the true value. Thus, the root $\hat{\theta}_1$ is the correct one. This is Method 1.

Using Method 2, we see that

$$-\dot{G}(\theta) = a + 2(\bar{X} - \theta),$$

and $E_\theta \dot{G}(\theta) = -a$ so that the ratio $\dot{G}(\theta)/E_\theta \dot{G}(\theta)$ tends to 1 when $\theta = \hat{\theta}_1$ and to -1 when $\theta = \hat{\theta}_2$, identifying $\hat{\theta}_1$ as the correct root.

For Method 3 we use the criterion

$$S(\theta) = n^{-1} \sum_{i=1}^{n} (X_i - \theta)^2,$$

and see that

$$S(\hat{\theta}_2) - S(\hat{\theta}_1) = 2a\Delta^{\frac{1}{2}} \sim a^2\, n$$

as $n \to \infty$, from which we again prefer $\hat{\theta}_1$ to $\hat{\theta}_2$.

The next example concerns *estimation of the angle in circular data.* This problem, which has been studied by McCullagh (1991), concerns the model in which $(X_{1i}, X_{2i}),\ i = 1, \dots, n$ are i.i.d. multivariate normal, with mean vector $(r\,\cos\theta, r\,\sin\theta)$ and covariance matrix I.

First we consider $r = 1$ and use the quasi-score (also the score) function

$$\begin{aligned} Q(\theta) &= -(\bar{X}_1 - \cos\theta)\,\sin\theta + (\bar{X}_2 - \sin\theta)\,\cos\theta \\ &= -\bar{X}_1\,\sin\theta + \bar{X}_2\,\cos\theta. \end{aligned}$$

Then $Q(\theta) = 0$ possesses two solutions $\hat{\theta}_1$ and $\hat{\theta}_2$ with $\hat{\theta}_2 = \hat{\theta}_1 + \pi,\ \ \hat{\theta}_1 = \tan^{-1}(\bar{X}_2/\bar{X}_1)$ being the root in $(-\pi/2, \pi/2)$.

Assume that $-\pi/2 < \theta_0 < \pi/2$ is the true value. Then the law of large numbers shows that $\hat{\theta}_1$ is the consistent root. This is the approach of Method 1.

For Method 2, we see that

$$\dot{Q}(\theta) = -\bar{X}_2\,\sin\theta - \bar{X}_1\,\cos\theta,$$

while $E_\theta \dot{Q}(\theta) = -1$ and $\dot{Q}(\hat{\theta}_1)/E_{\hat{\theta}_1}\,\dot{Q}(\hat{\theta}_1) \to 1$, $\dot{Q}(\hat{\theta}_2)/E_{\hat{\theta}_2}\,\dot{Q}(\hat{\theta}_2) \to -1$, so we choose $\hat{\theta}_1$.

Finally, for Method 3 we can take the sum of squares criterion

$$S(\theta) = \sum_{i=1}^{n} \left[(X_{1i} - \cos\theta)^2 + (X_{2i} - \sin\theta)^2 \right],$$

and

$$S(\hat{\theta}_2) - S(\hat{\theta}_1) = 4n(\bar{X}_1\,\cos\hat{\theta}_1 + \bar{X}_2\,\sin\hat{\theta}_2) \sim 4n$$

if $\hat{\theta}_1$ is in the same half circle as θ_0.

There is also an interesting sequel to this example. If r is also to be estimated, then the constraint $r > 0$ removes the ambiguity in θ and there is only one solution, which is $\hat{\theta}_1$. In this case, two parameters are simpler than one!

These have been toy examples and we now move to something more substantial. Here we follow up on the example on a *functional normal regression model* treated in Section 2.3 of Stefanski and Carroll (1987). The setting is one in which Y has a normal distribution with mean $\alpha + \boldsymbol{\beta}'\boldsymbol{u}$ and variance σ^2. It is supposed that $\boldsymbol{u}$ cannot be observed but that independent measurements $\boldsymbol{X}$ on it are available and $\text{var}\,(\boldsymbol{X} \mid \boldsymbol{u}) = \sigma^2\,\boldsymbol{\Omega}$ where $\boldsymbol{\Omega}$ is known.

For this setting, Stefanski and Carroll derive the estimating function (cf. their equation (2.15))

$$\boldsymbol{G}(\boldsymbol{\beta}) = -\boldsymbol{\beta}' S_{YX} \boldsymbol{\Omega} \boldsymbol{\beta} + (S_{YY}\boldsymbol{\Omega} - S_{XX})\boldsymbol{\beta} + S_{YX}$$

for $\boldsymbol{\beta}$ where

$$S_{YY} = n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \qquad S_{YX} = (S_{YX_1}, \ldots, S_{YX_p})',$$

$$S_{XX} = (S_{X_iY_j}),$$

with

$$S_{YX_i} = n^{-1}\sum_{k=1}^{n}(Y_k - \bar{Y})(X_{ik} - \bar{X}_i), \qquad S_{X_iX_j} = n^{-1}\sum_{k=1}^{n}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j).$$

In contrast to Stefanski and Carroll we shall not specialize to the case where $\boldsymbol{\beta}$ is scalar ($p = 1$).

For Method 1 we need to examine the asymptotics. We suppose that

$$S_{U,U} = n^{-1}\sum_{i=1}^{n}(\boldsymbol{u}_i - \bar{\boldsymbol{u}})(\boldsymbol{u}_i - \bar{\boldsymbol{u}})' \to \operatorname{var}\boldsymbol{u},$$

say, as $n \to \infty$ and then, using the law of large numbers in an obvious notation which covers the structural model case as well as the functional one,

$$S_{YX} \xrightarrow{\text{a.s.}} \operatorname{cov}(\boldsymbol{X}, Y), \quad S_{XX} \xrightarrow{\text{a.s.}} \operatorname{var}\boldsymbol{X}, \quad S_{YY} \xrightarrow{\text{a.s.}} \operatorname{var} Y.$$

Also

$$\begin{aligned}
\operatorname{var} Y &= E(\operatorname{var}(Y \mid \boldsymbol{u})) + \operatorname{var}(E(Y \mid \boldsymbol{u})) \\
&= \sigma^2 + \boldsymbol{\beta}'(\operatorname{var}\boldsymbol{u})\,\boldsymbol{\beta}, \\
\operatorname{var}\boldsymbol{X} &= E(\operatorname{var}(\boldsymbol{X} \mid \boldsymbol{u})) + \operatorname{var}(E(\boldsymbol{X} \mid \boldsymbol{u})) \\
&= \sigma^2\boldsymbol{\Omega} + \operatorname{var}\boldsymbol{u}, \\
\operatorname{cov}(\boldsymbol{X}, Y) &= E(\operatorname{cov}(\boldsymbol{X}, Y) \mid \boldsymbol{u})) + \operatorname{cov}(E(\boldsymbol{X} \mid \boldsymbol{u}), E(Y \mid \boldsymbol{u})) \\
&= \operatorname{cov}(\boldsymbol{u}, \alpha + \boldsymbol{\beta}'\boldsymbol{u}) = (\operatorname{var}\boldsymbol{u})\,\boldsymbol{\beta}.
\end{aligned}$$

Thus, asymptotically,

$$\begin{aligned}
\boldsymbol{G}(\hat{\boldsymbol{\beta}}) &\sim -\hat{\boldsymbol{\beta}}'(\operatorname{var}\boldsymbol{u})\boldsymbol{\beta}\boldsymbol{\Omega}\hat{\boldsymbol{\beta}} + (\boldsymbol{\beta}'(\operatorname{var}\boldsymbol{u})\boldsymbol{\beta}\boldsymbol{\Omega} - \operatorname{var}\boldsymbol{u})\hat{\boldsymbol{\beta}} \\
&\quad + (\operatorname{var}\boldsymbol{u})\boldsymbol{\beta} \\
&= (-\Omega\hat{\boldsymbol{\beta}}\boldsymbol{\beta}' - I)(\operatorname{var}\boldsymbol{u})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \qquad (13.1)
\end{aligned}$$

One root of (13.1) is $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ and any other requires that $(\boldsymbol{I} + \boldsymbol{\Omega}\hat{\boldsymbol{\beta}}\boldsymbol{\beta}')$ be singular and hence that $\boldsymbol{\beta}'\boldsymbol{\Omega}\hat{\boldsymbol{\beta}} = -1$ since, otherwise,

$$(\boldsymbol{I} + \boldsymbol{\Omega}\hat{\boldsymbol{\beta}}\boldsymbol{\beta}')^{-1} = \boldsymbol{I} - \frac{\boldsymbol{\Omega}\hat{\boldsymbol{\beta}}\boldsymbol{\beta}'}{1 + \boldsymbol{\beta}'\boldsymbol{\Omega}\hat{\boldsymbol{\beta}}}.$$

Any such other root satisfies

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\operatorname{var}\boldsymbol{u})^{-1}\,[(\boldsymbol{I} + \boldsymbol{\Omega}\hat{\boldsymbol{\beta}}\boldsymbol{\beta}')^{-}(\boldsymbol{I} + \boldsymbol{\Omega}\hat{\boldsymbol{\beta}}\boldsymbol{\beta}') - \boldsymbol{I}]\boldsymbol{Z},$$

where $\boldsymbol{Z}$ is an arbitrary vector and, noting that we can take

$$(\boldsymbol{I} + \boldsymbol{\Omega}\hat{\boldsymbol{\beta}}\boldsymbol{\beta}')^{-} = \boldsymbol{I} + \boldsymbol{\Omega}\hat{\boldsymbol{\beta}}\boldsymbol{\beta}',$$

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\operatorname{var}\boldsymbol{u})^{-1}\boldsymbol{\Omega}\hat{\boldsymbol{\beta}}\boldsymbol{\beta}'\boldsymbol{Z}. \tag{13.2}$$

Next, using $\boldsymbol{\beta}'\boldsymbol{\Omega}\hat{\boldsymbol{\beta}} = -1$ in (13.2), we have

$$\boldsymbol{\beta}'(\operatorname{var}\boldsymbol{u})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = -\boldsymbol{\beta}'\boldsymbol{Z}, \tag{13.3}$$

and, setting $\boldsymbol{\beta}'\boldsymbol{Z} = s$, (13.2) can be rewritten as

$$\hat{\boldsymbol{\beta}} = [\boldsymbol{I} - (\operatorname{var}\boldsymbol{u})^{-1}\boldsymbol{\Omega}s]^{-1}\boldsymbol{\beta}, \tag{13.4}$$

provided $s\det\{(\operatorname{var}\boldsymbol{u})^{-1}\boldsymbol{\Omega}\} \neq 1$ so that $\boldsymbol{I} - (\operatorname{var}\boldsymbol{u})^{-1}\boldsymbol{\Omega}s$ is nonsingular.

Note that, using (13.4) in (13.3),

$$\boldsymbol{\beta}'\boldsymbol{\Omega}(\boldsymbol{I} - (\operatorname{var}\boldsymbol{u})^{-1}\boldsymbol{\Omega}s)^{-1}\boldsymbol{\beta} = -1, \tag{13.5}$$

and s is obtained from this equation. The equations (13.4), (13.5) then specify the second root.

When solving the estimating equation $\boldsymbol{G}(\boldsymbol{\beta}) = \boldsymbol{0}$, it is clear that the correct root is close to $(S_{XX} - \hat{\sigma}^2\boldsymbol{\Omega})^{-1}S_{YX}$ which consistently estimates $\boldsymbol{\beta}$ when

$$\hat{\sigma}^2 = S_{YY} - S'_{YX}(S_{XX} - \hat{\sigma}^2\boldsymbol{\Omega})^{-1}S_{YX}.$$

If we have an incorrect root $\hat{\boldsymbol{\beta}}$, then

$$S'_{YX}(S_{XX} - \hat{\sigma}^2\boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}\hat{\boldsymbol{\beta}} \simeq -1.$$

Thus, Method 1 should pick the correct root given a suitably large sample.

To use Method 2 in this context we find

$$\begin{aligned}\dot{\boldsymbol{G}}(\boldsymbol{\beta}) &= -\boldsymbol{\Omega}\boldsymbol{\beta}S'_{YX} - \boldsymbol{\beta}'S_{YX}\boldsymbol{\Omega} + (S_{YY}\boldsymbol{\Omega} - S_{XX}) \\ &\sim -\boldsymbol{\Omega}\boldsymbol{\beta}\boldsymbol{\beta}'_0(\operatorname{var}\boldsymbol{u}) - \boldsymbol{\beta}'(\operatorname{var}\boldsymbol{u})\boldsymbol{\beta}_0\boldsymbol{\Omega} \\ &\quad +(\boldsymbol{\beta}'_0(\operatorname{var}\boldsymbol{u})\boldsymbol{\beta}_0\boldsymbol{\Omega} - \operatorname{var}\boldsymbol{u})\end{aligned}$$

almost surely, $\boldsymbol{\beta}_0$ denoting the true value, and

$$E_{\beta}\dot{\boldsymbol{G}}(\boldsymbol{\beta}) \sim -(\boldsymbol{\Omega}\boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{I})\operatorname{var}\boldsymbol{u}$$

almost surely, so that if $\hat{\boldsymbol{\beta}}_0$ is a root which is consistent for $\boldsymbol{\beta}_0$,

$$\dot{\boldsymbol{G}}(\hat{\boldsymbol{\beta}}_0)(E_{\hat{\beta}_0}\dot{\boldsymbol{G}}(\hat{\boldsymbol{\beta}}_0))^{-1} \xrightarrow{\mathrm{p}} \boldsymbol{I},$$

while if $\hat{\boldsymbol{\beta}}_1$ is a root which is consistent for $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_0$,

$$\begin{aligned}
\dot{\boldsymbol{G}}(\hat{\boldsymbol{\beta}}_1)(E_{\hat{\beta}_1}\dot{\boldsymbol{G}}(\hat{\boldsymbol{\beta}}_1))^{-1} &\xrightarrow{\text{p}} (\boldsymbol{\Omega}\boldsymbol{\beta}_1\boldsymbol{\beta}_0' + \boldsymbol{I})(\boldsymbol{\Omega}\boldsymbol{\beta}_1\boldsymbol{\beta}_1' + \boldsymbol{I}) \\
&\quad + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)'(\operatorname{var}\boldsymbol{u})\,\boldsymbol{\beta}_0\,\boldsymbol{\Omega}(\operatorname{var}\boldsymbol{u})^{-1}(\boldsymbol{\Omega}\boldsymbol{\beta}_1\boldsymbol{\beta}_1' + \boldsymbol{I})^{-1} \\
&= \boldsymbol{I} + \boldsymbol{\Omega}\boldsymbol{\beta}_1\boldsymbol{\beta}_0' + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)'(\operatorname{var}\boldsymbol{u})\,\boldsymbol{\beta}_0\,\boldsymbol{\Omega}(\operatorname{var}\boldsymbol{u})^{-1} \\
&\qquad (\boldsymbol{\Omega}\boldsymbol{\beta}_1\boldsymbol{\beta}_1' + \boldsymbol{I})^{-1} \\
&= \boldsymbol{I} - s(\boldsymbol{\Omega}\boldsymbol{\beta}_1\boldsymbol{\beta}_1' + \boldsymbol{I})\,\boldsymbol{\Omega}(\operatorname{var}\boldsymbol{u})(\boldsymbol{\Omega}\boldsymbol{\beta}_1\boldsymbol{\beta}_1' + \boldsymbol{I})^{-1} \\
&\neq \boldsymbol{I}
\end{aligned}$$

since $s \neq 0$. Here we have used, in the algebra, the intermediate results

$$(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)'(\operatorname{var}\boldsymbol{u})\,\boldsymbol{\beta}_0 = -s$$

and

$$\boldsymbol{\Omega}\boldsymbol{\beta}_1\boldsymbol{\beta}_0'(\boldsymbol{\Omega}\boldsymbol{\beta}_1\boldsymbol{\beta}_1' + \boldsymbol{I})\operatorname{var}\boldsymbol{u} = -s\boldsymbol{\Omega}\boldsymbol{\beta}_1\boldsymbol{\beta}_1'\boldsymbol{\Omega}.$$

The ratio thus detects the correct root.

13.3.3 Theory

Informal general explanations of the properties observed in the examples treated in Section 13.3.2 are straightforward to provide.

The asymptotics of an estimating function $\boldsymbol{G}$ may usually be examined directly. Suppose, for example, that the data are $\{\boldsymbol{X}_t,\, t \in T\}$, T being some index set, and that $\boldsymbol{G}$ is of the form

$$\boldsymbol{G}_T = \boldsymbol{G}(\boldsymbol{A}_T(\{\boldsymbol{X}_t,\, t \in T\});\, \boldsymbol{\theta})$$

with $\boldsymbol{A}_T(\{\boldsymbol{X}_t,\, t \in T\}) \xrightarrow{\text{a.s.}} \boldsymbol{A}(\boldsymbol{\theta}_0)$ in the limit, $\boldsymbol{\theta}_0$ denoting the true value of $\boldsymbol{\theta}$. Then, the equation $\boldsymbol{G}(\boldsymbol{A}(\boldsymbol{\theta}_0);\, \boldsymbol{\theta}) = \boldsymbol{0}$ gives a clear indication of the large sample behavior of the roots of $\boldsymbol{G}(\boldsymbol{\theta}) = \boldsymbol{0}$. Results of the kind $\boldsymbol{A}_T(\{\boldsymbol{X}_t,\, t \in T\}) \xrightarrow{\text{a.s.}} \boldsymbol{A}(\boldsymbol{\theta}_0)$ are typically established using law of large number or ergodic theorem arguments. This is the basis of Method 1.

To use Method 2, we assume that the matrix derivative $\dot{\boldsymbol{G}}(\boldsymbol{\theta}) \sim \boldsymbol{H}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ in probability, where

$$\boldsymbol{H}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = E_{\theta_0}(\dot{\boldsymbol{G}}(\boldsymbol{\theta})).$$

Both quantities should be unbounded as sampling continues and their difference has zero expectation. Then, if $\hat{\boldsymbol{\theta}}_0$ is a consistent estimator of $\boldsymbol{\theta}_0$,

$$(\boldsymbol{H}(\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_0))^{-1}\dot{\boldsymbol{G}}(\hat{\boldsymbol{\theta}}_0) \sim \boldsymbol{I}$$

in probability. But, if $\hat{\boldsymbol{\theta}}_1$ is a consistent estimator of $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0$, then

$$(\boldsymbol{H}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_1))^{-1}\dot{\boldsymbol{G}}(\hat{\boldsymbol{\theta}}_1) \sim (\boldsymbol{H}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1))^{-1}\boldsymbol{H}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1),$$

which in most problems is different from $\boldsymbol{I}_s$ the main exception being where $\boldsymbol{H}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ is constant in $\boldsymbol{\theta}_0$. Apart from any such exception, $(\boldsymbol{H}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}))^{-1}\dot{\boldsymbol{G}}(\hat{\boldsymbol{\theta}})$ can be used to identify the correct root. The matrix may not need to be examined in detail in order to discard incorrect roots. We could, for example, inspect the signs of the diagonal elements of $\dot{\boldsymbol{G}}(\hat{\boldsymbol{\theta}})$ and $\boldsymbol{H}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})$. If we have a quasi-score, $\boldsymbol{Q}(\boldsymbol{\theta})$, the asymptotic positive definiteness of $-\dot{\boldsymbol{Q}}(\hat{\boldsymbol{\theta}})$ for $\hat{\boldsymbol{\theta}}$ a consistent estimator of $\boldsymbol{\theta}_0$ may be a useful diagnostic. We note that for likelihood scores, $-\dot{Q}(\hat{\boldsymbol{\theta}}) > \mathbf{0}$ identifies a local maximum of the likelihood. Thus, Method 2 could be regarded as a generalization of this idea to non-standardized and non-optimal estimating functions. The positive definiteness of $-\dot{\boldsymbol{Q}}$ is particularly useful when $\dot{\boldsymbol{Q}}$ depends on further variance parameters which are estimated through its magnitude.

Finally, for the goodness of fit approach of Method 3 we assume that we have some acceptable criterion for the goodness of fit of the data. A modest requirement might be that the expectation of the criterion is a minimum when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

While we do not use the minimization of the criterion for estimating $\boldsymbol{\theta}_0$, it can be used to determine preference between the roots of $\boldsymbol{G}$. Least squares, or some generalization of it, seems appropriate for this purpose.

An informal justification can easily be given. We shall consider the situation in which there are consistent roots and possibly others. Let $\boldsymbol{\theta}_0$ denote the true value, for which there is a consistent estimator $\hat{\boldsymbol{\theta}}_0$ and suppose that $\hat{\boldsymbol{\theta}}_1 \xrightarrow{\text{p}} \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0$. Then, for a least squares type setting with minimization criterion

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{f}(\boldsymbol{\theta}))'(\boldsymbol{X}_i - \boldsymbol{f}(\boldsymbol{\theta})),$$

say, where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is the data and $E\boldsymbol{X}_i = \boldsymbol{f}(\boldsymbol{\theta})$, we have

$$\begin{aligned} n^{-1}[S(\hat{\boldsymbol{\theta}}_1) - S(\hat{\boldsymbol{\theta}}_0)] &= n^{-1}(\boldsymbol{f}(\hat{\boldsymbol{\theta}}_0) - \boldsymbol{f}(\hat{\boldsymbol{\theta}}_1))' \sum_{i=1}^{n} (2\boldsymbol{X}_i - \boldsymbol{f}(\hat{\boldsymbol{\theta}}_0) - \boldsymbol{f}(\hat{\boldsymbol{\theta}}_1)) \\ &\xrightarrow{\text{p}} (\boldsymbol{f}(\boldsymbol{\theta}_0) - \boldsymbol{f}(\boldsymbol{\theta}_1))'(\boldsymbol{f}(\boldsymbol{\theta}_0) - \boldsymbol{f}(\boldsymbol{\theta}_1)), \end{aligned}$$

using law of large number considerations.

If, on the other hand, $\hat{\boldsymbol{\theta}}_2$ is not a consistent root, then

$$\begin{aligned} n^{-1}[S(\hat{\boldsymbol{\theta}}_2) - S(\hat{\boldsymbol{\theta}}_0)] &= n^{-1}(\boldsymbol{f}(\hat{\boldsymbol{\theta}}_0) - \boldsymbol{f}(\hat{\boldsymbol{\theta}}_2))' \sum_{i=1}^{n} (2\boldsymbol{X}_i - \boldsymbol{f}(\hat{\boldsymbol{\theta}}_0) - \boldsymbol{f}(\hat{\boldsymbol{\theta}}_2)) \\ &\sim (\boldsymbol{f}(\boldsymbol{\theta}_0) - \boldsymbol{f}(\hat{\boldsymbol{\theta}_2}))'(\boldsymbol{f}(\boldsymbol{\theta}_0) - \boldsymbol{f}(\hat{\boldsymbol{\theta}}_2)) \end{aligned}$$

in probability, which is $O_p(1)$ and the discrimination between roots should still be clear. Note also that $n^{-1}S(\hat{\boldsymbol{\theta}}_0)$ will typically converge in probability to a constant.

13.4 Resampling Methods

Resampling methods, especially the bootstrap and jackknife, have proved to be highly useful in practice for obtaining confidence intervals for unknown $\boldsymbol{\theta}$, particularly when the sample size is limited. The methods were originally developed for i.i.d. random variables but considerable effort has been expended recently in providing adaptions to various dependent variable contexts. Nevertheless, the ideas behind resampling always require a focus on variables within the model of interest which do not depart drastically from stationarity.

A detailed discussion on this subject is outside the scope of this monograph, in part because only a small proportion of the current literature involves estimating functions, but the prospects for future work are clear. Both jackknife and bootstrap methods of providing confidence intervals for estimating functions which include martingales are discussed by Lele (1991b). The jackknife results come from Lele (1991a) and the bootstrap results, which are rather less complete, adapt work of Wu (1986) and Liu (1988).

One may expect that the technology associated with the moving block bootstrap to be able to deal with estimating functions based on stationary variables (e.g. Künsch (1989) for weakly dependent variables; Lahiri (1993), (1995) for strongly dependent ones). The idea behind the moving block bootstrap is to resample blocks of observations, rather than single observations as with the ordinary bootstrap, thereby adequately preserving the dependence of the data within blocks.

References

Aalen, O. O. (1977). Weak convergence of stochastic integrals related to counting processes. *Z. Wahrsch. Verw. Geb.* **38**, 261–277.

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6**, 701–726.

Aase, K. K. (1983). Recursive estimation in nonlinear time series of autoregressive type. *J. Roy. Statist. Soc. Ser. B* **45**, 228–237.

Adenstadt, R. K. (1974). On large-sample estimation for the mean of a stationary random sequence. *Ann. Statist.* **2**, 1095–1107.

Adenstadt, R. K., and Eisenberg, B. (1974). Linear estimation of regression coefficients. *Quart. Appl. Math.* **32**, 317–327.

Aitchison, J., and Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraint. *Ann. Math. Statist.* **29**, 813–828.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1982). Linear nonparametric tests for comparison of counting processes, with applications to censored survival data. *Internat. Statist. Rev.* **50**, 219–258.

Anh, V. V. (1988). Nonlinear least squares and maximum likelihood estimation for a heteroscedastic regression model. *Stochastic Process. Appl.* **29**, 317–333.

Baddeley, A. J. (1995). Time-invariance estimating equations. Research Report. Dept. Mathematics, Univ. Western Australia.

Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases.* Griffin, London.

Barndorff-Nielsen, O. E., and Cox, D. R. (1994). *Inference and Asymptotics.* Chapman and Hall, London.

Barndorff-Nielsen, O. E., and Sørensen, M. (1994). A review of some aspects of asymptotic likelihood theory for stochastic processes. *Internat. Statist. Rev.* **62**, 133–165.

Basawa, I. V. (1985). Neyman-Le Cam tests based on estimating functions. In L. Le Cam and R. A. Olshen Eds., *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer 2*, Wadsworth, Belmont, CA, 811–825.

Basawa, I. V. (1991). Generalized score tests for composite hypotheses. In V. P. Godambe, Ed. *Estimating Functions,* Oxford Science Publications, Oxford, 121–131.

Basawa, I. V., and Brockwell, P. J. (1984). Asymptotic conditional inference for regular non-ergodic models with an application to autoregressive processes. *Ann. Statist.* **12**, 161–171.

Basawa, I. V., Huggins, R. M., and Staudte, R. G. (1985). Robust tests for time series with an application to first-order autoregressive processes. *Biometrika* **72**, 559–571.

Basawa, I. V., and Koul, H. L. (1988). Large-sample statistics based on quadratic dispersion. *Internat. Statist. Rev.* **56**, 199–219.

Basawa, I. V., and Prakasa Rao, B. L. S. (1980). *Statistical Inference for Stochastic Processes.* Academic Press, London.

Basawa, I. V., and Scott, D. J. (1983). *Asymptotic Optimal Inference for Non-Ergodic Models.* Lecture Notes in Statistics **17**, Springer, New York.

Becker, N. G., and Heyde, C. C. (1990). Estimating population size from multiple recapture-experiments. *Stochastic Process. Appl.* **36**, 77–83.

Beran, J. (1989). A test of location for data with slowly decaying serial correlations. *Biometrika* **76**, 261–269.

Berliner, L. M. (1991). Likelihood and Bayesian prediction for chaotic systems. *J. Amer. Statist. Assoc.* **86**, 938–952.

Besag, J. E. (1975). Statistical analysis of non-lattice data. *The Statistician* **24**, 179–195.

Bhapkar, V. P. (1972). On a measure of efficiency in an estimating equation. *Sankhyā Ser. A* **34**, 467–472.

Bhapkar, V. P. (1989). On optimality of marginal estimating equations. Technical Report No. 274, Dept. Statistics, Univ. Kentucky.

Bibby, B. M., and Sørensen, M. (1995). Martingale estimation functions for discretely observed diffusion processes. *Bernoulli* **1**, 17–39.

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

Bradley, E. L. (1973). The equivalence of maximum likelihood and weighted least squares estimates in the exponential family. *J. Amer. Statist. Assoc.* **68**, 199–200.

Bustos, O. H. (1982). General M-estimates for contaminated p-th order autoregressive processes: consistency and asymptotic normality. Robustness in autoregressive processes. *Z. Wahrsch. Verw. Geb.* **59**, 491–504.

Carroll, R. J., and Stefanski, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *J. Amer. Statist. Assoc.* **85**, 652–663.

Chan, N. H., and Wei, C. Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Ann. Statist.* **16**, 367–401.

Chandrasekar, B., and Kale, B. K. (1984). Unbiased statistical estimation for parameters in presence of nuisance parameters. *J. Statistical Planning Inf.* **9**, 45–54.

Chen, K., and Heyde, C. C. (1995). On asymptotic optimality of estimating functions. *J. Statistical Planning Inf.* **48**, 102–112.

Chen, Y. (1991). On Quasi-likelihood Estimations. PhD thesis, University of Wisconsin-Madison.

Cheng, R. C. H., and Traylor, L. (1995). Non-regular maximum likelihood problems (with discussion). *J. Roy. Statist. Soc. Ser. B* **57**, 3–44.

Choi, Y. J., and Severo, N. C. (1988). An approximation for the maximum likelihood estimator for the infection rate in a simple stochastic epidemic. *Biometrika* **75**, 392–394.

Chung, K. L., and Williams, R. J. (1983). *Introduction to Stochastic Integration.* Birkhäuser, Boston.

Comets, F., and Janžura, M. (1996). A central limit theorem for conditionally centered random fields with an application to Markov fields (Preprint).

Cox, D. R., and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

Cox, J. S., Ingersoll, J. E., and Ross, S. A. (1985). A theory of term structure of interest rates. *Econometrica* **53**, 363–384.

Cramér, H. (1946). *Methods of Mathematical Statistics.* Princeton University Press, Princeton.

Crowder, M. (1987). On linear and quadratic estimating functions. *Biometrika* **74**, 591–597.

Cutland, N. J., Kopp, P. E., and Willinger, W. (1995). Stock price returns and the Joseph effect: a fractional version of the Black-Scholes model. *Seminar on Stochastic Analysis, Random Fields and Applications (Ascona 1993)*, Progr. Probab. **36**, Birkhäuser, Basel, 327–351.

Dahlhaus, R. (1988). Efficient parameter estimation for self-similar processes. *Ann. Statist.* **17**, 1749–1766.

Daley, D. J. (1969). Integral representations of transition probabilities and serial covariances of certain Markov chains. *J. Appl. Prob.* **6**, 648–659.

Davidian, M., and Carroll, R. J. (1987). Variance function estimation. *J. Amer. Statist. Assoc.* **82**, 1079–1091.

Davis, R. A., and McCormick, W. P. (1989). Estimation for first order autoregressive processes with positive or bounded innovations. *Stochastic Process. Appl.* **31**, 237–250.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the E-M algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1–38.

Denby, L., and Martin, R. L. (1979). Robust estimation of the first order autoregressive parameter. *J. Amer. Statist. Assoc.* **74**, 140–146.

Desmond, A. F. (1991). Quasi-likelihood, stochastic processes and optimal estimating functions. In V. P. Godambe, Ed., *Estimating Functions,* Oxford Science Publications, Oxford, 133–146.

Desmond, A. F. (1996). Optimal estimating functions quasi-likelihood and statistical modelling. *J. Statistical Planning Inf.*, in press.

Dion, J.-P., and Ferland, R. (1995). Absolute continuity, singular measures and asymptotics for estimators. *J. Statistical Planning Inf.* **43**, 235–242.

Doukhan, P. (1994). *Mixing.* Lecture Notes in Statistics **85**, Springer, New York.

Duffie, D. (1992). *Dynamic Asset Pricing Theory.* Princeton Univ. Press, Princeton.

Durbin, J. (1960). Estimation of parameters in time series regression models. *J. Roy. Statist. Soc. Ser. B* **22**, 139–153.

Efron, B., and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus Fisher information. *Biometrika* **65**, 457–487.

Elliott, R. J. (1982). *Stochastic Calculus and Applications.* Springer, New York.

Elliott, R. J., Aggoun, L., and Moore, J. B. (1994). *Hidden Markov Model Estimation and Control.* Springer, New York.

Feigin, P. D. (1977). A note on maximum likelihood estimation for simple branching processes. *Austral. J. Statist.* **19**, 152–154.

Feigin, P. D. (1985). Stable convergence of semimartingales. *Stochastic Process. Appl.* **19**, 125–134.

Feigin, P. D., Tweedie, R. L., and Belyea, C. (1983). Weighted area techniques for explicit parameter estimation in multi-stage models. *Austral. J. Statist.* **25**, 1–16.

Firth, D. (1987). On efficiency of quasi-likelihood estimation. *Biometrika* **74**, 233–246.

Firth, D. (1993). Bias reduction for maximum likelihood estimates. *Biometrika* **80**, 27–38.

Firth, D., and Harris, I. R. (1991). Quasi-likelihood for multiplicative random effects. *Biometrika* **78**, 545–555.

Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses. *Statistical Science* **8**, 284–309.

Fox, R., and Taqqu, M. S. (1986). Large sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Ann. Statist.* **14**, 517–532.

Fox, R., and Taqqu, M. S. (1987). Central limit theorems for quadratic forms in random variables having long-range dependence. *Probab. Th. Rel. Fields* **74**, 213–240.

Fürth, R. (1918). Statistik und Wahrscheinlichkeitsnachwirkung. *Physik Z.* **19**, 421–426.

Gastwirth, J. I., and Rubin, H. (1975). The behavior of robust estimators on dependent data. *Ann. Statist.* **3**, 1070–1100.

Gauss, C. F. (1880). *Theoria Combationis Observatorium Erroribus Minimus Obnaie* Part 1, 1821, Part 2, 1823; Suppl. 1826. In Werke 4, Göttingen, 1–108.

Gay, D. M., and Welsh, R. E. (1988). Maximum-likelihood and quasi-likelihood for nonlinear exponential family regression models. *J. Amer. Statist. Assoc.* **83**, 990–998.

Glynn, P., and Iglehart, D. L. (1990). Simultaneous output analysis using standardized time series. *Math. Oper. Res.* **15**, 1–16.

Godambe, V. P. (1960). An optimum property of regular maximum-likelihood estimation. *Ann. Math. Statist.* **31**, 1208–1211.

Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63**, 277–284.

Godambe, V. P. (1985). The foundations of finite sample estimation in stochastic processes. *Biometrika* **72**, 419–428.

Godambe, V. P. (Ed.) (1991) *Estimating Functions.* Oxford Science Publications, Oxford.

Godambe, V. P. (1994). Linear Bayes and optimal estimation. Technical Report STAT-94-11, Dept. Statistical & Actuarial Sciences, University of Waterloo.

Godambe, V. P., and Heyde C. C. (1987). Quasi-likelihood and optimal estimation. *Internat. Statist. Rev.* **55**, 231–244.

Godambe, V. P., and Kale, B. K. (1991). Estimating functions: an overview. In V. P. Godambe, Ed., *Estimating Functions,* Oxford Science Publications, Oxford, 3–20.

Godambe, V. P., and Thompson, M. E. (1974). Estimating equations in the presence of a nuisance parameter. *Ann. Statist.* **2**, 568–571.

Godambe, V. P., and Thompson, M. E. (1986). Parameters of superpopulation and survey population: their relationship and estimation. *Internat. Statist. Rev.* **54**, 127–138.

Godambe, V. P., and Thompson, M. E. (1989). An extension of quasi-likelihood estimation. *J. Statistical Planning Inf.* **22**, 137–152.

Greenwood, P. E., and Wefelmeyer, W. (1991). On optimal estimating functions for partially specified counting process models. In V. P. Godambe, Ed., *Estimating Functions,* Oxford Science Publications, Oxford, 147–168.

Grenander, U. (1981). *Abstract Inference.* Wiley, New York.

Grenander, U., and Szegö, G. (1958). *Toeplitz Forms and their Applications.* Univ. Calif. Press, Berkeley and Los Angeles.

Guyon, X. (1982). Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika* **69**, 95–105.

Guyon, X. (1995). *Random Fields on a Network, Modeling, Statistics and Applications*. Springer, New York.

Halfin, S. (1982). Linear estimators for a class of stationary queueing processes. *Oper. Res.* **30**, 515–529.

Hall, P., and Heyde, C. C. (1980). *Martingale Limit Theory and its Application.* Academic Press, New York.

Hanfelt, J. J., and Liang, K.-Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika* **82**, 461–477.

Hannan, E. J. (1970). *Multiple Time Series.* Wiley, New York.

Hannan, E. J. (1973). The asymptotic theory of linear time series models. *J. Appl. Prob.* **10**, 130–145.

Hannan, E. J. (1974). Correction to Hannan (1973). *J. Appl. Prob.* **11**, 913.

Hannan, E. J. (1976). The asymptotic distribution of serial covariances. *Ann. Statist.* **4**, 396–399.

Harville, D. A. (1977). Maximum likelihood approaches to variance components estimation and related problems. *J. Amer. Statist. Assoc.* **72**, 320–338.

Heyde, C. C. (1986). Optimality in estimation for stochastic processes under both fixed and large sample conditions. In Yu. V. Prohorov, V. A. Statulevicius, V. V. Sazonov and B. Grigelionis, Eds., *Probability Theory and Mathematical Statistics. Proceedings of the Fourth Vilnius Conference*, Vol. **1**, VNU Sciences Press, Utrecht, 535–541.

Heyde, C. C. (1987). On combining quasi-likelihood estimating functions. *Stochastic Process. Appl.* **25**, 281–287.

Heyde, C. C. (1988a). Fixed sample and asymptotic optimality for classes of estimating functions. *Contemp. Math.* **80**, 241–247.

Heyde, C. C. (1988b). Asymptotic efficiency results for the method for moments with application to estimation for queueing processes. In O. J. Boxma and R. Syski, Eds., *Queueing Theory and its Application. Liber Amicorum for J. W. Cohen.* CWI Monograph No. **7**, North-Holland, Amsterdam, 405–412.

Heyde, C. C. (1989a). On efficiency for quasi-likelihood and composite quasi-likelihood methods. In Y. Dodge, Ed., *Statistical Data Analysis and Inference,* Elsevier, Amsterdam, 209–213.

Heyde, C. C. (1989b). Quasi-likelihood and optimality for estimating functions: some current unifying themes. *Bull. Internat. Statist. Inst.* **53**, Book 1, 19–29.

Heyde, C. C. (1992a). On best asymptotic confidence intervals for parameters of stochastic processes. *Ann. Statist.* **20**, 603–607.

Heyde, C. C. (1992b). Some results on inference for stationary processes and queueing systems. In U. N. Bhat and I. V. Basawa, Eds., *Queueing and Related Models*. Oxford Univ. Press, Oxford, 337–345.

Heyde, C. C. (1993). Quasi-likelihood and general theory of inference for stochastic processes. In A. Obretenov and V. T. Stefanov, Eds., *7th International Summer School on Probability Theory and Mathematical Statistics, Lecture Notes*, Science Culture Technology Publishing, Singapore, 122–152.

Heyde, C. C. (1994a). A quasi-likelihood approach to estimating parameters in diffusion type processes. In J. Galambos and J. Gani, Eds., *Studies in Applied Probability*, *J. Applied Prob.* **31A**, 283–290.

Heyde, C. C. (1994b). A quasi-likelihood approach to the REML estimating equations. *Statistics & Probability Letters* **21**, 381–384.

Heyde, C. C. (1996). On the use of quasi-likelihood for estimation in hidden-Markov random fields. *J. Statistical Planning Inf.* **50**, 373–378.

Heyde, C. C., and Gay, R. (1989). On asymptotic quasi-likelihood. *Stochastic Process. Appl.* **31**, 223–236.

Heyde, C. C., and Gay, R. (1992). Thoughts on modelling and identification of random processes and fields subject to possible long-range dependence. In L. H. Y. Chen, K. P. Choi, K. Hu and J.-H. Lou, Eds., *Probability Theory*, de Gruyter, Berlin, 75–81.

Heyde, C. C., and Gay, R. (1993). Smoothed periodogram asymptotics and estimation for processes and fields with possible long-range dependence. *Stochastic Process. Appl.* **45**, 169–182.

Heyde, C. C., and Lin, Y.-X. (1991). Approximate confidence zones in an estimating function context. In V. P. Godambe, Ed., *Estimating Functions*, Oxford Science Publications, Oxford, 161–168.

Heyde, C. C., and Lin, Y.-X. (1992). On quasi-likelihood methods and estimation for branching processes and heteroscedastic regression models. *Austral. J. Statist.* **34**, 199–206.

Heyde, C. C., and Morton, R. (1993). On constrained quasi-likelihood estimation. *Biometrika* **80**, 755–761.

Heyde, C. C., and Morton, R. (1995). Quasi-likelihood and generalizing the E-M algorithm. *J. Roy. Statist. Soc. Ser. B* **58**, 317-327.

Heyde, C. C., and Morton, R. (1997). Multiple roots and dimension reduction issues for general estimating equations (Preprint).

Heyde, C. C., and Seneta, E. (1972). Estimation theory for growth and immigration rates in a multiplicative process. *J. Appl. Prob.* **9**, 235–256.

Heyde, C. C., and Seneta, E. (1977). *I. J. Bienaymé: Statistical Theory Anticipated.* Springer, New York.

Hoffmann-Jørgensen, J. (1994). *Probability With a View Towards Statistics.* Vol. II. Chapman and Hall, New York.

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika* **28**, 321–377.

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

Hutton, J. E., and Nelson, P. I. (1984). A mixing and stable central limit theorem for continuous time martingales. *Technical Report No.* **42**, Kansas State University.

Hutton, J. E., and Nelson, P. I. (1986). Quasi-likelihood estimation for semimartingales. *Stochastic Process. Appl.* **22**, 245–257.

Hutton, J. E., Ogunyemi, O. T., and Nelson, P. I. (1991). Simplified and two-stage-quasi-likelihood estimators. In V. P. Godambe, Ed., *Estimating Functions,* Oxford Science Publications, Oxford, 169–187.

Ibragimov, I. A., and Linnik, Yu. V. (1971). *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff, Groningen.

Jiang, J. (1996). REML estimation: asymptotic behaviour and related topics. *Ann. Statist.* **24**, 255–286.

Judge, G. G., and Takayama, T. (1966). Inequality restrictions in regression analysis. *J. Amer. Statist. Assoc.* **61**, 166–181.

Kabaila, P. V. (1980). An optimal property of the least squares estimate of the parameter of the spectrum of a purely non-deterministic time series. *Ann. Statist.* **8**, 1082–1092.

Kabaila, P. V. (1983). On estimating time series parameters using sample autocorrelations. *J. Roy. Statist. Soc. Ser. B* **45**, 107–119.

Kallianpur, G. (1983). On the diffusion approximation to a discountinuous model for a single neuron. In P. K. Sen, Ed., *Contributions to Statistics: Essays in Honor of Norman L. Johnson.* North-Holland, Amsterdam, 247–258.

Kallianpur, G., and Selukar, R. S. (1993). Estimation of Hilbert space valued random variables by the method of sieves. In J. K. Ghost et al., Eds., *Statistics and Probability, A Raghu Rag Bahadur Festschrift*, Wiley Eastern, New Delhi, 325–347.

Karlin, S., and Taylor, H. M. (1981). *A Second Course in Stochastic Processes.* Academic Press, New York.

Karr, A. F. (1987). Maximum likelihood estimation in the multiplicative intensity model via sieves. *Ann. Statist.* **15**, 473–490.

Karson, M. J. (1982). *Multivariate Statistical Methods.* Iowa State Univ. Press, Ames, IA.

Kaufmann, H. (1987). On the strong law of large numbers for multivariate martingales. *Stochastic Process. Appl.* **26**, 73–85.

Kaufmann, H. (1989). On weak and strong consistency of the maximum likelihood estimator in stochastic processes (Preprint).

Kessler, M., and Sørensen, M. (1995). Estimating equations based on eigenfunctions for a discretely observed diffusion process. Research Report No. 332, Dept. Theoretical Statistics, Univ. Aarhus.

Kimball, B. F. (1946). Sufficient statistical estimation functions for the parameters of the distribution of maximum values. *Ann. Math. Statist.* **17**, 299–309.

Kloeden, P. E., and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations.* Springer, Berlin.

Kloeden, P. E., Platen, E., Schurz, H., and Sørensen, M. (1996). On effects of discretization on estimators of drift parameters for diffusion processes. *J. Appl. Prob.* **33**, 1061-1076.

Künsch, H. (1984). Infinitesimal robustness for autoregressive processes. *Ann. Statist.* **12**, 843–863.

Künsch, H. R. (1989). The jackknife and bootstrap for general stationary observations. *Ann. Statist.* **17**, 1217–1241.

Kulkarni, P. M., and Heyde, C. C. (1987). Optimal robust estimation for discrete time stochastic processes. *Stochastic Process. Appl.* **26**, 267–276.

Kulperger, R. (1985). On an optimal property of Whittle's Gaussian estimate of the parameter of the spectrum of a time series. *J. Time Ser. Anal.* **6**, 253–259.

Kutoyants, Yu., and Vostrikova, L. (1995). On non-consistency of estimators. *Stochastics and Stochastic Reports* **53**, 53–80.

Lahiri, S. N. (1993). On the moving block bootstrap under long range dependence. *Statistics & Probability Letters* **18**, 405–413.

Lahiri, S. N. (1995). On the asymptotic behaviour of the moving block bootstrap for normalized sums of heavy-tail random variables. *Ann. Statist.* **23**, 1331–1349.

Lai, T. L., and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification of dynamic systems. *Ann. Statist.* **10**, 154–166.

Laird, N. (1985). Missing information principle. In N. L. Johnson and S. Kotz, Eds., *Encyclopedia of Statistical Sciences*, Wiley, New York, **5**, 548–552.

Le Cam, L. (1990a). On the standard asymptotic confidence ellipsoids of Wald. *Internat. Statist. Rev.* **58**, 129–152.

Le Cam, L. (1990b). Maximum likelihood: an introduction. *Internat. Statist. Rev.* **58**, 153–171.

Lele, S. (1991a). Jackknifing linear estimating equations: asymptotic theory and applications in stochastic processes. *J. Roy. Statist. Soc. Ser. B* **53**, 253–267.

Lele, S. (1991b). Resampling using estimating equations. In V. P. Godambe, Ed., *Estimating Functions*, Oxford Science Publications, Oxford, 295–304.

Lele, S. (1994). Estimating functions in chaotic systems. *J. Amer. Statist. Assoc.* **89**, 512–516.

Lépingle, D. (1977). Sur le comportement asymptotique des martingales locales. *Springer Lecture Notes in Mathematics* **649**, 148–161.

Leskow, J. (1989). A note on kernel smoothing of an estimator of a periodic function in the multiplicative model. *Statistics & Probability Letters* **7**, 395–400.

Leskow, J., and Rozanski, R. (1989). Histogram maximum likelihood estimation in the multiplicative intensity model. *Stochastic Process. Appl.* **31**, 151–159.

Li, B. (1993). A deviance function for the quasi-likelihood method. *Biometrika* **80**, 741–753.

Li, B. (1996a). A minimax approach to the consistency and efficiency for estimating equations. *Ann. Statist.* **24**, 1283–1297.

Li, B. (1996b). An optimal estimating equation based on the first three cumulants (Preprint).

Li, B., and McCullagh, P. (1994). Potential functions and conservative estimating equations. *Ann. Statist.* **22**, 340–356.

Liang, K. Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lin, Y.-X. (1992). The quasi-likelihood method. PhD thesis, Australian National University.

Lin, Y.-X. (1994a). The relationship between quasi-score estimating functions and E-sufficient estimating functions. *Austral. J. Statist.* **36**, 303–311.

Lin, Y.-X. (1994b). On the strong law of large numbers of multivariate martingales with random norming. *Stochastic Process. Appl.* **54**, 355–360.

Lin, Y.-X. (1996). Quasi-likelihood estimation of variance components of heteroscedastic random ANOVA model (Preprint).

Lin, Y.-X., and Heyde, C. C. (1993). Optimal estimating functions and Wedderburn's quasi-likelihood. *Comm. Statist. Theory Meth.* **22**, 2341–2350.

Lindsay, B. (1982). Conditional score functions: some optimality results. *Biometrika* **69**, 503–512.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemp. Math.* **80**, 221–239.

Liptser, R. S., and Shiryaev, A. N. (1977). *Statistics of Random Processes I. General Theory.* Springer, New York.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Ann. Statist.* **16**, 1696–1708.

Mak, T. K. (1993). Solving non-linear estimating equations. *J. Roy. Statist. Soc. Ser. B* **55**, 945–955.

Martin, R. D. (1980). Robust estimation of autoregressive models (with discussion). In D. R. Brillinger and G. C. Tiao, Eds, *Directions of Time Series*, Inst. Math. Statist., Hayward, CA, 228–262.

Martin, R. D. (1982). The Cramér-Rao bound and robust M-estimates for autoregressions. *Biometrika* **69**, 437–442.

Martin, R. D., and Yohai, V. J. (1985). Robustness in time series and estimating ARMA models. In E. J. Hannan, P. R. Krishnaiah and M. M. Rao, Eds., *Handbook of Statistics* **5**, Elsevier Science Publishers, New York, 119–155.

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59–67.

McCullagh, P. (1991). Quasi-likelihood and estimating functions. In D. V. Hinkley, N. Reid and E. J. Snell, Eds., *Statistical Theory and Modelling. In Honour of Sir David Cox, FRS.* Chapman and Hall, London, 265–286.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd Ed., Chapman and Hall, New York.

McKeague, I. W. (1986). Estimation for a semimartingale model using the method of sieves. *Ann. Statist.* **14**, 579–589.

McLeish, D. L., and Small, C. G. (1988). *The Theory and Applications of Statistical Inference Functions.* Lecture Notes in Statistics **44**, Springer, New York.

Merkouris, T. (1992). A transform method for optimal estimation in stochastic processes: basic aspects. In J. Chen, Ed. *Proceedings of a Symposium in Honour of Professor V. P. Godambe*, University of Waterloo, Waterloo, Canada, 42 pp.

Morton, R. (1981a). Efficiency of estimating equations and the use of pivots. *Biometrika* **68**, 227–233.

Morton, R. (1981b). Estimating equations for an ultrastructural relationship. *Biometrika* **68**, 735–737.

Morton, R. (1987). A generalized linear model with nested strata of extra-Poisson variation. *Biometrika* **74**, 247–257.

Morton, R. (1988). Analysis of generalized linear models with nested strata of variation. *Austral. J. Statist.* **30A**, 215–224.

Morton, R. (1989). On the efficiency of the quasi-likelihood estimators for exponential families with extra variation. *Austral. J. Statist.* **31**, 194–199.

Mtundu, N. D., and Koch, R. W. (1987). A stochastic differential equation approach to soil moisture. *Stochastic Hydrol. Hydraul.* **1**, 101–116.

Mykland, P. A. (1995). Dual likelihood. *Ann. Statist.* **23**, 386–421.

Naik-Nimbalkar, U. V., and Rajarshi, M. B. (1995). Filtering and smoothing via estimating functions. *J. Amer. Statist. Assoc.* **90**, 301–306.

Nelder, J. A., and Lee, Y. (1992). Likelihood, quasi-likelihood and pseudo-likelihood: some comparisons, *J. Roy. Statist. Soc. Ser. B* **54**, 273–284.

Nelder, J. A., and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* **74**, 221–232.

Nguyen, H. T. and Pham, D. P. (1982). Identification of the nonstationary diffusion model by the method of sieves. *SIAM J. Optim. Control* **20**, 603–611.

Osborne, M. R. (1992). Fisher's method of scoring, *Internat. Statist. Rev.* **60**, 99–117.

Parzen, E. (1957). On consistent estimates of the spectrum of a stationary time series. *Ann. Math. Statist.* **28**, 329–348.

Pedersen, A. R. (1995). Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli* **1**, 257–279.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–1048.

Priestley, M. B. (1981). *Spectral Analysis and Time Series.* Academic Press, London.

Pukelsheim, F. (1993). *Optimal Design of Experiments.* Wiley, New York.

Qian, W., and Titterington, D. M. (1990). Parameter estimation for hidden Markov chains. *Statistics & Probability Letters* **10**, 49–58.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd Ed., Wiley, New York.

Rao, C. R., and Mitra, S. K. (1971). *Generalized Inverse of Matrices and its Applications*, Wiley, New York.

Rebolledo, R. (1980). Central limit theorems for local martingales. *Z. Wahrsch. Verw. Geb.* **51**, 269–286.

Reynolds, J. F. (1975). The covariance structure of queues and related processes - a survey of recent work. *Adv. Appl. Prob.* **7**, 383–415.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes.* Cambridge Univ. Press, Cambridge.

Rogers, L. C. G., and Williams, D. (1987). *Diffusions, Markov Processes and Martingales, Vol. 2, Ito Calculus.* Wiley, Chichester.

Rosenblatt, M. (1985). *Stationary Sequences and Random Fields.* Birkhäuser, Boston.

Samarov, A., and Taqqu, M. S. (1988). On the efficiency of the sample mean in long-memory noise. *J. Time Series Anal.* **9**, 191–200.

Samuel, E. (1969). Comparison of sequential rules for estimation of the size of a population. *Biometrics* **25**, 517–527.

Schuh, H.-J., and Tweedie, R. L. (1979). Parameter estimation using transform estimation in time-evolving models. *Math. Biosciences* **45**, 37–67.

Shen, X., and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22**, 580–615.

Shiryaev, A. N. (1981). Martingales, recent developments, results and applications. *Internat. Statist. Rev.* **49**, 199–233.

Shumway, R. H., and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the E-M algorithm. *J. Time Series Anal.* **3**, 253–264.

Small, C. G., and McLeish, D. L. (1989). Projection as a method for increasing sensitivity and eliminating nuisance parameters. *Biometrika* **76**, 693–703.

Small, C. G., and McLeish, D. L. (1991). Geometrical aspects of efficiency criteria for spaces of estimating functions. In V. P. Godambe, Ed., *Estimating Functions*, Oxford Science Publications, Oxford, 267–276.

Small, C. G., and McLeish, D. L. (1994). *Hilbert Space Methods in Probability and Statistical Inference,* Wiley, New York.

Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika* **72**, 67–92.

Smith, R. L. (1989). A survey of nonregular problems. *Bull. Internat. Statist. Inst.* **53**, Book 3, 353–372.

Sørensen, M. (1990). On quasi-likelihood for semimartingales. *Stochastic Process. Appl.* **35**, 331–346.

Sørensen, M. (1991). Likelihood methods for diffusions with jumps. In N. U. Prabhu and I. V. Basawa, Eds., *Statistical Inference in Stochastic Processes,* Marcel Dekker, New York, 67–105.

Stefanski, L. A., and Carroll, R. J. (1987). Conditional score and optimal scores for generalized linear measurement-error models. *Biometrika* **74**, 703–716.

Sweeting, T. J. (1986). Asymptotic conditional inference for the offspring mean of a supercritical Galton-Watson process. *Ann. Statist.* **14**, 925–933.

Thavaneswaran, A. (1991). Tests based on an optimal estimate. In V. P. Godambe, Ed., *Estimating Functions,* Oxford Science Publications, Oxford, 189–197.

Thavaneswaran, A., and Abraham, B. (1988). Estimation for non-linear time series using estimating equations, *J. Time Ser. Anal.* **9**, 99–108.

Thavaneswaran, A., and Thompson, M. E. (1986). Optimal estimation for semimartingales. *J. Appl. Prob.* **23**, 409–417.

Thisted, R. A. (1988). *Elements of Statistical Computing.* Chapman and Hall, New York.

Thompson, M. E., and Thavaneswaran, A. (1990). Optimal nonparametric estimation for some semimartingale stochastic differential equations. *Appl. Math. Computation* **37**, 169–183.

Tjøstheim, D. (1978). Statistical spatial series modelling. *Adv. Appl. Prob.* **10**, 130–154.

Tjøstheim, D. (1983). Statistical spatial series modelling II: some further results on unilateral lattice processes. *Adv. Appl. Prob.* **15**, 562–584.

Vajda, I. (1995). Conditions equivalent to consistency of approximate MLE's for stochastic processes. *Stochastic Process. Appl.* **56**, 35–56.

Verbyla, A. P. (1990). A conditional derivation of residual maximum likelihood. *Austral. J. Statist.* **32**, 227–230.

Vitale, R. A. (1973). An asymptotically efficient estimate in time series analysis. *Quart. Appl. Math.* **30**, 421–440.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20**, 595–601.

Watson, R. K., and Yip, P. (1992). A note on estimation of the infection rate. *Stochastic Process. Appl.* **41**, 257–260.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447.

Wei, C. Z., and Winnicki, J. (1989). Some asymptotic results for branching processes with immigration. *Stochastic Process. Appl.* **31**, 261–282.

Whittle, P. (1951). *Hypothesis Testing in Time Series Analysis.* Almqvist and Wicksell, Uppsala.

Whittle, P. (1952). Tests of fit in time series. *Biometrika* **39**, 309–318.

Whittle, P. (1953). The analysis of multiple time series. *J. Roy. Statist. Soc. Ser. B* **15**, 125–139.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika* **41**, 434–449.

Winnicki, J. (1988). Estimation theory for the branching process with immigration. *Contemp. Math.* **80**, 301-322.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261–1295.

Yanev, N. M., and Tchoukova-Dantcheva, S. (1980). On the statistics of branching processes with immigration. *C. R. Acad. Sci. Bulg.* **33**, 463–471.

Zeger, S. L., and Liang, K.-Y. (1986). Longitudial data analysis for discrete and continuous outcomes. *Biometrics* **44**, 1033–1048.

Zehnwirth, B. (1988). A generalization of the Kalman filter for models with state-dependent observation variance. *J. Amer. Statist. Assoc.* **83**, 164–167.

Zygmund, A. (1959). *Trigonometric Series*, Vol. 1, 2nd Ed., Cambridge Univ. Press, Cambridge.

Index

Aalen, O.O., 18, 19, 150, 166, 211
Aase, K.K., 177, 211
Abraham, B., 176, 177, 225
acceptable estimating function, 74
Adenstadt, R.K., 155, 156, 211
Aggoun, L., 136, 138, 214
Aitchison, J., 183, 211
algorithm, 127, 177
  E-M, 116–119, 123, 127
  P-S, 116–119, 123, 127
ancillary, 43, 60, 107 (see also E-ancillary)
Anderson, P.K., 18, 211
Anh, V.V., 160, 161, 211
ARMA model, 23
asymptotically non-negative definite, 71, 72
asymptotic mixed normality, 27, 60, 62–64
asymptotic normality, 26, 27, 40, 41, 54–56, 60, 62–64, 110, 136, 147, 159, 161, 163, 164, 179, 191, 196, 197, 201
asymptotic first order efficiency, 72
asymptotic quasi-likelihood (see quasi-likelihood, asymptotic)
asymptotic relative efficiency, 74, 125, 126, 167, 168, 175
autoregressive process, 19, 23, 31, 56, 79, 98, 170, 173, 174, 176
  spatial, 137
  conditional, 137
  random coefficient, 176
  threshold, 176

Baddeley, A.J., 201, 211
Bailey, N.T.J., 162, 163, 211
Barndorff-Nielsen, O.E., 35, 56, 59, 62, 86, 211
Banach space, 147
Basawa, I.V., 54, 60, 63, 132, 135, 141, 142, 169, 173, 185, 212
Becker, N.G., 165, 168, 212
Belyea, C., 200, 215
Beran, J., 158, 212
Berliner, L.M., 37, 212
Bernoulli distribution, 87
Besag, J.E., 91, 212
best linear unbiased estimator (BLUE), 153, 154
beta distribution, 124
Bhapkhar, V.P., 115, 212
bias correction, 61, 67
Bibby, B.M., 135, 212
binary response, 25
Billingsley, P., 65, 212
binomial distribution, 122, 167
birth and death process, 18
birth process, 56
Black-Scholes model, 31
bootstrap, 9, 210
Borgan, O., 18, 211
Bradley, E.L., 7, 212
branching process, 182
  Galton-Watson, 2, 15, 27, 31, 35, 36, 69, 87, 195
Brockwell, P.J., 60, 212
Brouwer fixed point theorem, 183
Brownian motion, 17, 31, 33, 34, 131–133, 136, 196
Burkholder, Davis, Gundy inequality, 149
Bustos, O.H., 169, 213

cadlag, 191
Carroll, R.J., 131, 139, 203, 205, 213, 214, 225
Cauchy distribution, 40, 200
Cauchy-Schwarz inequality, 5, 24, 167
censored data, 150
central limit results, 4, 9, 26, 27, 64, 149, 155, 166, 167, 179, 190–195
Chan, N.H., 56, 213
Chandrasekar, B., 19, 213
characteristic function, 199
Chebyshev inequality, 189
Chen, K., 88, 213
Chen, Y., 182, 213
Cheng, R.C.H., 54, 213
chi-squared distribution, 58, 110, 143–145, 172
Choi, Y.J., 163, 213
Cholesky square root matrix, 71, 72
Chung, K.L., 97, 213
coefficient of variation, 121, 124
colloidal solution, 69
Comets, F., 179, 213
complete probability space, 92
conditional centering, 179
conditional inference, 107
confidence intervals (or zones), 1, 4, 8, 9, 24, 27, 53–67, 69, 71, 88, 110, 131, 158, 163, 171, 191, 210
  average, 56, 66
  simulteneous, 58
consistency, 6, 8, 10, 26, 38, 40, 54, 55, 63, 64, 70, 131, 135, 147–150, 161, 163, 179–186, 190, 196, 197, 201, 203
constrained parameter estimation, 8, 107–112, 142
control, 41
convergence
  stable, 191, 193, 195
  mixing, 57, 191, 192
convex, 14, 46, 156–158
correlation, 6, 13, 25
counting process, 18, 112, 150, 151, 165
covariate, 25, 148
Cox, D.R., 35, 41, 54, 58, 61, 62, 107, 180, 211, 213
Cox, J.S., 131, 133, 213
Cox-Ingersoll-Ross model, 131, 133
Cramér, H., 2, 7, 180, 185, 195, 213
Cramér-Rao inequality, 2, 7
Cramér-Wold device, 195
cross-sectional data, 159
Crowder, M., 95, 105, 213
cumulant spectal density, 159
cumulative hazard function, 9, 150
curvature, 62
Cutland, N.J., 31, 213

Dahlhaus, R., 86, 214
Daley, D.J., 156, 214
Davidian, M., 131, 214
Davis, R.A., 19, 214
demographic stochasticity, 36
Dempster, A.P., 116, 214
Denby, L., 169, 214
dererminant criterion for optimality, 19
Desmond, A.F., 2, 25, 214
differentiability (stochastic), 56
differential geometry, 62
diffusion, 8, 17, 129, 131, 132, 135, 148, 151, 200
Dion, J.-P., 10, 214
discrete exponential family, 2, 16
dispersion, 12, 22, 119, 124, 129
distance, 7, 12
Doob-Meyer decomposition, 27
drift coefficient, 131, 148
Doukhan, P., 179, 214
Duffie, D., 31, 214
Durbin, J., 1, 214
dynamical system, 37, 38, 40, 131, 136

E-ancillary, 8, 43–51, 113, 114
Edgeworth expansion, 62
efficient score statistic, 9, 141, 142

Efron, B., 59, 214
eigenfunction, 200, 201
Eisenberg, B., 155, 211
Elliott, R.J., 50, 136, 138, 214
ellipsoids (of Wald), 58
E-M algorithm, 8, 103, 107, 116, 117, 119–123, 127
epidemic, 9, 162, 163
equicontinuity (stochastic), 56
ergodic, 134, 142, 153
ergodic theorem, 134, 153, 208
error
  approximation, 148, 151
  contrasts, 139
  estimation, 148, 151
estimating functions, 1, 2
  combination, 2, 8, 35, 103, 137, 138
  optimal, 4–6
  robust, 9, 169–175
  standardized, 3–5, 118, 138
estimating function space, 3–5, 8, 11, 13, 15–19, 22–25, 28, 32, 36–40, 43–51, 71, 75, 88, 89, 92, 94–96, 99, 100, 104, 105, 111, 130, 132, 137, 138, 153, 162–164, 166, 169–171, 176, 195, 199–201
  Hutton-Nelson, 32, 33, 162–164
  convex, 14, 46–48
estimation
  constrained parameter, 8, 107–112, 142
  function, 147–151
  recursive, 176–177
  robust, 169–175
E-sufficient, 8, 43–51, 113, 114
Euclidean space, 11, 92, 94, 141, 182
Euler scheme, 151
exponential distribution, 41, 121, 125
exponential family, 2, 7, 16, 24, 38, 53, 79, 125

failure, 38
F-distribution, 110
Feigin, P.D., 17, 200, 214, 215
Fejer kernel, 157
Ferland, R., 10, 214
field (see random field)
filtering, 8, 103
filtration, 27, 30, 54, 94, 132
finance, 31, 135
finite variation process, 31, 148
first order efficient, 72, 73
Firth, D., 61, 67, 95, 102, 105, 215
Fisher, R.A., 1, 2, 12, 40, 59, 72, 97, 107, 141, 202
Fisher information, 2, 12, 40, 59, 72, 141
  conditional, 97
Fisher method of scoring, 120, 202
Fitzmaurice, G.M., 25, 26, 215
Fourier methods, 85
Fox, R., 86, 215
fractional Brownian motion, 31, 86
function estimation, 9, 147–151
functional relationships, 112, 205
Fürth, R., 87, 215

Galton-Watson process, 2, 15, 27, 31, 36, 69, 195
gamma distribution, 56, 134
Gastwirth, J.L., 169, 215
Gauss, C.F. i, 1, 3, 25, 83, 84, 86, 127, 138, 158, 159, 161, 215
Gaussian distribution, 83, 86, 127, 138, 158, 159 (see also normal distribution)
Gauss-Markov theorem, 3, 4, 161
Gay, D.M., 202, 215
Gay, R., 74, 82, 88, 218
generalized estimating equation (GEE), 8, 25, 26, 89
generalized inverse, 30, 108, 109, 130, 132, 184
generalized linear model (GLIM), 21, 22, 79, 100, 104, 202
geometric distribution, 38
Gibbs field, 138
Gibbs sampler, 201
Gill, R.D., 18, 211

Girsanov transformation, 138
Glynn, P., 64, 215
Godambe, V.P., 1, 2, 21, 38, 48, 97, 105, 107, 116, 172, 215, 216
Gram-Schmidt orthogonalization, 93, 96
Greenwood, P.E., 56, 216
Grenander, U., 147, 156, 216
Guyon, X., 179, 201, 216

Hajek convolution theorem, 63
Halfin, S., 157, 216
Hall, P., 54, 57, 63, 87, 156, 180, 181, 187, 192, 195, 216
Hanfelt, J.J., 203, 216
Hannan, E.J., 83, 85, 86, 153, 216, 217
Harris, I.R., 102, 215
Harville, D., 130, 217
hazard function, 150
Hermite-Chebyshev polynomials, 97
heteroscedastic
  autoregression, 79
  regression, 9, 159–161
Heyde, C.C., 1, 2, 13, 38, 48, 54, 57, 62, 63, 69, 72, 74, 82, 87, 88, 92, 94, 97, 107, 116, 126, 131, 136, 153, 156, 159, 161, 165, 168, 169, 172, 180, 181, 187, 192, 195, 203, 212, 213, 216–218, 220, 222
hidden Markov models, 9, 93, 136, 139
Hinkley, D.V., 35, 41, 54, 58, 59, 61, 107, 180, 213, 214
Hilbert Space, 13, 44
Hoffmann-Jørgensen, J., 56, 218
Holder inequality, 86
Hotelling, H., 13, 218
Huber, P.J., 169, 173, 219
Huber function, 173
Huggins, R.M., 169, 173, 212
Hutton, J.E., 32–36, 56, 61, 97, 150, 151, 162–164, 184, 191, 196, 219
Hutton-Nelson estimating function, 32–36, 150, 151, 162–164, 196
hypothesis testing, 9, 141–145

Ibragimov, I.A., 155, 219
idempotent matrix, 99, 100, 143
Iglehart, D.L., 64, 215
immigration distribution, 69, 70, 87
Ingersoll, J.E., 131, 133, 213
infection rate, 9, 162
infectives, 162, 164
infinitesimal generator, 9, 200, 201
information, 2, 7, 8, 12, 40, 41, 55, 72, 92, 108, 113, 114, 118, 126, 142, 159
  empirical, 59, 204
  expected, 59, 204
  Fisher, 2, 12, 40, 59, 72, 97, 141
  martingale, 28, 96–98, 160, 166, 167, 172
  observed, 59
integration by parts, 190
intensity, 18, 34, 56, 135, 148, 165
interest rate, 133
invariance, 2, 58
Ito formula, 32, 97, 135, 194

Janžura, M., 179, 213
jackknife, 9, 210
Jensen inequality, 66
Jiang, J., 131, 219
Judge, G.G., 111, 219

Kabaila, P.V., 85, 87, 219
Kale, B.K., 2, 19, 211, 216
Kallianpur, G.K., 33, 148, 219
Kalman filter, 8, 103
Karlin, S., 200, 219
Karr, A., 148, 219
Karson, M.J., 130, 219
Kaufmann, H., 186, 219, 220
Keiding, N., 18, 211
kernel estimation, 147
Kessler, M., 135, 200, 201, 220

Kimball, B.F., 1, 220
Kloeden, P.E., 134, 135, 151, 220
Koch, R.W., 197, 223
Kopp, P.E., 31, 213
Kronecker lemma, 190
Kronecker product, 80
Künsch, H., 169, 210, 220
Kulkarni, P.M., 169, 220
Kulperger, R., 85, 220
Kunita-Watanabe inequality, 50
Kutoyants, Yu., 62, 220
kurtosis, 62, 98, 104, 130, 139

Lagrange multiplier, 108, 111–113
Lahiri, S.N., 208, 220
Laird, N.M., 25, 26, 116, 122–124, 212, 215, 220
Langevin model, 131, 135
Laplace, P.S. de, 1
Laplace transform, 200
lattice, 81, 136
law of large numbers, 120, 204, 205, 207, 208 (see also martingale strong law)
least squares, 1, 2, 3, 5, 7, 10, 21, 87, 161, 202–205, 209
Le Cam, L., 53, 58, 220, 221
Lee, Y., 23, 223
Legendre, A., 1
Lele, S., 37, 38, 210, 221
Lépingle, D., 187, 196, 221
Leskow, J., 148, 221
lexicographic order, 101
Li, B., 62, 142, 180, 203, 221
Liang, K.Y., 21, 25, 221, 226
lifetime distribution, 150
likelihood, 6, 8, 10, 16, 25, 35, 40, 41, 53, 54, 58, 72, 83, 91, 111, 117, 118, 122, 123, 127, 129, 133, 138, 141, 142, 165, 179, 180, 184, 202, 203
    conditional, 107
    constrained
    non-regular cases, 40, 41
    partial, 107
likelihood ratio, 58, 63, 131, 134, 141, 142, 145, 203
Lin, Y.-X., 43, 88, 105, 159, 186, 218, 221
Lindeberg-Feller theorem, 4
Lindsay, B., 91, 107, 139, 222
Linnik, Yu.V., 155, 219
Liptser, R.S., 133, 222
Little, R.J.A., 127, 222
Liu, R.Y., 208, 222
Loewner optimality, 12, 20
Loewner ordering, 12, 55, 118, 144
logistic map, 37, 40
logit link function, 25
longitudinal data, 8, 25
long-range dependence, 82, 86, 158, 159

Mak, T.K., 202, 222
Markov, A.A. i, 3, 9, 93, 139, 157, 161, 162, 200, 201
Markov process, 9, 157, 162, 200, 201
Martin, R.D., 169, 172, 214, 222
martingale, 15, 17, 18, 26–28, 35, 48, 49, 51, 59, 61, 62, 69, 70, 93, 94–98, 131–133, 135, 136, 148–150, 159, 160, 162–167, 169–171, 176, 180, 181, 186–196, 200, 210
    central limit theorem, 55, 179, 186, 190–195
    continuous part, 34, 54
    information (see information, martingale)
    strong law, 55, 150, 174, 179, 181, 182, 186–190, 195, 196
maximum likelihood, 1, 2, 16–19, 21, 34, 35, 38, 39, 41, 53, 54, 57, 58, 61, 70, 92, 98, 116, 124, 129, 131, 134, 136, 141, 148, 160, 163, 165, 166, 169, 180, 200, 202
    non-parametric, 17
    regularity conditions, 40, 41, 54, 62

restricted, 141
McCormick, W.P., 19, 214
McCullagh, P., 21, 22, 125, 142, 184, 203, 205, 221, 222
McKeague, I.W., 148, 222
McLeish, D.L., 1, 3, 8, 13, 43–45, 49, 127, 128, 222, 225
measurement errors, 139
membrane potential, 33, 147
Merkouris, T., 13, 199, 223
M-estimation, 1, 142
method of moments, 1, 153, 154, 156, 169, 202
metric space, 200
minimum chi-squared, 1
minimal sufficient, 1
missing data, 8, 107, 116, 117, 127
Mitra, S.K., 131, 224
mixed normality, 27, 62, 63, 192
mixing conditions, 84, 156, 179
mixture densities, 127
models
branching (see branching process)
epidemic, 162–164
hidden Markov, 93
interest rate, 133
logistic, 37, 38, 40
membrane potential, 33, 147
multi-stage, 200
nearest neighbour, 91
nested, 99–102
particles in a fluid, 87
physician services, 127
population, 35–37
queueing, 157, 158
recapture, 164–168
risky asset, 31
soil moisture, 196–197
moment generating function, 199
Moore, J.B., 136, 138, 214
Moore-Penrose inverse, 30
Morton, R., 21, 23, 100–102, 107, 112, 116, 126, 203, 218, 223
Mtundu, N.D., 197, 223
multiple roots, 202–209
multiplicative errors, 100, 102
mutual quadratic characteristic, 26
Mykland, P.A., 62, 223

Naik-Nimbalkar, U.V., 103, 104, 222
Nelder, J.A., 21, 22, 125, 184, 202, 222, 223
Nelson, P.I., 32–36, 56, 61, 97, 150, 151, 162–164, 184, 191, 196, 219
nested strata, 8, 99
neurophsiological model, 32, 147
Newton-Raphson method, 202
Nguyen, H.T., 148, 223
noise, 17, 30, 31, 33, 127, 132, 133, 135, 136
additive, 36, 37
multiplicative, 36
non-ergodic models, 59, 60, 63, 144
nonparametric estimation, 147, 150
norm (Euclidean), 55, 71, 184, 186
normal distribution, 4, 27, 28, 41, 55–57, 62–66, 88, 91, 105, 111, 116, 121, 129–131, 139, 156, 161, 163, 164, 166, 173
nuisance parameter, 8, 57, 59, 60, 70, 71, 88, 107, 113–115, 139, 159

offspring distribution, 16, 27, 69, 70, 87, 182, 195
Ogunyemi, O.T., 56, 184, 196, 219
on-line signal processing, 9, 176
optimal
asymptotic (O_A), 1, 2, 7, 8, 11, 12, 29, 30
experimental design, 8, 12, 20
fixed sample (O_F), 1, 5, 7, 11–21, 28, 30, 55
orthogonal, 13, 71, 93, 100, 102–104, 138, 156, 160
orthonormal sequence, 147
Osborne, M.R., 120, 223
outliers, 169

parameters

constrained, 107–112, 142
incidental, 112
nuisance (see nuisance parameters)
scale, 200
parametric model, 1, 11
parameter space, 41, 43, 50, 53
partial likelihood, 107
Parzen, E., 84, 223
Pearson, K., 1
Pedersen, A.R., 135, 223
periodogram, 82
Pham, D.P., 148, 223
pivot, 66
Platen, E., 134, 135, 151, 220
point process, 148 (see also counting process, Poisson process)
Poisson distribution, 33, 67, 70, 87, 100, 121, 167
Poisson process, 19, 31, 34, 35, 111, 112, 135, 196, 197
compound, 35, 196
Pollard, D., 56, 223
population process, 35 (see also branching process)
power series family, 2, 16
Prakasa-Rao, B.L.S., 54, 132, 135, 141, 212
predictable process, 18, 31, 51, 94, 97, 162
prediction variance, 82
Pregibon, D., 202, 223
Prentice, R.L., 25, 224
Priestley, M.B., 84, 224
probability generating function, 167
projected estimating function, 108–110, 143
projected estimator, 13, 108, 143, 144
projection, 13, 47, 107–109, 114, 125, 129
P-S algorithm, 107, 116
Pukelsheim, F., 12, 20, 38, 224
purely nondeterministic, 153

Qian, W., 136, 224
quadratic characteristic, 27, 29, 59, 94, 132, 144, 151, 166, 187, 192, 196
quadratic form, 130
quadratic variation process, 33, 54, 133
quasi-likelihood, 1, 7–9, 12, 16, 18, 19, 21, 22, 36, 38, 40, 53, 61, 69, 70, 73, 87, 107, 116, 121, 129, 130, 131, 134, 138, 139, 142, 147, 148, 150, 151, 160–165, 169, 180–182, 195–197
asymptotic, 8
composite, 91, 92
quasi-score, 7–9, 12, 13, 22, 23, 26, 35, 39, 43, 46, 48, 50, 51, 55, 58, 62, 67, 69–72, 80, 82, 83, 85, 87, 92–96, 98–100, 102, 104, 105, 108, 109–112, 116–118, 122, 125, 127, 128, 130, 132, 134, 135, 137–139, 142, 144, 145, 154, 159, 162, 163, 166, 170–173, 175, 176, 179, 182, 195, 196
asymptotic, 26, 35, 37, 58, 69, 72—76, 78–80, 88, 154
combined, 98, 99, 103, 160
conservative, 142, 203
existence of, 15, 21, 36–38
Hutton-Nelson, 33–36, 61, 150, 151, 162–164
robust, 170–174
sub-, 48, 49
queueing models, 157, 158

Radon-Nikodym derivitive, 129, 131–133
Rajarshi, M.B., 103, 104, 223
random coefficient autoregression, 176
random effects model, 105
random environment, 35, 36
random field, 8, 82, 93, 136, 137, 179, 201
random norming, 8, 56, 63, 189

Rao, C.R., 2, 7, 20, 28, 29, 72, 73, 95, 108, 128, 129, 131, 141, 224
Rao-Blackwell theorem, 128
Rebolledo, R., 166, 224
recapture experiment, 9, 164
recursive estimation, 9, 176, 177
regression, 9, 21, 23, 25, 60, 89, 111, 113, 148, 159, 170, 173, 174, 205
REML estimation, 8, 129–131
removal rate, 164
resampling, 9, 210
residuals, 99
Reynolds, J.F., 157, 158, 224
Riemann zeta function, 126
Ripley, B.D., 137, 224
robust methods, 9, 169
Rogers, L.C.G., 26, 31, 54, 133, 149, 224
Rosenblatt, M., 82, 84, 85, 179, 224
Rotnitzky, A.G., 25, 26, 215
Rozanski, R., 148, 221
Rubin, D.B., 116, 127, 214, 222
Rubin, H., 169, 215

Samarov, A., 158, 224
sample space, 2, 43
Samuel, E., 165, 224
Schuh, H.-J., 200, 224
score function, 2, 4, 6–8, 17, 24, 41, 62, 67, 91, 93, 94, 107, 111, 113, 115–118, 122, 127, 134, 139, 141, 160, 181, 205
Scott, D.J., 63, 142, 185, 212
screening tests, 122
Selukar, R.S., 148, 219
semimartingale, 8, 9., 30–36, 54, 69, 93, 97, 131, 135, 148
semiparametric model, 1, 9
Seneta, E., 69, 87, 218
service time, 157
Severo, N., 163, 213
Shen, X., 148, 224
Shiryaev, A.N., 26, 133, 222, 224
short-range dependence, 82, 158, 159
Shumway, R.H., 127, 224
sieve, 9, 147, 148, 151
signal, 30–32, 132, 136
Silvey, S.D., 183, 211
simulation, 64
simultaneous reduction lemma, 20, 21
skewness, 62, 98, 104
Slutsky theorem, 65
Small, C.G., 1, 3, 8, 13, 43–45, 49, 127, 128, 222, 225
Smith, R.L., 54, 225
smoothed periodogram, 82
smoothing, 8, 103, 104
smoothing function, 82, 84
soil moisture, 197
Sørensen, M., 34, 35, 56, 59, 135, 136, 191, 197, 200, 201, 212, 220, 225
sources of variation, 33–35, 101, 136, 137, 139
spatial process, 93
spectral density, 82, 86, 153, 155, 156, 158, 159
square root matrix, 71, 72
standardization, 3–5, 11, 71
state space models, 103
stationary process, 1, 153, 156, 158, 210
Staudte, R.G., 169, 173, 212
Stefansky, L.A., 139, 203, 205, 206, 213, 225
stochastic differential equation, 32, 49, 133, 135, 196, 197
stochastic disturbance (see noise)
stochastic integral, 133, 163
Stoffer, D.S., 127, 224
stopping time, 189, 193
strata, 99, 100, 102
strong consistency, 56, 70, 135, 148, 161, 163, 181, 182, 185, 186, 190, 195
structural relationships, 112, 206
sufficiency, 2, 3, 43 (see also E-sufficient)
surrogate predictors, 139
Sweeting, T.J., 60, 225

Szegö, G., 156, 216

Takayama, T., 111, 219
Taqqu, M.S., 86, 158, 215, 224
Taylor, H.M., 200, 219
Tchoukova-Dantcheva, S., 70, 226
Thavaneswaran, A., 18, 142, 147, 176, 177, 225
Thisted, R.A., 202, 225
Thompson, M.E., 18, 105, 116, 147, 216, 225
time series, 8, 103, 176
  moving average, 156, 159, 180 (see also autoregressive process)
  standardized, 64
Titterington, D.M., 136, 224
Toeplitz matrix, 154
trace criterion (for optimality), 19
traffic intensity, 158
transform martingale families, 199
trapping rate, 168
Traylor, L., 54, 213
treatment, 4
t-statistic, 64, 66, 168
Tweedie, R.L., 200, 215, 22

unacceptable estimating function, 74, 75
uniform convergence, 57
uniform distribution, 41

Vajda, I., 180, 226
Verbyla, A.P., 130, 226
Vitale, R.A., 156, 226
Vostrikova, L., 62, 220

Wald, A., 9, 58, 141–143, 180, 226
Wald test, 9, 141–143
Watson, R.K., 162, 226
Wedderburn, R.W.M. i, 7, 21, 23–25, 101, 226
Wedderburn estimating function, 24
Wefelmeyer, W., 56, 216
Wei, C.Z., 56, 70, 87, 211, 226
Welsh, R.E., 202, 215
Whittle, P. ii, 82, 83, 86, 226
Whittle estimator, 83, 86
Williams, D., 26, 31, 54, 133, 149, 224
Williams, R.J., 97, 213
Willinger, W., 31, 213
Winnicki, J., 69, 70, 87, 226
Wong, W.H., 148, 224
Wu, C.F.J., 208, 226

Yanev, N.M., 70, 226
Yip, P., 162, 226
Yohai, V.J., 169, 172, 222
Yule-Walker equations, 23

Zeger, S.L., 21, 25, 200, 226
Zehnwirth, B., 103, 226
Zygmund, A., 157, 226

Springer Series in Statistics

(continued from p. ii)

Mosteller/Wallace: Applied Bayesian and Classical Inference: The Case of *The Federalist Papers.*
Pollard: Convergence of Stochastic Processes.
Pratt/Gibbons: Concepts of Nonparametric Theory.
Ramsay/Silverman: Functional Data Analysis.
Read/Cressie: Goodness-of-Fit Statistics for Discrete Multivariate Data.
Reinsel: Elements of Multivariate Time Series Analysis, 2nd edition.
Reiss: A Course on Point Processes.
Reiss: Approximate Distributions of Order Statistics: With Applications to Non-parametric Statistics.
Rieder: Robust Asymptotic Statistics.
Rosenbaum: Observational Studies.
Ross: Nonlinear Estimation.
Sachs: Applied Statistics: A Handbook of Techniques, 2nd edition.
Särndal/Swensson/Wretman: Model Assisted Survey Sampling.
Schervish: Theory of Statistics.
Seneta: Non-Negative Matrices and Markov Chains, 2nd edition.
Shao/Tu: The Jackknife and Bootstrap.
Siegmund: Sequential Analysis: Tests and Confidence Intervals.
Simonoff: Smoothing Methods in Statistics.
Small: The Statistical Theory of Shape.
Tanner: Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd edition.
Tong: The Multivariate Normal Distribution.
van der Vaart/Wellner: Weak Convergence and Empirical Processes: With Applications to Statistics.
Vapnik: Estimation of Dependences Based on Empirical Data.
Weerahandi: Exact Statistical Methods for Data Analysis.
West/Harrison: Bayesian Forecasting and Dynamic Models, 2nd edition.
Wolter: Introduction to Variance Estimation.
Yaglom: Correlation Theory of Stationary and Related Random Functions I: Basic Results.
Yaglom: Correlation Theory of Stationary and Related Random Functions II: Supplementary Notes and References.