for the parameter ξ, and the value obtained is $\hat{\xi}_{\rm obs}$. The goal is to determine an interval $[a, b]$ given the data $x_1, \ldots, x_n$ such that the probabilities $P[a < \xi] = \alpha$ and $P[\xi < b] = \beta$ hold for fixed α and β regardless of the true value ξ.

The confidence interval is found by solving equations (9.9) for a and b,

$$\begin{aligned} \alpha &= \int_{\hat{\xi}_{\rm obs}}^{\infty} g(\hat{\xi}; a)\, d\hat{\xi}, \\ \beta &= \int_{-\infty}^{\hat{\xi}_{\rm obs}} g(\hat{\xi}; b)\, d\hat{\xi}. \end{aligned} \tag{10.30}$$

Figure 10.2 shows the 68.3% confidence intervals for various values of n assuming a measured value $\hat{\xi}_{\rm obs} = 1$. Also shown are the intervals one would obtain from the measured value plus or minus the estimated standard deviation. As n becomes larger the p.d.f. $g(\hat{\xi}; n, \xi)$ becomes Gaussian (as it must by the central limit theorem) and the 68.3% central confidence interval approaches $[\hat{\xi}_{\rm obs} - \hat{\sigma}_{\hat{\xi}}, \hat{\xi}_{\rm obs} + \hat{\sigma}_{\hat{\xi}}]$.

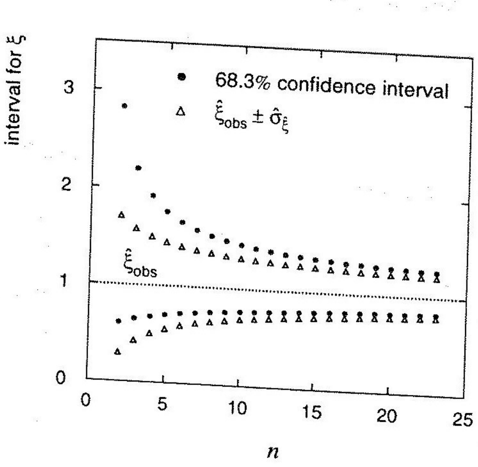

Fig. 10.2 Classical confidence intervals for the parameter of the exponential distribution ξ (between solid points) and the interval $[\hat{\xi}_{\rm obs} - \hat{\sigma}_{\hat{\xi}}, \hat{\xi}_{\rm obs} + \hat{\sigma}_{\hat{\xi}}]$ (between open triangles) for different values of the number of measurements n, assuming an observed value $\hat{\xi}_{\rm obs} = 1$.

11 Unfolding

Up to now we have considered random variables such as particle energies, decay times, etc., usually with the assumption that their values can be measured without error. The present chapter concerns the distortions to distributions which occur when the values of these variables are subject to additional random fluctuations due to the limited resolution of the measuring device. The procedure of correcting for these distortions is known as **unfolding**. The same mathematics can be found under the general heading of **inverse problems**, and is also called **deconvolution** or **unsmearing**. Although the presentation here is mainly in the context of particle physics, the concepts have been developed and applied in fields such as optical image reconstruction, radio astronomy, crystallography and medical imaging.

The approach here, essentially that of classical statistics, follows in many ways that of [Any91, Any92, Bel85, Zhi83, Zhi88]. Some of the methods have a Bayesian motivation as well, however, cf. [Siv96, Ski85, Ski86, Jay86].

In Section 11.1 the unfolding problem is formulated and the notation defined. Unfolding by inversion of the response matrix is discussed in Section 11.2. This technique is rarely used in practice, but is a starting point for better solutions. A simple method based on correction factors is shown in Section 11.3. The main topic of this chapter, regularized unfolding, is described in Sections 11.4 through 11.7. This includes the strategy used to find the solution, a survey of several regularization functions, and methods for estimating the variance and bias of the solution. These points are illustrated by means of examples in Section 11.8, and information on numerical implementation of the methods is given in Section 11.9.

It should be emphasized that in many problems it is not necessary to unfold the measured distribution, in particular if the goal is to compare the result with the prediction of an existing theory. In that case one can simply modify the prediction to include the distortions of the detector, and this can be directly compared with the measurement. This procedure is considerably simpler than unfolding the measurement and comparing it with the original (unmodified) theory.

Without unfolding, however, the measurement cannot be compared with the results of other experiments, in which the effects of resolution will in general be different. It can also happen that a new theory is developed many years after a measurement has been carried out, and the information needed to modify the theory for the effects of resolution, i.e. the response function or matrix (see below), may no longer be available. If a particularly important measured distribution is to retain its value, then both the measurement and the response matrix should be preserved. Unfortunately, this is often impractical, and it is rarely done.

By unfolding the distribution one provides a result which can directly be compared with those of other experiments as well as with theoretical predictions. Other reasons for unfolding exist in applications such as image reconstruction, where certain features may not be recognizable in the uncorrected distribution. In this chapter we will assume that these arguments have been considered and that the decision has been made to unfold.

11.1 Formulation of the unfolding problem

Consider a random variable x whose p.d.f. we would like to determine. In this chapter we will allow for limited resolution in the measurement of x, as well as detection efficiency less than 100% and the presence of background processes. As an example, we could consider the distribution of electron energies resulting from the beta decay of radioactive nuclei, i.e. the variable x refers to the energy of the emitted electron.

By 'limited resolution' we mean that because of measurement errors, the measured values of x may differ in a random way from the values that were actually created. For example, a particular beta decay may result in an electron with a certain energy, but because of the resolution of the measuring device, the recorded value will in general be somewhat different. Each observed event is thus characterized by two quantities: a true value y (which we do not know) and an observed value x.

In general one must also allow for the occurrence of a true value that does not result in any measured value at all. For the example of beta decay, it could be that an emitted electron escapes completely undetected, since the detector may not cover the entire solid angle surrounding the radioactive source, or electron energies below a certain minimum threshold may not produce a sufficiently large signal to be detected. The probability that an event leads to some measured value is called the detection **efficiency**[1] $\varepsilon(y)$, which in general depends on the true value of the event, y.

Suppose the true values are distributed according to the p.d.f. $f_{\text{true}}(y)$. In order to construct a usable estimator for $f_{\text{true}}(y)$, it is necessary to represent it by means of some finite set of parameters. If no functional form for $f_{\text{true}}(y)$ is known a priori, then it can still be represented as a normalized histogram with M bins. The probability to find y in bin j is simply the integral over the bin,

[1] If the reason that the event went undetected is related to the geometry, e.g. limited solid angle of the detector, then the efficiency is often called acceptance. The term efficiency is sometimes used to refer to the conditional probability that an event is detected given that it is contained in the sensitive region of the detector. Here we will use efficiency in the more general sense, meaning the overall probability for an event to be detected.

$$p_j = \int_{\text{bin } j} f_{\text{true}}(y)\, dy. \tag{11.1}$$

Suppose we perform an experiment in which a certain total number of events m_{tot} occur; this will differ in general from the number observed. The number m_{tot} could be treated as fixed or as a random variable. In either case, we will call the expectation value of the total number of events $\mu_{\text{tot}} = E[m_{\text{tot}}]$, so that the expected number of events in bin j is

$$\mu_j = \mu_{\text{tot}}\, p_j. \tag{11.2}$$

We will refer to the vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$ as the 'true histogram'. Note that these are not the actual numbers of events in the various bins, but rather the corresponding expectation values, i.e. the μ_i are not in general integers. One could, for example, regard the true number of events in bin i as a random variable m_i with mean μ_i. Because of the limited resolution and efficiency, however, m_i is not directly observable, and it does not even enter the present formulation of the problem. Instead, we will construct estimators directly for the parameters μ_i.

For reasons of convenience one usually constructs a histogram of the observed values as well. Suppose that we begin with a sample of measured values of x, and that these are entered into a histogram with N bins, yielding $\mathbf{n} = (n_1, \ldots, n_N)$. These values could also be sample moments, Fourier coefficients, etc. In fact, the variable x could be multidimensional, containing not only a direct measurement of the true quantity of interest y, but also correlated quantities which provide additional information on y.

The number of bins N may in general be greater, less than, or equal to the number of bins M in the true histogram. Suppose the ith bin contains n_i entries, and that the total number of entries is $\sum_i n_i = n_{\text{tot}}$. It is often possible to regard the variables n_i as independent Poisson variables with expectation values ν_i. That is, for this model the probability to observe n_i entries in bin i is given by

$$P(n_i; \nu_i) = \frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!}. \tag{11.3}$$

Since a sum of Poisson variables is itself a Poisson variable (cf. Section 10.4), n_{tot} will then follow a Poisson distribution with expectation value $\nu_{\text{tot}} = \sum_i \nu_i$. We may also consider the case where n_{tot} is regarded as a fixed parameter, and where the n_i follow a multinomial distribution. Whatever the distribution, we will call the expectation values

$$\nu_i = E[n_i]. \tag{11.4}$$

The form of the probability distribution for the data $\mathbf{n} = (n_1, \ldots, n_N)$ (Poisson, multinomial, etc.) will be needed in order to construct the likelihood function, used in unfolding methods based on maximum likelihood. Alternatively, we may be given the covariance matrix,

$$V_{ij} = \mathrm{cov}[n_i, n_j], \tag{11.5}$$

which is needed in methods based on least squares. We will assume that either the form of the probability law or the covariance matrix is known.

By using the law of total probability, (1.27), the expectation values $\nu_i = E[n_i]$ can be expressed as

$$\begin{aligned}\nu_i &= \mu_{\mathrm{tot}} \, P(\text{event observed in bin } i) \\ &= \mu_{\mathrm{tot}} \int dy \, P\left(\begin{array}{c} \text{observed} \\ \text{in bin } i \end{array} \middle| \begin{array}{c} \text{true value } y \text{ and} \\ \text{event detected} \end{array} \right) \varepsilon(y) \, f_{\mathrm{true}}(y) \\ &= \mu_{\mathrm{tot}} \int_{\mathrm{bin}\, i} dx \int dy \, s(x|y) \, \varepsilon(y) \, f_{\mathrm{true}}(y). \end{aligned} \tag{11.6}$$

Here $s(x|y)$ is the conditional p.d.f. for the measured value x given that the true value was y, and given that the event was observed somewhere, i.e. it is normalized such that $\int s(x|y)dx = 1$. We will call s the **resolution function** or in imaging applications the **point spread function**. One can also define a **response function**,

$$r(x|y) = s(x|y) \, \varepsilon(y), \tag{11.7}$$

which gives the probability to observe x, including the effect of limited efficiency, given that the true value was y. Note that this is not normalized as a conditional p.d.f. for x. One says that the true distribution is **folded** with the response function, and thus the task of estimating f_{true} is called **unfolding**.

Breaking the integral over y in equation (11.6) into a sum over bins and multiplying both numerator and denominator by μ_j, the expected number of entries to be observed in bin i becomes

$$\begin{aligned}\nu_i &= \sum_{j=1}^{M} \frac{\int_{\mathrm{bin}\, i} dx \int_{\mathrm{bin}\, j} dy \, s(x|y) \, \varepsilon(y) \, f_{\mathrm{true}}(y)}{(\mu_j / \mu_{\mathrm{tot}})} \, \mu_j \\ &= \sum_{j=1}^{M} R_{ij} \, \mu_j, \end{aligned} \tag{11.8}$$

where the **response matrix** R is given by

$$\begin{aligned} R_{ij} &= \frac{\int_{\mathrm{bin}\, i} dx \int_{\mathrm{bin}\, j} dy \, s(x|y) \, \varepsilon(y) \, f_{\mathrm{true}}(y)}{\int_{\mathrm{bin}\, j} dy \, f_{\mathrm{true}}(y)} \\ &= \frac{P(\text{observed in bin } i \text{ and true value in bin } j)}{P(\text{true value in bin } j)} \\ &= P(\text{observed in bin } i \,|\, \text{true value in bin } j). \end{aligned} \tag{11.9}$$

The response matrix element R_{ij} is thus the conditional probability that an event will be found with measured value x in bin i given that the true value y was in bin j. The effect of off-diagonal elements in R is to smear out any fine structure. A peak in the true histogram concentrated mainly in one bin will be observed over several bins. Two peaks separated by less than several bins will be merged into a single broad peak.

As can be seen from the first line of equation (11.9), the response matrix depends on the p.d.f. $f_{\mathrm{true}}(y)$. This is a priori unknown, however, since the goal of the problem is to determine $f_{\mathrm{true}}(y)$. If the bins of the unfolded histogram are small enough that $s(x|y)$ and $\varepsilon(y)$ are approximately constant over the bin of y, then the direct dependence on $f_{\mathrm{true}}(y)$ cancels out. In the following we will assume that this approximation holds, and that the error in the response matrix due to any uncertainty in $f_{\mathrm{true}}(y)$ can be neglected. In practice, the response matrix will be determined using whatever best approximation of $f_{\mathrm{true}}(y)$ is available prior to carrying out the experiment.

Although $s(x|y)$ and $\varepsilon(y)$ are by construction independent of the probability that a given value y occurs (i.e. independent of $f_{\mathrm{true}}(y)$), they are not in general completely model independent. The variable y may not be the only quantity that influences the probability to obtain a measured value x. For the example of beta decay where y represents the true and x the measured energy of the emitted electron, $s(x|y)$ and $\varepsilon(y)$ will depend in general on the angular distribution of the electrons (some parts of the detector may have better resolution than others), and different models of beta decay might predict different angular distributions.

In the following we will neglect this model dependence and simply assume that the resolution function $s(x|y)$ and efficiency $\varepsilon(y)$, and hence the response matrix R_{ij}, depend only on the measurement apparatus. We will assume in fact that R can be determined with negligible uncertainty both from the standpoint of model dependence as well as from that of detector response. In practice, R is determined either by means of calibration experiments where the true value y is known a priori, or by using a Monte Carlo simulation based on an understanding of the physical processes that take place in the detector. In real problems the model dependence may not be negligible, and the understanding of the detector itself is never perfect. Both must be treated as a possible sources of systematic error.

Note that the response matrix R_{ij} is not in general symmetric (nor even square), with the first index $i = 1, \ldots, N$ denoting the bin of the observed histogram and the second index $j = 1, \ldots, M$ referring to a bin of the true histogram. Summing over the first index and using $\int s(x|y) dx = 1$ gives

$$\begin{aligned}
\sum_{i=1}^{N} R_{ij} &= \sum_{i=1}^{N} \frac{\int_{\text{bin } i} dx \int_{\text{bin } j} dy\, s(x|y)\, \varepsilon(y)\, f_{\text{true}}(y)}{(\mu_j/\mu_{\text{tot}})} \\
&= \frac{\int_{\text{bin } j} dy\, \varepsilon(y)\, f_{\text{true}}(y)}{\int_{\text{bin } j} f_{\text{true}}(y)\, dy} \\
&\equiv \varepsilon_j, \qquad (11.10)
\end{aligned}$$

i.e. one obtains the average value of the efficiency over bin j.

In addition to limited resolution and efficiency, one must also allow for the possibility that the measuring device produces a value when no true event of the type under study occurred, i.e. the measured value was caused by some **background** process. In the case of beta decay, this could be the result of spurious signals in the detector, the presence of radioactive nuclei in the sample other than the type under study, interactions due to particles coming from outside the apparatus such as cosmic rays, etc. Suppose that we have an expectation value β_i for the number of entries observed in bin i which originate from background processes. The relation (11.8) is then modified to be

$$\nu_i = \sum_{j=1}^{M} R_{ij}\, \mu_j + \beta_i. \qquad (11.11)$$

Note that the β_i include the effects of limited resolution and efficiency of the detector. They will usually be determined either from calibration experiments or from a Monte Carlo simulation of both the background processes and the detector response. In the following we will assume that the values β_i are known, although in practice this will only be true to a given accuracy. The uncertainty in the background is thus a source of systematic error in the unfolded result.

To summarize, we have the following vector quantities (referred to also in a general sense as histograms or distributions):

(1) the true histogram (expectation values of true numbers of entries in each bin), $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$,
(2) the normalized true histogram (probabilities), $\mathbf{p} = (p_1, \ldots, p_M) = \boldsymbol{\mu}/\mu_{\text{tot}}$,
(3) the expectation values of the observed numbers of entries, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$,
(4) the actual number of entries observed (the data), $\mathbf{n} = (n_1, \ldots, n_N)$,
(5) efficiencies $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_M)$, and
(6) expected background values $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)$.

It is assumed either that we know the form of the probability distribution for the data $\mathbf{n}$, which will allow us to construct the likelihood function, or that we have the covariance matrix $V_{ij} = \text{cov}[n_i, n_j]$, which can be used to construct a χ^2 function. In addition we have the response matrix R_{ij}, where $i = 1, \ldots, N$ represents the bin of the observed histogram, and $j = 1, \ldots, M$ gives the bin of the true histogram. We will assume that R and $\boldsymbol{\beta}$ are known. The vectors $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, $\boldsymbol{\beta}$ and the matrix R are related by

$$\boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}, \qquad (11.12)$$

where $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ and $\boldsymbol{\beta}$ should be understood as column vectors in matrix equations. The goal is to construct estimators $\hat{\boldsymbol{\mu}}$ for the true histogram, or estimators $\hat{\mathbf{p}}$ for the probabilities.

11.2 Inverting the response matrix

In this section we will examine a seemingly obvious method for constructing estimators for the true histogram $\boldsymbol{\mu}$, which, however, often leads to an unacceptable solution. Consider the case where the number of bins in the true and observed histograms are equal, $M = N$. For now we will assume that the matrix relation $\boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}$ can be inverted to give

$$\boldsymbol{\mu} = R^{-1}(\boldsymbol{\nu} - \boldsymbol{\beta}). \qquad (11.13)$$

An obvious choice for the estimators of $\boldsymbol{\nu}$ is given by the corresponding data values,

$$\hat{\boldsymbol{\nu}} = \mathbf{n}. \qquad (11.14)$$

The estimators for the $\boldsymbol{\mu}$ are then simply

$$\hat{\boldsymbol{\mu}} = R^{-1}(\mathbf{n} - \boldsymbol{\beta}). \qquad (11.15)$$

One can easily show that this is, in fact, the solution obtained from maximizing the log-likelihood function,

$$\log L(\boldsymbol{\mu}) = \sum_{i=1}^{N} \log P(n_i; \nu_i), \qquad (11.16)$$

where $P(n_i; \nu_i)$ is a Poisson or binomial distribution. It is also the least squares solution, where one minimizes

$$\chi^2(\boldsymbol{\mu}) = \sum_{i,j=1}^{N} (\nu_i - n_i)\,(V^{-1})_{ij}\,(\nu_j - n_j). \qquad (11.17)$$

Note that $\log L(\boldsymbol{\mu})$ and $\chi^2(\boldsymbol{\mu})$ can be written as functions of $\boldsymbol{\mu}$ or $\boldsymbol{\nu}$, since the relation $\boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}$ always holds. That is, when differentiating (11.16) or (11.17) with respect to μ_i one uses $\partial\nu_i/\partial\mu_j = R_{ij}$.

Before showing how the estimators constructed in this way can fail, it is interesting to investigate their bias and variance. The expectation value of $\hat{\mu}_j$ is given by

$$\begin{aligned} E[\hat{\mu}_j] &= \sum_{i=1}^{N} (R^{-1})_{ji}\, E[n_i - \beta_i] = \sum_{i=1}^{N} (R^{-1})_{ji}\, (\nu_i - \beta_i) \\ &= \mu_j, \end{aligned} \tag{11.18}$$

so the estimators $\hat{\mu}_j$ are unbiased, since by assumption, $\hat{\nu}_i = n_i$ is unbiased. For the covariance matrix we find

$$\begin{aligned} \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j] &= \sum_{k,l=1}^{N} (R^{-1})_{ik}\, (R^{-1})_{jl}\, \mathrm{cov}[n_k, n_l] \\ &= \sum_{k=1}^{N} (R^{-1})_{ik}\, (R^{-1})_{jk}\, \nu_k, \end{aligned} \tag{11.19}$$

where to obtain the last line we have used the covariance matrix for independent Poisson variables, $\mathrm{cov}[n_k, n_l] = \delta_{kl}\nu_k$.

In the following we will use the notation $V_{ij} = \mathrm{cov}[n_i, n_j]$ for the covariance matrix of the data, and $U_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j]$ for that of the estimators of the true distribution. Equation (11.19) can then be written in matrix notation,

$$U = R^{-1}\, V\, (R^{-1})^T. \tag{11.20}$$

Consider now the example shown in Fig. 11.1. The original true distribution $\boldsymbol{\mu}$ is shown in Fig. 11.1(a), and the expectation values for the observed distribution $\boldsymbol{\nu}$ are shown in the histogram of Fig. 11.1(b).

The histogram $\boldsymbol{\nu}$ has been computed according to $\boldsymbol{\nu} = R\boldsymbol{\mu}$, i.e. the background $\boldsymbol{\beta}$ is taken to be zero. The response matrix R is based on a Gaussian resolution function with a standard deviation equal to 1.5 times the bin width, and the efficiencies ε_i are all taken to be unity. This results in a probability of approximately 26% for an event to remain in the bin where it was created, 21% for the event to migrate one bin, and 16% to migrate two or more bins.

Figure 11.1(c) shows the data $\mathbf{n} = (n_1, \ldots, n_N)$. These have been generated by the Monte Carlo method using Poisson distributions with the mean values ν_i from Fig. 11.1(b). Since the number of entries in each bin ranges from around 10^2 to 10^3, the relative statistical errors (ratio of standard deviation to mean value) for the n_i are in the range from 3 to 10%.

Figure 11.1(d) shows the estimates $\hat{\boldsymbol{\mu}}$ obtained from matrix inversion, equation (11.15). The error bars indicate the standard deviations for each bin. Far from achieving the 3–10% precision that we had for the n_i, the $\hat{\mu}_j$ oscillate wildly from bin to bin, and the error bars are as large as the estimated values themselves. (Notice the increased vertical scale on this plot.) The correlation coefficients for neighboring bins are close to −1.

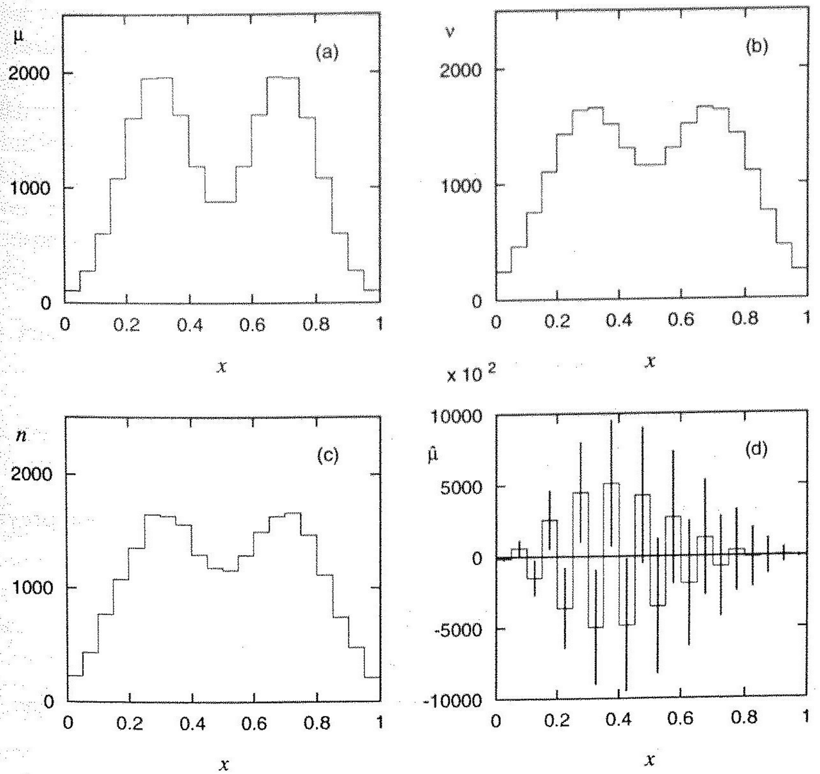

Fig. 11.1 (a) A hypothetical true histogram $\boldsymbol{\mu}$, (b) the histogram of expectation values $\boldsymbol{\nu} = R\boldsymbol{\mu}$, (c) the histogram of observed data $\mathbf{n}$, and (d) the estimators $\hat{\boldsymbol{\mu}}$ obtained from inversion of the response matrix.

The reason for the catastrophic failure stems from the fact that we do not have the expectation values $\boldsymbol{\nu}$; if we did, we could simply compute $\boldsymbol{\mu} = R^{-1}\boldsymbol{\nu}$. Rather, we only have the data $\mathbf{n}$, which are random variables and hence subject to statistical fluctuations. Recall that the effect of the response matrix is to smear out any fine structure. If there had been peaks close together in $\boldsymbol{\mu}$, then although these would be merged together in $\boldsymbol{\nu}$, there would still remain a certain residual fine structure. Upon applying R^{-1} to $\boldsymbol{\nu}$, this remnant of the original structure would be restored. The data $\mathbf{n}$ have indeed statistical fluctuations from bin to bin, and this leads to the same qualitative result as would a residual fine structure in $\boldsymbol{\nu}$. Namely, the unfolded result is given a large amount of fine structure, as is evident in Fig. 11.1(d).

It is interesting to compare the covariance matrix U (11.19) with that given by the RCF inequality (cf. Section 6.6); this gives the smallest possible variance for any choice of estimator. For this we will regard the n_i as independent Poisson variables with mean values ν_i. The log-likelihood function is thus

$$\log L(\boldsymbol{\mu}) = \sum_{i=1}^{N} \log P(n_i; \nu_i) = \sum_{i=1}^{N} \log \left(\frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!} \right). \qquad (11.21)$$

Dropping additive terms that do not depend on $\boldsymbol{\mu}$ gives

$$\log L(\boldsymbol{\mu}) = \sum_{i=1}^{N} (n_i \log \nu_i - \nu_i). \qquad (11.22)$$

One can check that by setting the derivatives of $\log L$ with respect to the components of $\boldsymbol{\mu}$ equal to zero,

$$\frac{\partial \log L}{\partial \mu_k} = \sum_{i=1}^{N} \frac{\partial \log L}{\partial \nu_i} \frac{\partial \nu_i}{\partial \mu_k} = \sum_{i=1}^{N} \left(\frac{n_i}{\nu_i} - 1 \right) R_{ik} = 0, \qquad (11.23)$$

one obtains in fact the same estimators, $\hat{\boldsymbol{\nu}} = \mathbf{n}$, as we have seen previously. Differentiating one more time gives

$$\frac{\partial^2 \log L}{\partial \mu_k \, \partial \mu_l} = - \sum_{i=1}^{N} \frac{n_i \, R_{ik} \, R_{il}}{\nu_i^2}. \qquad (11.24)$$

The RCF bound for the inverse covariance matrix for the case of zero bias (equation (6.19)) is therefore

$$\begin{aligned} (U^{-1})_{kl} &= -E \left[\frac{\partial^2 \log L}{\partial \mu_k \, \partial \mu_l} \right] \\ &= \sum_{i=1}^{N} \frac{E[n_i] \, R_{ik} \, R_{il}}{\nu_i^2} \\ &= \sum_{i=1}^{N} \frac{R_{ik} \, R_{il}}{\nu_i}. \qquad (11.25) \end{aligned}$$

By multiplying both sides of the equation once by U, twice by R^{-1}, and summing over the appropriate indices, one can solve (11.25) for the RCF bound for the covariance matrix,

$$U_{ij} = \sum_{k=1}^{N} (R^{-1})_{ik} \, (R^{-1})_{jk} \, \nu_k. \qquad (11.26)$$

This is the same as the result of the exact calculation (11.19), so we see that the maximum likelihood solution is both unbiased and efficient, i.e. it has the smallest possible variance for an estimator with zero bias. We would obtain the same result using the method of least squares; in that case, unbiased and efficient estimators are guaranteed by the Gauss–Markov theorem.

Although the solution in Fig. 11.1(d) bears little resemblance to the true distribution, it has certain desirable properties. It is simple to construct, has zero bias, and the variance is equal to the RCF bound. In order to be of use, however, the correlations must be taken into account. For example, one can test the compatibility of the estimators $\hat{\boldsymbol{\mu}}$ with a hypothesis $\boldsymbol{\mu}_0$ by constructing a χ^2 statistic,

$$\chi^2 = (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)^T \, U^{-1} \, (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0), \qquad (11.27)$$

which uses the full covariance matrix U of the estimators. This test would be meaningless if the χ^2 were to be computed with only the diagonal elements of U. We should also note that although the variances are extremely large in the example shown here, they would be significantly smaller if the bins are made large compared to the width of the resolution function.

Regardless of its drawbacks, response-matrix inversion indicates some important lessons and provides a starting point for other methods. Since the inverse-matrix solution has zero bias and minimum variance as given by the RCF inequality, any reduction in variance can only be achieved by introducing a bias. The art of unfolding consists of constructing biased estimators $\hat{\boldsymbol{\mu}}$ such that the bias will be small if our prior beliefs, usually some assumptions concerning smoothness, are in fact correct. Roughly speaking, the goal is to find an optimal trade-off between bias and variance, although we will see in Section 11.7 that there is a certain arbitrariness in determining how this optimum is achieved.

The need to incorporate prior knowledge suggests using the Bayesian approach, where the a priori probabilities are combined with the data to yield a posteriori probabilities for the true distribution (cf. Sections 1.2, 6.13). This is a common starting point in the literature on unfolding. It suffers from the difficulty, however, that prior knowledge is often of a complicated or qualitative nature and is thus difficult to express in terms of prior probabilities. The fact that prior beliefs are inherently subjective is not a real disadvantage here; in the classical approach as well there is a certain subjectivity as to how one chooses a biased estimator. In the following we will mainly follow classical statistics, using bias and variance as the criteria by which to judge the quality of a solution, while pointing out the connections with the Bayesian techniques wherever possible.

As a final remark on matrix inversion, we can consider the case where the number of bins M in the unfolded histogram is not equal to the number of measured bins N. For $M > N$, the system of equations (11.12), $\boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}$, is underdetermined, and the solution is not unique. The methods presented in Section 11.4 can be used to select a solution as the estimator $\hat{\boldsymbol{\mu}}$. For $M < N$, (11.12) is overdetermined, and an exact solution does not exist in general. An approximate solution can be constructed using, for example, the methods of maximum likelihood or of least squares, i.e. the problem is equivalent to parameter estimation as discussed in Chapters 5–8. If M is large, then correlations between the estimators can lead to large variances. In such a case it may be desirable to reduce the variances, at the cost of introducing bias, by using one of the regularization methods of Section 11.4.

11.3 The method of correction factors

Consider the case where the bins of the true distribution $\boldsymbol{\mu}$ are taken to be the same as those of the data $\mathbf{n}$. One of the simplest and perhaps most commonly used techniques is to take as the estimator for μ_i

$$\hat{\mu}_i = C_i(n_i - \beta_i), \tag{11.28}$$

where β_i is the expected background and C_i is a multiplicative **correction factor**. The correction factors can be determined using a Monte Carlo program which includes both a model of the process under study as well as a simulation of the measuring apparatus. The factors C_i are determined by running the Monte Carlo program once with and once without the detector simulation, yielding model predictions for the observed and true values of each bin, ν_i^{MC} and μ_i^{MC}. Here ν^{MC} refers to the signal process only, i.e. background is not included. The correction factor is then simply the ratio,

$$C_i = \frac{\mu_i^{\mathrm{MC}}}{\nu_i^{\mathrm{MC}}}. \tag{11.29}$$

For now we will assume that it is possible to generate enough Monte Carlo data so that the statistical errors in the correction factors are negligible. If this is not the case, the uncertainties in the C_i can be incorporated into those of the estimates $\hat{\mu}_i$ by the usual procedure of error propagation.

If the effects of resolution are negligible, then the response matrix is diagonal, i.e. $R_{ij} = \delta_{ij}\varepsilon_j$, and therefore one has

$$\nu_i^{\mathrm{sig}} = \nu_i - \beta_i = \varepsilon_i \mu_i, \tag{11.30}$$

where ν_i^{sig} is the expected number of entries in bin i without background. Thus the correction factors become simply $C_i = 1/\varepsilon_i$, so that $1/C_i$ plays the role of a generalized efficiency. When one has off-diagonal terms in the response matrix, however, the values of $1/C_i$ can be greater than unity. That is, because of migrations between bins, it is possible to find more entries in a given bin than the number of true entries actually created there.

The expectation value of the estimator $\hat{\mu}_i$ is

$$\begin{aligned} E[\hat{\mu}_i] &= C_i\, E[n_i - \beta_i] = C_i(\nu_i - \beta_i) = \frac{\mu_i^{\mathrm{MC}}}{\nu_i^{\mathrm{MC}}}\, \nu_i^{\mathrm{sig}} \\ &= \left(\frac{\mu_i^{\mathrm{MC}}}{\nu_i^{\mathrm{MC}}} - \frac{\mu_i}{\nu_i^{\mathrm{sig}}} \right) \nu_i^{\mathrm{sig}} + \mu_i . \end{aligned} \tag{11.31}$$

The estimator $\hat{\mu}_i$ thus has a bias which is only zero if the ratios $\mu_i/\nu_i^{\mathrm{sig}}$ are the same for the Monte Carlo model and for the real experiment.

The covariance matrix for the estimators is given by

$$\begin{aligned} \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j] &= C_i^2\, \mathrm{cov}[n_i, n_j] \\ &= C_i^2\, \delta_{ij}\, \nu_i . \end{aligned} \tag{11.32}$$

The last line uses the covariance matrix for the case where the n_i are independent Poisson variables with expectation values ν_i. For many practical problems, the C_i are of order unity, and thus the variances of the estimates $\hat{\mu}_i$ are approximately the same as what one would achieve with perfect resolution. In addition, the technique is simple to implement, not even requiring a matrix inversion. The price that one pays is the bias,

$$b_i = \left(\frac{\mu_i^{\mathrm{MC}}}{\nu_i^{\mathrm{MC}}} - \frac{\mu_i}{\nu_i^{\mathrm{sig}}} \right) \nu_i^{\mathrm{sig}}. \tag{11.33}$$

A rough estimate of the systematic uncertainty due to this bias can be obtained by computing the correction factors with different Monte Carlo models. Clearly a better model leads to a smaller bias, and therefore it is often recommended that the estimated distribution $\hat{\boldsymbol{\mu}}$ be used to tune the Monte Carlo, i.e. by adjusting its parameters to improve the agreement between $\boldsymbol{\nu}^{\mathrm{MC}}$ and the background subtracted data $\mathbf{n} - \boldsymbol{\beta}$. One can then iterate the procedure and obtain improved correction factors from the tuned model.

A danger in the method of correction factors is that the bias often pulls the estimates $\hat{\boldsymbol{\mu}}$ towards the model prediction $\boldsymbol{\mu}^{\mathrm{MC}}$. This complicates the task of testing the model, which may have been the purpose of carrying out the measurement in the first place. In such cases one must ensure that the uncertainty in the unfolded result due to the model dependence of the correction factors is taken into account in the estimated systematic errors, and that these are incorporated into any model tests.

11.4 General strategy of regularized unfolding

Although the method of correction factors is simple and widely practiced, it has a number of disadvantages, primarily related to the model dependence of the result. An alternative approach is to impose in some way a measure of smoothness on the estimators for the true histogram $\boldsymbol{\mu}$. This is known as **regularization** of the unfolded distribution.

As a starting point, let us return to the oscillating solution of Section 11.2 obtained from inversion of the response matrix. This estimate for $\boldsymbol{\mu}$ is characterized by a certain maximum value of the log-likelihood function $\log L_{\max}$, or a minimum value of the χ^2. In the following we will usually refer only to the log-likelihood function; the corresponding relations using χ^2 can be obtained by the replacement $\log L = -\chi^2/2$.

One can consider a certain region of $\boldsymbol{\mu}$-space around the maximum likelihood (or least squares) solution as representing acceptable solutions, in the sense that they have an acceptable level of agreement between the predicted expectation values $\boldsymbol{\nu}$ and the data $\mathbf{n}$. The extent of this region can be defined by requiring that $\log L$ stay within some limit of its maximum value. That is, one determines the acceptable region of $\boldsymbol{\mu}$-space by

$$\log L(\boldsymbol{\mu}) \geq \log L_{\max} - \Delta \log L \tag{11.34}$$

or for the case of least squares,

$$\chi^2(\boldsymbol{\mu}) \leq \chi^2_{\min} + \Delta\chi^2 \tag{11.35}$$

for appropriately chosen $\Delta \log L$ or $\Delta\chi^2$. The values of $\Delta \log L$ or $\Delta\chi^2$ will determine the trade-off between bias and variance achieved in the unfolded histogram; we will return to this point in detail in Section 11.7.

In addition to the acceptability of the solution, we need to define a measure of its smoothness by introducing a function $S(\boldsymbol{\mu})$, called the **regularization function**. Several possible forms for $S(\boldsymbol{\mu})$ will be discussed in the next section. The general strategy is to choose the solution with the highest degree of smoothness out of the acceptable solutions determined by the inequalities (11.34) or (11.35).

Maximizing the regularization function $S(\boldsymbol{\mu})$ with the constraint that $\log L(\boldsymbol{\mu})$ remain equal to $\log L_{\max} - \Delta \log L$ is equivalent to maximizing the quantity

$$\alpha\,[\log L(\boldsymbol{\mu}) - (\log L_{\max} - \Delta \log L)] + S(\boldsymbol{\mu}) \tag{11.36}$$

with respect to both $\boldsymbol{\mu}$ and α. Here α is a Lagrange multiplier called the **regularization parameter**, which can be chosen to correspond to a specific value of $\Delta \log L$. For a given α, the solution is thus determined by finding the maximum of a weighted combination of $\log L$ and the $S(\boldsymbol{\mu})$,

$$\Phi(\boldsymbol{\mu}) = \alpha \, \log L(\boldsymbol{\mu}) + S(\boldsymbol{\mu}). \tag{11.37}$$

Setting $\alpha = 0$ leads to the smoothest distribution possible; this ignores completely the data $\mathbf{n}$. A very large α leads to the oscillating solution from inversion of the response matrix, corresponding to having the likelihood function equal to its maximum value.

In order for the prescription of maximizing $\Phi(\boldsymbol{\mu})$ to be in fact equivalent to the general strategy stated above, the surfaces of constant $\log L(\boldsymbol{\mu})$ and $S(\boldsymbol{\mu})$ must be sufficiently well behaved; in the following we will assume this to be the case. In particular, they should not change from convex to concave or have a complicated topology such that multiple local maxima exist.

Recall that we can write $\log L$ and S as functions of $\boldsymbol{\mu}$ or $\boldsymbol{\nu}$, since the relation $\boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}$ always holds. In a similar way, we will always take the relation

$$\hat{\boldsymbol{\nu}} = R\hat{\boldsymbol{\mu}} + \boldsymbol{\beta} \tag{11.38}$$

to define the estimators for $\boldsymbol{\nu}$; knowing these is equivalent to knowing the estimators $\hat{\boldsymbol{\mu}}$. Note, however, that in contrast to the method of Section 11.2, we will no longer have $\hat{\boldsymbol{\nu}} = \mathbf{n}$. It should also be kept in mind that $\mu_{\rm tot} = \sum_j \mu_j$ and $\nu_{\rm tot} = \sum_i \nu_i = \sum_{i,j} R_{ij}\mu_j$ are also functions of $\boldsymbol{\mu}$.

Here we will only consider estimators $\hat{\boldsymbol{\mu}}$ for which the estimated total number of events $\hat{\nu}_{\rm tot}$ is equal to the number actually observed,

$$\hat{\nu}_{\rm tot} = \sum_{i=1}^{N} \hat{\nu}_i = \sum_{i=1}^{N}\sum_{j=1}^{M} R_{ij}\,\hat{\mu}_j + \beta_i = n_{\rm tot}. \tag{11.39}$$

This condition is not in general fulfilled automatically. It can be imposed by modifying equation (11.37) to read

$$\varphi(\boldsymbol{\mu}, \lambda) = \alpha \, \log L(\boldsymbol{\mu}) + S(\boldsymbol{\mu}) + \lambda \left[n_{\rm tot} - \sum_{i=1}^{N} \nu_i \right], \tag{11.40}$$

where λ is a Lagrange multiplier. Setting $\partial\varphi/\partial\lambda = 0$ then leads to $\sum_i \nu_i = n_{\rm tot}$.

As a technical aside, note that it does not matter whether the regularization parameter α is attached to the regularization function $S(\boldsymbol{\mu})$ (as it is in many references) or with the likelihood function. In the particular numerical implementation given in Section 11.9, it is more convenient to associate α with the likelihood.

11.5 Regularization functions

11.5.1 Tikhonov regularization

A commonly used measure of smoothness is the mean value of the square of some derivative of the true distribution. This technique was suggested independently by Phillips [Phi62] and Tikhonov [Tik63, Tik77], and is usually called **Tikhonov regularization**. If we consider the p.d.f. $f_{\rm true}(y)$ before being discretized as a histogram, then the regularization function is

$$S[f_{\rm true}(y)] = -\int \left(\frac{d^k f_{\rm true}(y)}{dy^k} \right)^2 dy, \tag{11.41}$$

where the integration is over all allowed values of y. The minus sign comes from the convention taken here that we maximize φ as defined by (11.40). That is, greater S corresponds to more smoothness. (Equivalently one can of course minimize a combination of regularization and log-likelihood functions with the opposite sign; this convention as well is often encountered in the literature.)

In principle, a linear combination of terms with different derivatives could be used; in practice, one value of k is usually chosen. When $f_{\text{true}}(y)$ is represented as a histogram, the derivatives are replaced by finite differences. For equal bin widths, one can use for $k = 1$ (cf. [Pre92])

$$S(\boldsymbol{\mu}) = -\sum_{i=1}^{M-1} (\mu_i - \mu_{i+1})^2, \tag{11.42}$$

for $k = 2$

$$S(\boldsymbol{\mu}) = -\sum_{i=1}^{M-2} (-\mu_i + 2\mu_{i+1} - \mu_{i+2})^2, \tag{11.43}$$

or for $k = 3$

$$S(\boldsymbol{\mu}) = -\sum_{i=1}^{M-3} (-\mu_i + 3\mu_{i+1} - 3\mu_{i+2} + \mu_{i+3})^2. \tag{11.44}$$

A common choice for the derivative is $k = 2$, so that $S(\boldsymbol{\mu})$ is related to the average curvature.

If the bin widths Δy_i are all equal, then they can be ignored in (11.42)–(11.44). This would only give a constant of proportionality, and can be effectively absorbed into the regularization parameter α. If the Δy_i are not all equal, then this can be included in the finite differences in a straightforward manner. For $k = 2$, for example, one can assume a parabolic form for $f_{\text{true}}(y)$ within each group of three adjacent bins,

$$f_i(y) = a_{0i} + a_{1i}y + a_{2i}y^2. \tag{11.45}$$

There are $M - 2$ such groups, centered around bins $i = 2, \ldots, M - 1$. The coefficients can be determined in each group by setting the integrals of $f_i(y)$ over bins $i - 1$, i and $i + 1$ equal to the corresponding values of μ_{i-1}, μ_i and μ_{i+1}. The second derivative for the group centered around bin i is then $f_i'' = 2a_{2i}$, and the regularization function can thus be taken to be

$$S(\boldsymbol{\mu}) = -\sum_{i=2}^{M-1} f_i''^2 \, \Delta y_i. \tag{11.46}$$

Note that the second derivative cannot be determined in the first and last bins. Here they are not included in the sum (11.46), i.e. they are taken to be zero; alternatively one could set them equal to the values obtained in bins 2 and $M-1$.

For any value of the derivative k and regardless of the bin widths, the functions $S(\boldsymbol{\mu})$ given above can be expressed as

$$S(\boldsymbol{\mu}) = -\sum_{i,j=1}^{M} G_{ij}\,\mu_i\,\mu_j = -\boldsymbol{\mu}^T G\,\boldsymbol{\mu}, \tag{11.47}$$

where G is a symmetric matrix of constants. For $k = 2$ with equal bin widths (11.43), for example, G is given by

$$\begin{array}{l} \left.\begin{array}{l} G_{ii} = 6 \\ G_{i,i\pm1} = G_{i\pm1,i} = -4 \\ G_{i,i\pm2} = G_{i\pm2,i} = 1 \end{array}\right\} \quad 3 \le i \le M-2, \\ G_{11} = G_{MM} = 1, \\ G_{22} = G_{M-1,M-1} = 5, \\ G_{12} = G_{21} = G_{M,M-1} = G_{M-1,M} = -2, \end{array} \tag{11.48}$$

with all other G_{ij} equal to zero.

In order to obtain the estimators and their covariance matrix (Section 11.6), we will need the first and second derivatives of S. These are

$$\frac{\partial S}{\partial \mu_i} = -2\sum_{j=1}^{M} G_{ij}\,\mu_j \tag{11.49}$$

and

$$\frac{\partial^2 S}{\partial \mu_i \partial \mu_j} = -2\,G_{ij}. \tag{11.50}$$

Tikhonov regularization using $k = 2$ has been widely applied in particle physics for the unfolding of structure functions (distributions of kinematic variables in lepton–nucleon scattering). Further descriptions can be found in [Blo85, Höc96, Roe92, Zec95].

11.5.2 Regularization functions based on entropy

Another commonly used regularization function is based on the **entropy** H of a probability distribution $\mathbf{p} = (p_1, \ldots, p_M)$, defined as [Sha48]

$$H = -\sum_{i=1}^{M} p_i \log p_i. \tag{11.51}$$

The idea here is to interpret the entropy as a measure of the smoothness of a histogram $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$, and to use

$$S(\boldsymbol{\mu}) = H(\boldsymbol{\mu}) = -\sum_{i=1}^{M} \frac{\mu_i}{\mu_{\text{tot}}} \log \frac{\mu_i}{\mu_{\text{tot}}} \tag{11.52}$$

as a regularization function. Estimators based on (11.52) are said to be constructed according to the **principle of maximum entropy** or **MaxEnt**. To see how entropy is related to smoothness, consider the number of ways in which a particular histogram $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$ can be constructed out of $\mu_{\rm tot}$ entries (here the values μ_j are integers). This is given by

$$\Omega(\boldsymbol{\mu}) = \frac{\mu_{\rm tot}!}{\mu_1!\,\mu_2!\,\ldots\,\mu_M!}. \tag{11.53}$$

(Recall that the same factor appears in the multinomial distribution (2.6).) By taking the logarithm of (11.53) and using Stirling's approximation, $\log n! \approx n(\log n - 1)$, valid for large n, one obtains

$$\begin{aligned}\log\Omega &\approx \mu_{\rm tot}(\log\mu_{\rm tot} - 1) - \sum_{i=1}^{M} \mu_i(\log\mu_i - 1) \\ &= -\sum_{i=1}^{M} \mu_i \log\frac{\mu_i}{\mu_{\rm tot}} \\ &= \mu_{\rm tot}\, S(\boldsymbol{\mu}).\end{aligned} \tag{11.54}$$

We will use equation (11.54) to generalize $\log\Omega$ to the case where the μ_i are not integers.

If all of the events are concentrated in a single bin, i.e. the histogram has the minimum degree of smoothness, then there is only one way of arranging them, and hence the entropy is also a minimum. At the other extreme, one can show that the entropy is maximum for the case where all μ_i are equal, i.e. the histogram corresponds to a uniform distribution. To maximize H with the constraint $\sum_i p_i = 1$, a Lagrange multiplier can be used.

For later reference, we list here the first and second derivatives of the entropy-based $S(\boldsymbol{\mu})$:

$$\frac{\partial S}{\partial \mu_i} = -\frac{1}{\mu_{\rm tot}} \log\frac{\mu_i}{\mu_{\rm tot}} - \frac{S(\boldsymbol{\mu})}{\mu_{\rm tot}} \tag{11.55}$$

and

$$\frac{\partial^2 S}{\partial\mu_i \partial\mu_j} = \frac{1}{\mu_{\rm tot}^2}\left[1 - \frac{\delta_{ij}\,\mu_{\rm tot}}{\mu_i} + \log\left(\frac{\mu_i\mu_j}{\mu_{\rm tot}^2}\right) + 2S(\boldsymbol{\mu})\right]. \tag{11.56}$$

11.5.3 Bayesian motivation for the use of entropy

In much of the literature on unfolding problems, the principle of maximum entropy is developed in the framework of Bayesian statistics. (See, for example, [Siv96, Jay86, Pre92].) This approach to unfolding runs into difficulties, however, as we will see below. It is nevertheless interesting to compare Bayesian MaxEnt with the classical methods of the previous section.

In the Bayesian approach, the values $\boldsymbol{\mu}$ are treated as random variables in the sense of subjective probability (cf. Section 1.2), and the joint probability density $f(\boldsymbol{\mu}|\mathbf{n})$ represents the degree of belief that the true histogram is given by $\boldsymbol{\mu}$. To update our knowledge about $\boldsymbol{\mu}$ in light of the data $\mathbf{n}$, we use Bayes' theorem,

$$f(\boldsymbol{\mu}|\mathbf{n}) \propto L(\mathbf{n}|\boldsymbol{\mu})\,\pi(\boldsymbol{\mu}), \tag{11.57}$$

where $L(\mathbf{n}|\boldsymbol{\mu})$ is the likelihood function (the conditional probability for the data $\mathbf{n}$ given $\boldsymbol{\mu}$) multiplied by the prior density $\pi(\boldsymbol{\mu})$. The prior density represents our knowledge about $\boldsymbol{\mu}$ before seeing the data $\mathbf{n}$.

Here we will regard the total number of events $\mu_{\rm tot}$ as an integer. This is in contrast to the classical approach, where $\mu_{\rm tot}$ represents an expectation value of an integer random variable, and thus is not necessarily an integer itself. Suppose we have no prior knowledge about how these $\mu_{\rm tot}$ entries are distributed in the histogram. One can then argue that by symmetry, each of the possible ways of placing $\mu_{\rm tot}$ entries into M bins is equally likely. The probability for a certain histogram $(\mu_1, \ldots, \mu_M)$ therefore should be, in the absence of any other prior information, proportional to the number of ways in which it can be made; this is just the number Ω given by equation (11.53). The total number of ways of distributing the entries $\Omega(\boldsymbol{\mu})$ is thus interpreted as the prior probability $\pi(\boldsymbol{\mu})$,

$$\begin{aligned}\pi(\boldsymbol{\mu}) &= \Omega(\boldsymbol{\mu}) = \frac{\mu_{\rm tot}!}{\mu_1!\,\mu_2!\,\ldots\,\mu_M!} \\ &= \exp(\mu_{\rm tot}\,H),\end{aligned} \tag{11.58}$$

where H is the entropy given by equation (11.51).

From the strict Bayesian standpoint, the job is finished when we have determined $f(\boldsymbol{\mu}|\mathbf{n})$. It is not practical to report $f(\boldsymbol{\mu}|\mathbf{n})$ completely, however, since this is a function of as many variables as there are bins M in the unfolded distribution. Therefore some way of summarizing it must be found; to do this one typically selects a single vector $\hat{\boldsymbol{\mu}}$ as the Bayesian estimator. The usual choice is the $\boldsymbol{\mu}$ for which the probability $f(\boldsymbol{\mu}|\mathbf{n})$, or equivalently its logarithm, is a maximum. According to equation (11.57), this is determined by maximizing

$$\begin{aligned}\log f(\boldsymbol{\mu}|\mathbf{n}) &\propto \log L(\boldsymbol{\mu}|\mathbf{n}) + \log\pi(\boldsymbol{\mu}) \\ &= \log L(\boldsymbol{\mu}|\mathbf{n}) + \mu_{\rm tot}H(\boldsymbol{\mu}) \\ &= \log L(\boldsymbol{\mu}|\mathbf{n}) + \mu_{\rm tot}H(\boldsymbol{\mu}).\end{aligned} \tag{11.59}$$

The Bayesian prescription thus corresponds to using a regularization function

$$S(\boldsymbol{\mu}) = \mu_{\rm tot}H(\boldsymbol{\mu}) = -\sum_{i=1}^{M} \mu_i \log\frac{\mu_i}{\mu_{\rm tot}}. \tag{11.60}$$

Furthermore, the regularization parameter α is no longer an arbitrary factor but is set equal to 1. If all of the efficiencies ε_i are equal, then the requirement $\nu_{\rm tot} = n_{\rm tot}$ also implies that $\mu_{\rm tot}$ is constant. This is then equivalent to using the previous regularization function $S(\boldsymbol{\mu}) = H$ with $\alpha = 1/\mu_{\rm tot}$.

If the efficiencies are not all equal, however, then constant $\nu_{\rm tot}$ does not imply constant $\mu_{\rm tot}$, and as a result, the distribution of maximum $S(\boldsymbol{\mu}) = \mu_{\rm tot}H(\boldsymbol{\mu})$ is no longer uniform. This is because S can increase simply by increasing $\mu_{\rm tot}$, and thus in the distribution of maximum S, bins with low efficiency are enhanced. In this case, then, using H and $\mu_{\rm tot}H$ as regularization functions will lead to somewhat different results, although the difference is in practice not great if the efficiencies are of the same order of magnitude. In any event, $S = H$ is easier to justify as a measure of smoothness, since the distribution of maximum H is always uniform.

We will see in Section 11.9 that the Bayesian estimator (11.59) gives too much weight to the entropy term (see Fig. 11.3(a) and [Ski86]). From the classical point of view one would say that it does not represent a good trade-off between bias and variance, having an unreasonably large bias. One can modify the Bayesian interpretation by replacing $\mu_{\rm tot}$ in (11.59) by an effective number of events $\mu_{\rm eff}$, which can be adjusted to be smaller than $\mu_{\rm tot}$. The estimator is then given by the maximum of

$$\log L(\boldsymbol{\mu}|\mathbf{n}) + \mu_{\rm eff}H(\boldsymbol{\mu}). \tag{11.61}$$

This is equivalent to using $S(\boldsymbol{\mu}) = H(\boldsymbol{\mu})$ as before, and the parameter $\mu_{\rm eff}$ plays the role of the regularization parameter.

The problem with the original Bayesian solution stems from our use of $\Omega(\boldsymbol{\mu})$ as the prior density. From either the Bayesian or classical points of view, the quantities $\mathbf{p} = \boldsymbol{\mu}/\mu_{\rm tot}$ are given by some set of unknown, constant numbers, e.g. the electron energy distribution of specific type of beta decay. In either case, our prior knowledge about the *form* of the distribution (i.e. about $\mathbf{p}$, not $\boldsymbol{\mu}$) should be independent of the number of observations in the data sample that we obtain. This points to a fundamental problem in using $\pi(\boldsymbol{\mu}) = \Omega(\boldsymbol{\mu})$, since this becomes increasingly concentrated about a uniform distribution (i.e. all p_i equal) as $\mu_{\rm tot}$ increases.

It is often the case that we have indeed some prior beliefs about the form of the distribution $\mathbf{p}$, but that these are difficult to quantify. We could say, for example, that distributions with large amounts of structure are a priori unlikely, since it may be difficult to imagine a physical theory predicting something with lots of peaks. On the other hand, a completely flat distribution may not seem very physical either, so $\Omega(\boldsymbol{\mu})$ does not really reflect our prior beliefs. Because of these difficulties with the interpretation of $\Omega(\boldsymbol{\mu})$ as a prior p.d.f., we will stay with the classical approach here, and simply regard the entropy as one of the possible regularization functions.

11.5.4 Regularization function based on cross-entropy

Recall that the distribution of maximum entropy is flat, and thus the bias introduced into the estimators $\hat{\boldsymbol{\mu}}$ will tend to pull the result towards a more uniform distribution. Suppose we know a distribution $\mathbf{q} = (q_1, \ldots, q_M)$ that we regard as the most likely a priori shape for the true distribution $\mathbf{p} = \boldsymbol{\mu}/\mu_{\rm tot}$. We will call $\mathbf{q}$ the **reference distribution**. Suppose that we do not know how to quantify our degree of belief in $\mathbf{q}$, however, and hence we do not have a prior density $\pi(\boldsymbol{\mu})$ for use with Bayes' theorem. That is, $\mathbf{q}$ represents the normalized histogram $\boldsymbol{\mu}/\mu_{\rm tot}$ for which the prior density is a maximum, but it does not specify the entire prior density.

In this case, the regularization function can be taken as

$$S(\boldsymbol{\mu}) = K(\mathbf{p};\mathbf{q}), \tag{11.62}$$

where $K(\mathbf{p};\mathbf{q})$ is called the **cross-entropy** [Kul64] or **Shannon–Jaynes entropy** [Jay68], defined as

$$K(\mathbf{p};\mathbf{q}) = -\sum_{i=1}^{M} p_i \log \frac{p_i}{M q_i}. \tag{11.63}$$

The cross-entropy is often defined without the factor of M, and also without the minus sign, in which case the principle of maximum entropy becomes the principle of minimum cross-entropy. We will keep the minus sign so as to maintain the similarity between $K(\mathbf{p};\mathbf{q})$ and the Shannon entropy $H(\mathbf{p})$ (11.51). Note that $K(\mathbf{p};\mathbf{q}) = H(\mathbf{p})$ when the reference distribution is uniform, i.e. $q_i = 1/M$ for all i.

One can easily show that the cross-entropy $K(\mathbf{p};\mathbf{q})$ is a maximum when the probabilities $\mathbf{p}$ are equal to those of the reference distribution $\mathbf{q}$. The effect of using the regularization function (11.62) is that the bias of the estimators $\hat{\boldsymbol{\mu}}$ will be zero (or small) if the true distribution is equal (or close) to the reference distribution.

11.6 Variance and bias of the estimators

The estimators $\hat{\boldsymbol{\mu}}$ are functions of the data $\mathbf{n}$, and are hence themselves random variables. In order to obtain the covariance matrix $U_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j]$, we can calculate an approximate expression for $\hat{\boldsymbol{\mu}}$ as a function of $\mathbf{n}$, and then use the error propagation formula (1.54) to relate U to the covariance matrix for the data, $V_{ij} = \mathrm{cov}[n_i, n_j]$.

The estimators $\hat{\boldsymbol{\mu}}$ are found by maximizing the function $\varphi(\boldsymbol{\mu}, \lambda)$ given by (11.40), which uses a given log-likelihood or χ^2 function and some form of the regularization function $S(\boldsymbol{\mu})$ (Tikhonov, entropy, etc.). The estimators $\hat{\boldsymbol{\mu}}$ and the Lagrange multiplier λ are thus solutions to the system of $M + 1$ equations

$$F_i(\boldsymbol{\mu}, \lambda, \mathbf{n}) = 0, \quad i = 1, \ldots, M+1, \tag{11.64}$$

where

$$F_i(\boldsymbol{\mu}, \lambda, \mathbf{n}) = \begin{cases} \frac{\partial \varphi}{\partial \mu_i} & i = 1, \ldots, M, \\ \frac{\partial \varphi}{\partial \lambda} & i = M+1. \end{cases} \qquad (11.65)$$

Suppose the data actually obtained are given by the vector $\tilde{\mathbf{n}}$, the corresponding estimates are $\tilde{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}(\tilde{\mathbf{n}})$, and the Lagrange multiplier λ has the value $\tilde{\lambda}$. We would like to know how $\hat{\boldsymbol{\mu}}$ and λ would change if the data were given by some different values $\mathbf{n}$. Expanding the functions $F_i(\boldsymbol{\mu}, \lambda, \mathbf{n})$ to first order in a Taylor series about the values $\tilde{\boldsymbol{\mu}}$, $\tilde{\lambda}$ and $\tilde{\mathbf{n}}$ gives

$$\begin{aligned} F_i(\boldsymbol{\mu}, \lambda, \mathbf{n}) \quad &\approx \quad F_i(\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\mathbf{n}}) + \sum_{j=1}^{M} \left[\frac{\partial F_i}{\partial \mu_j} \right]_{\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\mathbf{n}}} (\mu_j - \tilde{\mu}_j) \\ &+ \left[\frac{\partial F_i}{\partial \lambda} \right]_{\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\mathbf{n}}} (\lambda - \tilde{\lambda}) + \sum_{j=1}^{N} \left[\frac{\partial F_i}{\partial n_j} \right]_{\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\mathbf{n}}} (n_j - \tilde{n}_j). \quad (11.66) \end{aligned}$$

The first term $F_i(\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\mathbf{n}})$ as well as the entire expression $F_i(\boldsymbol{\mu}, \lambda, \mathbf{n})$ are both equal to zero, since both sets of arguments should represent solutions. Solving equation (11.66) for $\boldsymbol{\mu}$ gives

$$\hat{\boldsymbol{\mu}}(\mathbf{n}) \approx \tilde{\boldsymbol{\mu}} \; - \; A^{-1} B (\mathbf{n} - \tilde{\mathbf{n}}), \qquad (11.67)$$

where the $M+1$ component of $\boldsymbol{\mu}$ refers to the Lagrange multiplier λ. The symmetric $(M+1) \times (M+1)$ matrix A is given by

$$A_{ij} = \begin{cases} \frac{\partial^2 \varphi}{\partial \mu_i \partial \mu_j}, & i, j = 1, \ldots, M, \\ \frac{\partial^2 \varphi}{\partial \mu_i \partial \lambda} = -1, & i = 1, \ldots, M, j = M+1, \\ \frac{\partial^2 \varphi}{\partial \lambda^2} = 0, & i = M+1, j = M+1, \end{cases} \qquad (11.68)$$

and the $(M+1) \times N$ matrix B is

$$B_{ij} = \begin{cases} \frac{\partial^2 \varphi}{\partial \mu_i \partial n_j}, & i = 1, \ldots, M, j = 1, \ldots, N, \\ \frac{\partial^2 \varphi}{\partial \lambda \partial n_j} = 1, & i = M+1, j = 1, \ldots, N. \end{cases} \qquad (11.69)$$

By using the error propagation formula (1.54), the covariance matrix for the estimators $U_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j]$ is obtained from the covariance matrix for the data $V_{ij} = \mathrm{cov}[n_i, n_j]$ by

$$\mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j] = \sum_{k,l=1}^{N} \frac{\partial \hat{\mu}_i}{\partial n_k} \frac{\partial \hat{\mu}_j}{\partial n_l} \mathrm{cov}[n_k, n_l]. \qquad (11.70)$$

The derivatives in (11.70) can be computed using (11.67) to be

$$\frac{\partial \hat{\mu}_i}{\partial n_k} = -(A^{-1} B)_{ik} \equiv C_{ik}, \qquad (11.71)$$

where the matrices A and B are given by equations (11.68) and (11.69). What we will use here is not the entire matrix C, but rather only the $M \times N$ submatrix, excluding the row $i = M+1$, which refers to the Lagrange multiplier λ. The final expression for the covariance matrix U can thus be expressed in the more compact form,

$$U = C V C^T. \qquad (11.72)$$

The derivatives in (11.68) and (11.69) depend on the choice of regularization function and on the particular log-likelihood function used to define $\varphi(\boldsymbol{\mu}, \lambda)$ (11.40), e.g. Poisson, Gaussian ($\log L = -\chi^2/2$), etc. In the case, for example, where the data are treated as independent Poisson variables with covariance matrix $V_{ij} = \delta_{ij} \nu_i$, and where the entropy-based regularization function (11.54) is used, one has

$$\begin{aligned} \frac{\partial^2 \varphi}{\partial \mu_i \partial \mu_j} \quad &= \quad -\alpha \sum_{k=1}^{N} R_{ki} R_{kj} \frac{n_k}{\nu_k^2} \\ &+ \quad \frac{1}{\mu_{\mathrm{tot}}^2} \left[1 - \frac{\delta_{ij}\, \mu_{\mathrm{tot}}}{\mu_i} + \log\left(\frac{\mu_i \mu_j}{\mu_{\mathrm{tot}}^2} \right) + 2S(\boldsymbol{\mu}) \right] \qquad (11.73) \end{aligned}$$

and

$$\frac{\partial^2 \varphi}{\partial \mu_i \partial n_j} = \frac{\alpha R_{ji}}{\nu_j}. \qquad (11.74)$$

The matrices A and B (and hence C) can be determined by evaluating the derivatives (11.73) and (11.74) with the estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ obtained in the actual experiment. Table 11.1 summarizes the necessary ingredients for Poisson and Gaussian log-likelihood functions. Note that for the Gaussian case, i.e. for the method of least squares, the quantities always refer to $\log L = -\frac{1}{2}\chi^2$, and not to χ^2 itself. The derivatives of Tikhonov and entropy-based regularization functions are given in Sections 11.5.1 and 11.5.2.

In order to determine the biases $b_i = E[\hat{\mu}_i] - \mu_i$, we can compute the expectation values $E[\hat{\mu}_i]$ by means of the approximate relation (11.67),

$$b_i = E[\hat{\mu}_i] - \mu_i \approx \tilde{\mu}_i + \sum_{j=1}^{N} C_{ij} (\nu_j - \tilde{n}_j) \; - \; \mu_i. \qquad (11.75)$$

This can be estimated by substituting the estimator from equation (11.67) for μ_i and replacing ν_j by its corresponding estimator $\hat{\nu}_j = \sum_k R_{jk} \hat{\mu}_k$, which yields

$$\hat{b}_i = \sum_{j=1}^{N} C_{ij}(\hat{\nu}_j - n_j) = \sum_{j=1}^{N} \frac{\partial \hat{\mu}_i}{\partial n_j}(\hat{\nu}_j - n_j). \tag{11.76}$$

The approximations used to construct $\hat{b}_i$ are valid for small $(\hat{\nu}_j - n_j)$, or equivalently, large values of the regularization parameter α. For small α, the matrix C in fact goes to zero, since the estimators $\hat{\mu}_i$ are then decoupled from the measurements n_j, cf. equation (11.71). In this case, however, the bias is actually at its largest. But since we will only use $\hat{b}_i$ and its variance in order to determine the regularization parameter α, the approximation is sufficient for our purposes.

By error propagation (neglecting the variance of the matrix C), one obtains the covariance matrix W for the $\hat{b}_i$,

$$W_{ij} = \mathrm{cov}[\hat{b}_i, \hat{b}_j] = \sum_{k,l=1}^{N} C_{ik}\, C_{jl}\, \mathrm{cov}[\,(\hat{\nu}_k - n_k),\, (\hat{\nu}_l - n_l)\,]. \tag{11.77}$$

This can be computed by using $\hat{\nu}_k = \sum_m R_{km}\hat{\mu}_m$ to relate the covariance matrix $\mathrm{cov}[\hat{\nu}_k, \hat{\nu}_l]$ to that of the estimators for the true distribution, $U_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j]$, which is in turn related by equation (11.72) to the covariance matrix of the data by $U = CVC^T$. Putting this all together gives

$$\begin{aligned} W &= (CRC - C)\, V\, (CRC - C)^T \\ &= (CR - I)\, U\, (CR - I)^T, \end{aligned} \tag{11.78}$$

where I is the $M \times M$ unit matrix. The variances $V[\hat{b}_i] = W_{ii}$ can be used to tell whether the estimated biases are significantly different from zero; this in turn can be employed as a criterion to determine the regularization parameter.

Table 11.1 Log-likelihood functions and their derivatives for Poisson and Gaussian random variables.

	Poisson	Gaussian (least squares)
$\log L$	$\sum_i (n_i \log \nu_i - \nu_i)$	$-\frac{1}{2}\sum_{i,j}(\nu_i - n_i)(V^{-1})_{ij}(\nu_j - n_j)$
$\frac{\partial \log L}{\partial \mu_i}$	$\sum_j \left(\frac{n_j}{\nu_j} - 1\right) R_{ji}$	$-\sum_{j,k} R_{ji}\,(V^{-1})_{jk}\,(\nu_k - n_k)$
$\frac{\partial^2 \log L}{\partial \mu_i \partial \mu_j}$	$-\sum_k \frac{n_k R_{ki} R_{kj}}{\nu_k^2}$	$-(R^T\, V^{-1}\, R)_{ij}$
$\frac{\partial^2 \log L}{\partial n_i \partial \mu_j}$	$\frac{R_{ij}}{\nu_i}$	$(V^{-1}\, R)_{ij}$

Before proceeding to the question of the regularization parameter, however, it is important to note that the biases are in general nonzero for all regularized unfolding methods, in the sense that they are given by some functions, not everywhere zero, of the true distribution. Their numerical values, however, can in fact be zero for particular values of $\boldsymbol{\mu}$. A guiding principle in unfolding is to choose a method such that the bias will be zero (or small) if $\boldsymbol{\mu}$ has certain properties believed a priori to be true. For example, if the true distribution is uniform, then estimates based on Tikhonov regularization with $k = 1$ (11.42) will have zero bias; if the true distribution is linear, then $k = 2$ (11.43) gives zero bias, etc. If the true distribution is equal to a reference distribution $\mathbf{q}$, then unfolding using the cross-entropy (11.63) will yield zero bias.

11.7 Choice of the regularization parameter

The choice of the regularization parameter α, or equivalently the choice of $\Delta \log L$ (or $\Delta\chi^2$), determines the trade-off between the bias and variance of the estimators $\hat{\boldsymbol{\mu}}$. By setting α very large, the solution is dominated by the likelihood function, and one has $\log L = \log L_{\max}$ (or with least squares, $\chi^2 = \chi^2_{\min}$) and correspondingly very large variances. At the other extreme, $\alpha \to 0$ puts all of the weight on the regularization function and leads to a perfectly smooth solution.

Various definitions of an optimal trade-off are possible; these can incorporate the estimates for the covariance matrix $U_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j]$, the biases $\hat{b}_i$, and the covariance matrix of their estimators, $W_{ij} = \mathrm{cov}[\hat{b}_i, \hat{b}_j]$. Here U and W will refer to the estimated values, $\widehat{U}$ and $\widehat{W}$; the hats will not be written explicitly.

One possible measure of the goodness of the final result is the mean squared error, cf. equation (5.5), averaged over all bins,

$$\mathrm{MSE} = \frac{1}{M}\sum_{i=1}^{M}(U_{ii} + \hat{b}_i^2). \tag{11.79}$$

The method of determining α so as to obtain a particular value of the MSE will depend on the numerical implementation. Often it is simply a matter of trying a value α, maximizing $\varphi(\boldsymbol{\mu}, \lambda)$, and iterating the procedure until the desired solution is found.

One could argue, however, that the contribution to the mean squared error should be different for different bins depending on how accurately they are measured. Since the variance of a Poisson variable with mean value μ_i is equal to μ_i, one can define a weighted MSE,

$$\mathrm{MSE}' = \frac{1}{M}\sum_{i=1}^{M}\frac{U_{ii} + \hat{b}_i^2}{\hat{\mu}_i}, \tag{11.80}$$

in analogy with the χ^2 used in the method of least squares. For Poisson distributed data, the quantity MSE′ represents the mean squared increase in the errors due to limited resolution. It is thus reasonable to require that this quantity be small.

A popular choice for the regularization parameter is based on the idea that, on average, each bin should contribute approximately one unit to the χ^2, i.e. α is determined such that $\chi^2 = N$. This can be generalized to the log-likelihood case as $\Delta \log L = \log L_{\max} - \log L = N/2$, since for Gaussian distributed $\mathbf{n}$ one has $\log L = -\chi^2/2$.

Naively one might expect that an increase in the χ^2 of one unit would set the appropriate level of discrepancy between the data $\mathbf{n}$ and the estimates $\hat{\boldsymbol{\nu}}$. This typically leads, however, to solutions with unreasonably large variance. The problem can be traced to the fact that the estimator $\hat{\nu}_i$ receives contributions not only from n_i but also from neighboring bins as well. The coupling of the estimators $\hat{\nu}_i$ to the measurements n_j can be expressed by the matrix

$$\frac{\partial \hat{\nu}_i}{\partial n_j} = \frac{\partial}{\partial n_j} \sum_{k=1}^{M} R_{ik}\, \hat{\mu}_k = (RC)_{ij}. \tag{11.81}$$

A modification of the criterion $\Delta\chi^2 = 1$ has been suggested in [Sch94] which incorporates this idea. It is based on an increase of one unit in an effective χ^2,

$$\Delta\chi^2_{\text{eff}} = (\hat{\boldsymbol{\nu}} - \mathbf{n})^T\, RC\,V^{-1}(RC)^T\,(\hat{\boldsymbol{\nu}} - \mathbf{n}) = 1, \tag{11.82}$$

where the matrix RC effectively takes into account the reduced coupling between the estimators $\hat{\nu}_i$ and the data n_i.

Alternatively, one can look at the estimates of the biases and their variances. If the biases are significantly different from zero, then it is reasonable to subtract them. This is equivalent to going to a smaller value of $\Delta \log L$. As a measure of the deviation of the biases from zero, one can construct the weighted sum of squares,

$$\chi^2_b = \sum_{i=1}^{M} \frac{\hat{b}_i^2}{W_{ii}}. \tag{11.83}$$

The strategy is thus to reduce $\Delta \log L$ (i.e. increase α) until χ^2_b is equal to a sufficiently small value, such as the number of bins M. At this point the standard deviations of the biases are approximately equal to the biases themselves, and therefore any further bias reduction would introduce as much error as it removes.

The bias squared, the variance, and their sum, the mean squared error, are shown as a function of $\Delta \log L$ in Fig. 11.2. These are based on the example from Fig. 11.1, there unfolded by inverting the response matrix, and here treated using (a) maximum entropy and (b) Tikhonov regularization. The increase in the estimated bias for low $\Delta \log L$ reflects the variance of the estimators $\hat{b}_i$; the true bias decreases to zero as $\Delta \log L$ goes to zero. The arrows indicate solutions based on the various criteria introduced above; these are discussed further in the next section.

Further criteria for setting the regularization parameter have been proposed based on singular value analysis [Höc96], or using a procedure known as cross-validation [Wah79]. Unfortunately, the choice of α is still a somewhat open question. In practice, the final estimates are relatively stable as the value of $\Delta \log L$ decreases, until a certain point where the variances suddenly shoot up (see Fig. 11.2). The onset of the rapid increase in the variances indicates roughly the natural choice for setting α.

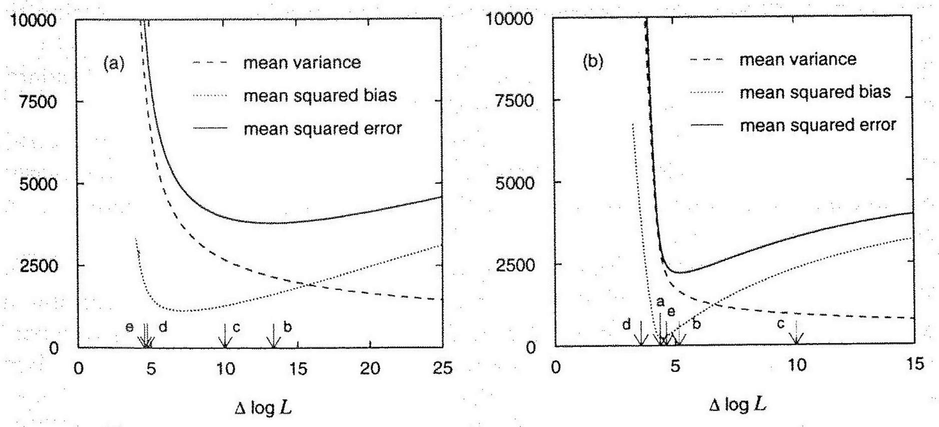

Fig. 11.2 The estimated mean variance, mean squared bias, and their sum, the mean squared error, as a function of $\Delta \log L$ for (a) MaxEnt and (b) Tikhonov regularization ($k = 2$). The arrows indicate the solutions from Figs 11.3 and 11.4: (b) is minimum MSE, (c) is $\Delta \log L = N/2$, (d) is $\Delta\chi^2_{\text{eff}} = 1$, and (e) is $\chi^2 b = M$. For the MaxEnt case, the Bayesian solution $\Delta \log L = 970$ is not shown. For Tikhonov regularization, (a) gives the solution for minimum weighted MSE.

11.8 Examples of unfolding

Figures 11.3 and 11.4 show examples based on maximum entropy and Tikhonov regularization, respectively. The distributions $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ and $\mathbf{n}$ are the same as seen previously in Figs 11.1(a)–(c), having $N = M = 20$ bins, all efficiencies ε equal to unity, and backgrounds $\boldsymbol{\beta}$ equal to zero. The estimators $\hat{\boldsymbol{\mu}}$ are found by maximizing the function $\varphi(\boldsymbol{\mu}, \lambda)$ (11.40), here constructed with a log-likelihood function based on independent Poisson distributions for the data. On the left, the original 'true' histograms $\boldsymbol{\mu}$ are shown along with the unfolded solutions $\hat{\mu}_i$ and error bars $\sqrt{U_{ii}}$ corresponding to a given value of the regularization parameter α, or equivalently to a given $\Delta \log L$. On the right are the corresponding estimates of the biases $\hat{b}_i$ with their standard deviations $\sqrt{W_{ii}}$. These should not be confused with the true residuals $\hat{\mu}_i - \mu_i$, which one could not construct without knowledge of the true histogram $\boldsymbol{\mu}$. The estimates $\hat{b}_i$, on the other hand, are determined from the data.

Consider first Fig. 11.3, with the entropy-based regularization function $S(\boldsymbol{\mu}) = H(\boldsymbol{\mu})$. Figure 11.3(a) corresponds to $\alpha = 1/\mu_{\text{tot}}$, i.e. the Bayesian prescription (11.59), and gives $\Delta \log L = 970$. We show this choice simply to illustrate that the prior density $\pi(\boldsymbol{\mu}) = \Omega(\boldsymbol{\mu})$ does not lead to a reasonable solution. Although the standard deviations $\sqrt{U_{ii}}$ are very small, there is a large bias. The estimates $\hat{b}_i$ shown on the right are indeed large, and from their error bars one can see that they are significantly different from zero. Note that the estimated biases here are not, in fact, in very good agreement with the true residuals $\hat{\mu}_i - \mu_i$, owing to the approximations made in constructing the estimators $\hat{b}_i$, cf. equation (11.67). The approximations become better as $\Delta \log L$ is decreased, until the standard deviations $\sqrt{W_{ii}}$ become comparable to the biases themselves.

Figure 11.3(b) shows the result based on minimum mean squared error (11.79). This corresponds to $\Delta \log L = 13.3$ and $\chi_b^2 = 154$. Although the estimated biases are much smaller than for $\alpha = 1/\mu_{\rm tot}$, they are still significantly different from zero.

The solution corresponding to $\Delta \log L = N/2 = 10$ is shown in Fig. 11.3(c). Here the biases are somewhat smaller than in the result based on minimum MSE, but are still significantly different from zero, giving $\chi_b^2 = 87$. In this particular example, requiring minimum weighted mean squared error (11.80) gives $\Delta \log L = 10.5$, and is thus similar to the result from $\Delta \log L = N/2 = 10$.

The results corresponding to $\Delta\chi^2_{\rm eff} = 1$ and $\chi_b^2 = M$ are shown in Figs 11.3(d) and (e), respectively. Both of these have biases which are consistent with zero, at the expense of larger variances compared to the results from $\Delta \log L = N/2$ or minimum MSE. The $\Delta\chi^2_{\rm eff} = 1$ case has $\chi_b^2 = 20.8$, and the $\chi_b^2 = M$ case has $\Delta\chi^2_{\rm eff} = 0.85$, so in this example they are in fact very similar.

Now consider Fig. 11.4, which shows examples based on the same distribution, but now using Tikhonov regularization with $k = 2$. The figures correspond to (a) minimum weighted MSE, (b) minimum MSE, (c) $\Delta \log L = N/2$, (d) $\Delta\chi^2_{\rm eff} = 1$, and (e) $\chi_b^2 = M$. Here in particular the solution from $\Delta \log L = N/2$ does not appear to go far enough; although the statistical errors $\sqrt{U_{ii}}$ are quite small, the biases are large and significantly different from zero ($b_i^2 \gg W_{ii}$). Reasonable results are achieved in (a), (b) and (e), but the requirement $\Delta\chi^2_{\rm eff} = 1$ (d) appears to go too far. The bias is consistent with zero, but no more so than in the case with $\chi_b^2 = M$. The statistical errors are, however, much larger.

A problem with Tikhonov regularization, visible in the right most bins in Fig. 11.4, is that the estimates can become negative. (All of the bins are positive only for Fig. 11.4(a).) There is in fact nothing in the algorithm to prevent negative values. If this must be avoided, then the algorithm has to be modified by, for example, artificially decreasing the errors on points where the negative estimates would occur. This problem is absent in MaxEnt unfolding, since there the gradient of $S(\boldsymbol{\mu})$ diverges if any μ_i approach zero. This penalty keeps all of the μ_i positive.

The techniques discussed in this chapter can easily be generalized to multidimensional distributions. For the case of two dimensions, for example, unfolding methods have been widely applied to problems of image restoration [Fri72, Fri80, Fri83, Ski85], particularly in astronomy [Nar86], and medical imaging [Lou92]. A complete discussion is beyond the scope of this book, and we will only illustrate some main ideas with a simple example.

Figure 11.5 shows an example of MaxEnt unfolding applied to a test photograph with 56×56 pixels. Figure 11.5(a) is taken as the 'true' image, representing the vector $\boldsymbol{\mu}$. In Fig. 11.5(b), the image has been blurred with a Gaussian resolution function with a standard deviation equal to 0.6 times the pixel size.

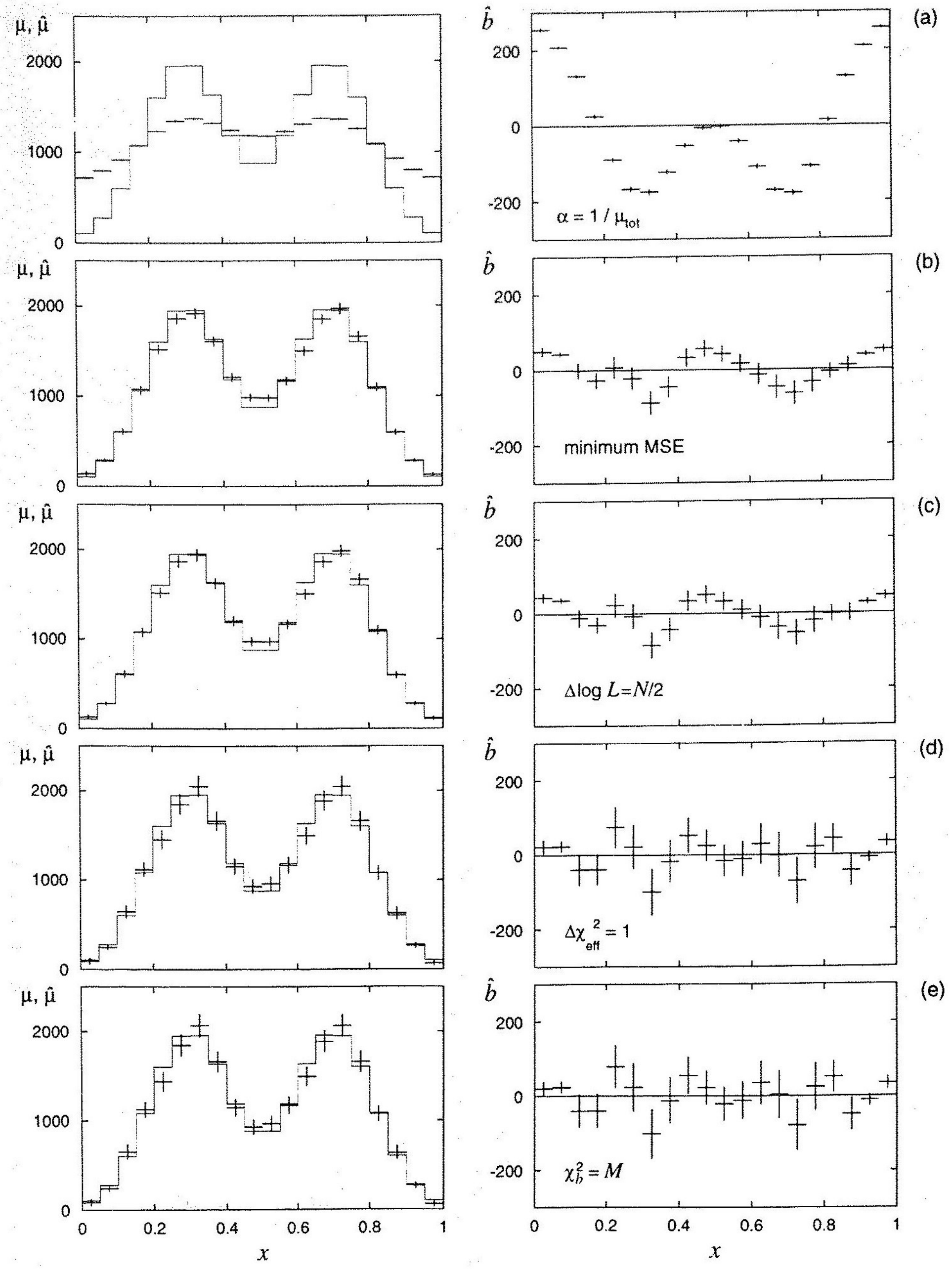

Fig. 11.3 MaxEnt unfolded distributions shown as points with the true distribution shown as a histogram (left) and the estimated biases (right) for different values of the regularization parameter α. The examples correspond to (a) the Bayesian prescription $\alpha = 1/\mu_{\rm tot}$, (b) minimum mean squared error, (c) $\Delta \log L = N/2$, (d) $\Delta\chi^2_{\rm eff} = 1$, and (e) $\chi_b^2 = M$. In this example, the solution of minimum weighted MSE turns out similar to case (c) with $\Delta \log L = N/2$.

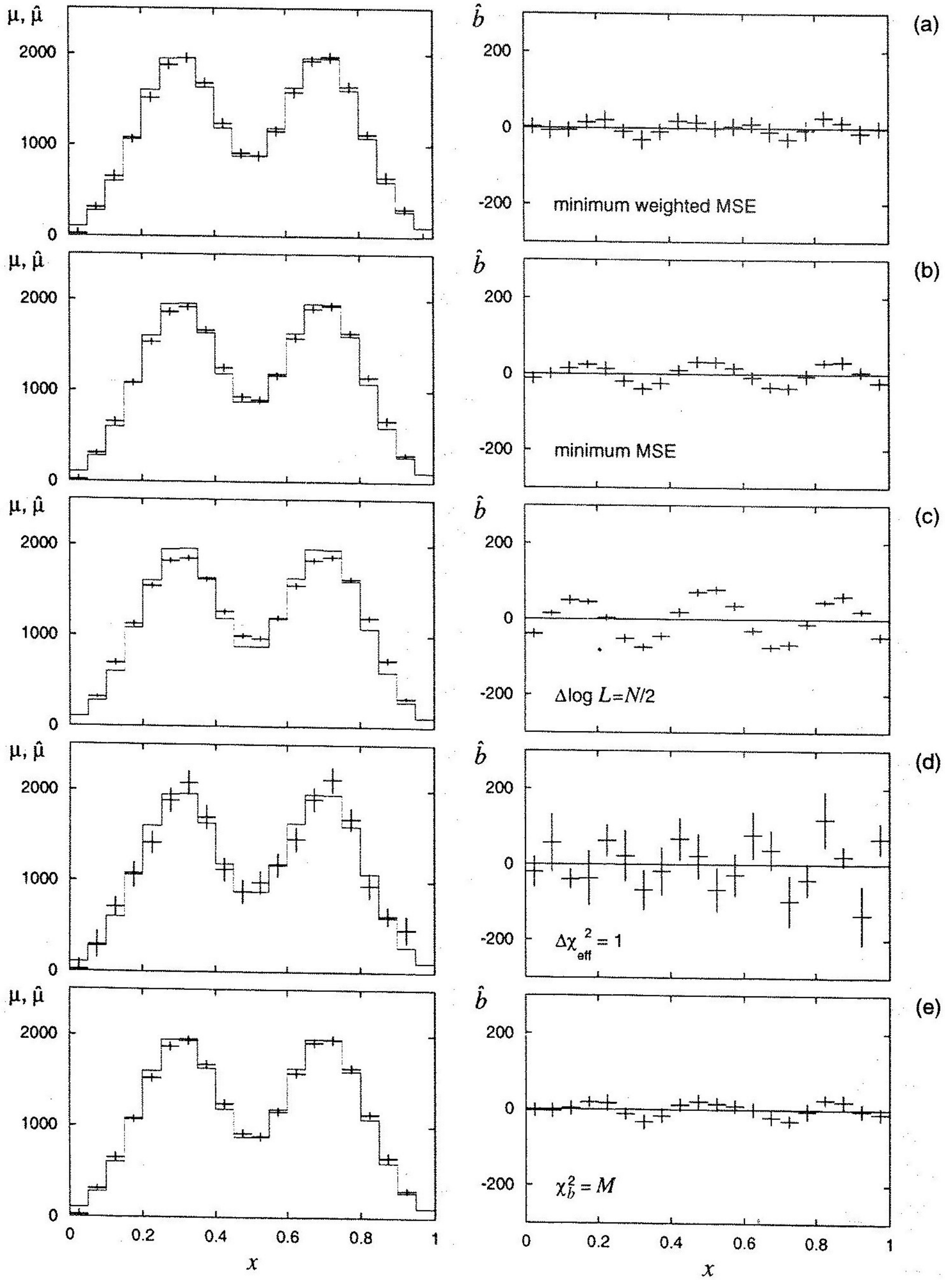

Fig. 11.4 Unfolded distributions using Tikhonov regularization ($k = 2$) shown as points with the true distribution shown as a histogram (left) and the estimated biases (right) for different values of the regularization parameter α. The examples correspond to (a) minimum weighted mean squared error, (b) minimum mean squared error, (c) $\Delta \log L = N/2$, (d) $\Delta\chi^2_{\rm eff} = 1$, and (e) $\chi^2_b = M$.

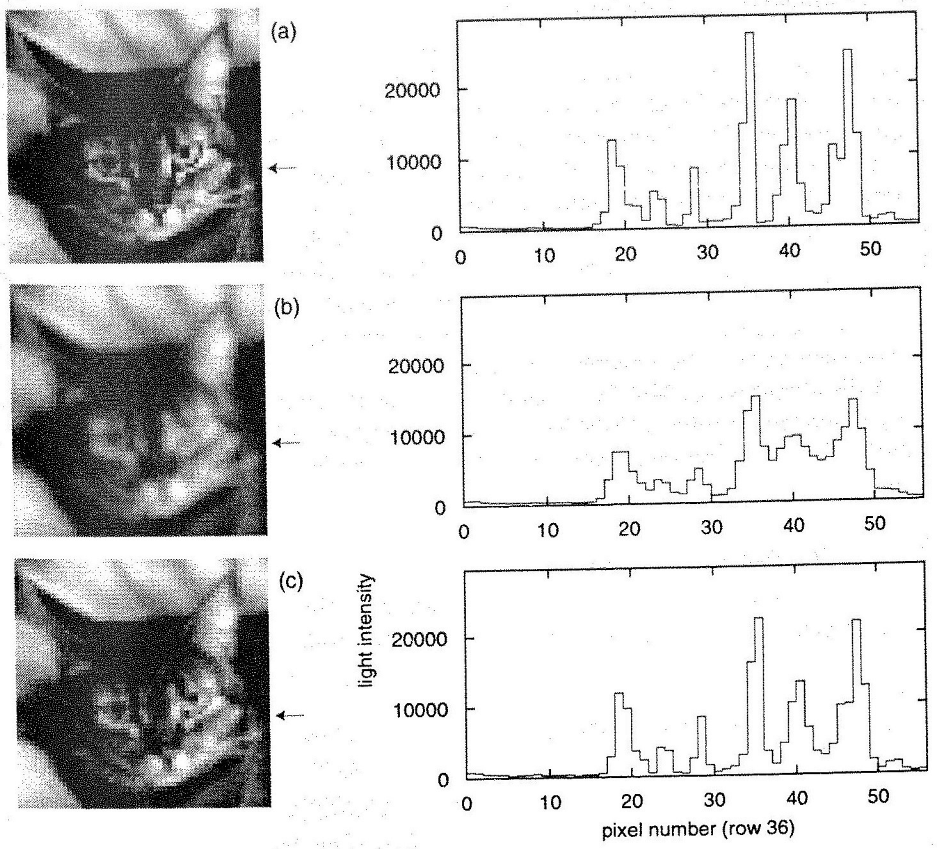

Fig. 11.5 (a) The original ‘true’ image $\boldsymbol{\mu}$. (b) The observed image $\mathbf{n}$, blurred with a Gaussian point spread function with a standard deviation equal to 60% of the pixel size. (c) The maximum entropy unfolded image. The histograms to the right show the light intensity in pixel row 36 (indicated by arrows).

For purposes of this exercise, the effective number of ‘photons’ (or, depending on the type of imaging system, silver halide grains, photoelectrons, etc.) was assigned such that the brightest pixels have on the order of 10^4 entries. Thus if the number of entries in pixel i is treated as a Poisson variable n_i with expectation value ν_i, the relative sizes of the fluctuations in the brighter regions are on the order of 1% ($\sigma_i/\nu_i = 1/\sqrt{\nu_i}$). Figure 11.5(c) shows the unfolded image according to maximum entropy with $\Delta \log L = N/2$ where $N = 3136$ is the number of pixels. The histograms to the right of Fig. 11.5 show the light intensity in pixel row 36 of the corresponding photographs.

For this particular example, the method of maximum entropy has certain advantages over Tikhonov regularization. First, there is the previously mentioned feature of MaxEnt unfolding that all of the bins remain positive by construction. Beyond that, one has the advantage that the entropy can be directly generalized to multidimensional distributions. This follows immediately from the fact that the entropy $H = -\sum_j p_j \log p_j$ is simply a sum over all bins, and pays no attention to the relative values of adjacent bins. For Tikhonov regularization, one can generalize the function $S(\boldsymbol{\mu})$ to two dimensions by using a finite-difference approximation of the Laplacian operator; see e.g. [Pre92], Chapter 18.

A consequence of the fact that entropy is independent of the relative bin locations is that the penalty against isolated peaks is relatively slight. Large peaks occur in images as bright spots such as stars, which is a reason for MaxEnt's popularity among astronomers. For relatively smooth distributions such as those in Figs 11.3 and 11.4, Tikhonov regularization leads to noticeably smaller variance for a given bias. This would not be the case for distributions with sharp peaks, such as the photograph in Fig. 11.5.

A disadvantage of MaxEnt is that it necessarily leads to nonlinear equations for $\boldsymbol{\mu}$. But the number of pixels in a picture is typically too large to allow for solution by direct matrix inversion, so that one ends up anyway using iterative numerical techniques.

11.9 Numerical implementation

The numerical implementation of the unfolding methods described in the previous sections can be a nontrivial task. Finding the maximum of the function

$$\varphi(\boldsymbol{\mu}, \lambda) = \alpha \log L(\boldsymbol{\mu}) + S(\boldsymbol{\mu}) + \lambda \left[n_{\text{tot}} - \sum_{j=1}^{N} \nu_j \right] \qquad (11.84)$$

with respect to $\boldsymbol{\mu}$ and the Lagrange multiplier λ implies solving the $M+1$ equations (11.64). If φ is a quadratic function of $\boldsymbol{\mu}$, then the equations (11.64) are linear. This occurs, for example, if one has a log-likelihood function for Gaussian distributed $\mathbf{n}$, giving $\log L = -\chi^2/2$, in conjunction with Tikhonov regularization. Methods of solution for this case based on singular value decomposition are discussed in [Höc96]. If φ contains, for example, a log-likelihood function based on the Poisson distribution, or an entropy-based regularization function, then the resulting equations are nonlinear and must be solved by iterative numerical techniques.

Consider as an example the case of a Poisson-based likelihood function, cf. equations (11.21), (11.22),

$$\log L(\boldsymbol{\mu}) = \sum_{i=1}^{N} (n_i \log \nu_i - \nu_i), \qquad (11.85)$$

with the regularization function $S = H$ where H is the entropy (11.51).

A possible method of solution for MaxEnt regularization is illustrated in Fig. 11.6. The three axes represent three dimensions of $\boldsymbol{\mu}$-space, and the diagonal plane is a subspace of constant $\sum_i \nu_i = n_{\text{tot}}$. The two points indicated in the plane are the point of maximum entropy (all μ_i equal) and the point of maximum likelihood. The curve connecting the two indicates possible solutions to (11.40) corresponding to different values of the regularization parameter α; for example, $\alpha = 0$ gives the point of maximum entropy; $\alpha \to \infty$ corresponds to the point of maximum likelihood. The curve passes through the points at which contours of constant entropy and constant likelihood touch. Note that the point of maximum likelihood is not in the allowed region with all $\mu_i > 0$. This is, in fact, typical of the oscillating maximum likelihood solution, cf. Fig. 11.1(d).

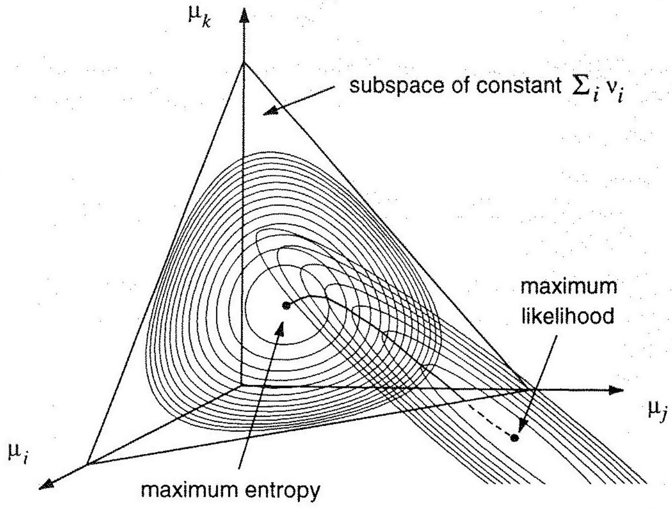

Fig. 11.6 Three dimensions of $\boldsymbol{\mu}$-space illustrating the numerical implementation of maximum entropy unfolding (see text).

The program used for the MaxEnt examples shown in Figs 11.3 and 11.5 employs the following algorithm, which includes some features of more sophisticated methods described in [Siv96, Ski85]. The point of maximum likelihood usually cannot be used for the initial value of $\boldsymbol{\mu}$, since there one often has negative μ_i, and hence the entropy is not defined. Instead, the point of maximum entropy is taken for the initial values. This is determined by requiring all μ_i equal, subject to the constraint

$$\begin{aligned} \nu_{\text{tot}} &= \sum_{i=1}^{N} \nu_i = \sum_{i=1}^{N} \sum_{j=1}^{M} R_{ij} \mu_j = \sum_{j=1}^{M} \varepsilon_j \mu_j \\ &= n_{\text{tot}}, \end{aligned} \qquad (11.86)$$

where ε_j is the efficiency for bin j. The point of maximum entropy is thus

$$\mu_i = \frac{n_{\text{tot}}}{\sum_{j=1}^{M} \varepsilon_j}. \tag{11.87}$$

If one uses $S(\boldsymbol{\mu}) = \mu_{\text{tot}} H$, and if the efficiencies are not all equal, then the distribution of maximum $S(\boldsymbol{\mu})$ is not uniform, but rather is given by the solution to the M equations,

$$\log \frac{\mu_i}{\mu_{\text{tot}}} + \frac{S(\boldsymbol{\mu})\,\varepsilon_i}{n_{\text{tot}}} = 0, \quad i = 1, \ldots, M. \tag{11.88}$$

Starting from the point of maximum $S(\boldsymbol{\mu})$, one steps along the curve of maximum φ in the subspace of constant ν_{tot}. As long as one remains in this subspace, it is only necessary to maximize the quantity

$$\Phi(\boldsymbol{\mu}) = \alpha \, \log L(\boldsymbol{\mu}) + S(\boldsymbol{\mu}), \tag{11.89}$$

i.e. the same as $\varphi(\boldsymbol{\mu})$ but without the Lagrange multiplier λ, cf. (11.40). Simply requiring $\nabla \Phi = 0$ will not, however, lead to the desired solution. Rather, $\nabla \Phi$ must be first projected into the subspace of constant ν_{tot} and the components of the resulting vector set equal to zero. In this way the Lagrange multiplier λ never enters explicitly into the algorithm. That is, the solution is found by requiring

$$D\Phi = \nabla \Phi - \mathbf{u}(\mathbf{u} \cdot \nabla \Phi) = 0, \tag{11.90}$$

where $\mathbf{u}$ is a unit vector in the direction of $\nabla \nu_{\text{tot}}$. This is given by (cf. (11.10))

$$\frac{\partial \nu_{\text{tot}}}{\partial \mu_k} = \sum_{i=1}^{N} \sum_{j=1}^{M} R_{ij} \frac{\partial \mu_j}{\partial \mu_k} = \varepsilon_k, \tag{11.91}$$

so that the vector $\mathbf{u}$ is simply given by the vector of efficiencies, normalized to unit length,

$$\mathbf{u} = \frac{\boldsymbol{\varepsilon}}{|\boldsymbol{\varepsilon}|}. \tag{11.92}$$

We will use the differential operator D to denote the projection of the gradient into the subspace of constant ν_{tot}, as defined by equation (11.90).

One begins thus at the point of maximum entropy and takes a small step in the direction of $D \log L$. The resulting $\boldsymbol{\mu}$ is in general not directly on the curve of maximum Φ, but it will be close, as long as the step taken is sufficiently small. As a measure of distance from this curve one can examine $|D\Phi|$. If this exceeds a given limit then the step was too far; it is undone and a smaller step is taken.

If the resulting point $\boldsymbol{\mu}$ were in fact on the curve of maximum Φ, then we would have $\alpha \, D \log L + DS = 0$ and the corresponding regularization parameter would be

$$\alpha = \frac{|DS|}{|D \log L|}. \tag{11.93}$$

The parameter α can simply be set equal to the right-hand side of (11.93), and a side step taken to return to the curve of $D\Phi = 0$. This can be done using standard methods of function maximization (usually reformulated as minimization; cf. [Bra92, Pre92]). These side steps as well are made such that one remains in the subspace of $\nu_{\text{tot}} = n_{\text{tot}}$, i.e. the search directions are projected into this subspace. One then proceeds in this fashion, increasing α by means of the forwards steps along $D \log L$ and moving to the solution $D\Phi = 0$ with the side steps, until the desired value of $\Delta \log L = \log L_{\max} - \log L$ is reached. Intermediate results can be stored and examined in order to determine the optimal stopping point.

Although the basic ideas of the algorithm outlined above can also be applied to Tikhonov regularization, the situation there is somewhat complicated by the fact that the solution of maximum $S(\boldsymbol{\mu})$ is not uniquely determined. For $k = 2$, for example, any linear function gives $S = 0$. One can simply start at $\mu_i = n_{\text{tot}}/M$ and set the regularization parameter α sufficiently large that a unique solution is found.

It is also possible with Tikhonov regularization to leave off the condition $\sum_i \nu_i = n_{\text{tot}}$, since here the regularization function does not tend to pull the solution to a very different total normalization. If the normalization condition is omitted, however, then one will not obtain exactly $\sum_i \hat{\nu}_i = n_{\text{tot}}$. One can argue that $\hat{\nu}_{\text{tot}}$ should be an unbiased estimator for the total number of events, but since the bias is not large, the constraint is usually not included.