# mmzm_linear

September 29, 2016

## 1 Lineární model

Ve zhuštěné podobě se řešení lineární regrese zapisuje pomocí matic.

Hledáme hodnoty *K* parametrů $\theta_j$ pomocí *N* měření: $E(Y_i|\theta) = \sum_j^N a_{ij}\theta_j$, či v maticové formě $E(\mathbf{Y}|\theta) = \mathbf{A}\theta$. Matice $a_{ij} = f_j(x_i)$ jsou hodnoty sady $j = 1..K$ funkcí (např. různé mocniny v případě polynomiálního modelu) vyjádřených v $i = 1..N$ měřených bodech $x_i$. Předpokládáme, že měření $y_i$ jsou nezávislá, tedy disperzní matice $D(\mathbf{Y}) = \sigma^2\mathbf{W}^{-1}$ je diagonální (váhy mohou být normovány na $Tr(\mathbf{W}) = 1$).

Zobecněním dvojpar. postupu dostaneme pro ML odhad soustavu rovnic

$$\mathbf{A}^T\mathbf{W}\mathbf{Y} = (\mathbf{A}^T\mathbf{W}\mathbf{A})\widehat{\theta}$$

kde součin v závorce je Hessián **H**, regulární symetrická matice; k ní inverzní označ. **D** určuje disperzi. Platí

$$D(\widehat{\theta}) = \sigma^2\mathbf{D}$$

kdy $\widehat{\theta} = D\mathbf{A}^T\mathbf{W}\mathbf{Y}$ je lineární kombinace normálně rozdělených NP (těmi jsou měřené hodnoty $Y$), tedy také normální NP.

Pokud $\sigma^2$ neznáme, odhadujeme ji pomocí "reziduálního součtu čtverců"

$$\widehat{\sigma^2} = \frac{1}{N-p}(\mathbf{Y}-\widehat{\mathbf{Y}})^T\mathbf{W}(\mathbf{Y}-\widehat{\mathbf{Y}}) = \frac{1}{N-p}\sum_i^N w_i(y_i-\widehat{y_i})^2 = \frac{S_0}{N-p}$$

kde $\widehat{\mathbf{Y}} = \mathbf{A}\widehat{\theta}$ (předpověď modelu) a $p$ je počet parametrů (dimenze $\theta_j$).

## 2 Mnohorozměrné problémy

Základní otázky otázka potřebnosti dalšího parametru - jaký nejmenší počet proměnných vysvětluje dostatečně data?

### 2.1 Hlavní komponenty - principal component analysis (PCA)

Matice $A$ popisuje transformaci (rotaci/inverzi) měřených veličin

$$Y = AX$$

- hledáme takovou kombinaci $a_1X$, kdy $V(a_1X)$ bude největší za normalizační podmínky $a_1b_1 = 1$ -> *první hlavní komponenta*

- pak hledáme takovou kombinaci $a_2X$ , kdy $V(a_2X)$ bude největší za podmínky $a_2b_2 = 1$ a $Cov(a_1X, a_2X) = 0$ -> *druhá hlavní komponenta*

Nechť proměnné X mají kovarianční matici $\Sigma$

řešení: najdeme vlastní čísla $\lambda_i$ a vlastní vektory $\pi_i$ kovar. matice, předpokládáme, že budou ortogonální (autom. splněno, pokud jsou vlastní čísla různá).

Matice W vlastních vektorů matice $X^TX$ a sdružená matice V vlastních vektorů matice $XX^T$ (ident. v případě čtvercové matice X) jsou transformačními maticemi **singulární dekompozice** matice X ve tvaru $X = WLV$, kde L je matice pouze s diagonálními nezápornými elementy.

Stopa kovar. matice je při transformaci zachována - součet variancí je součtem vlastních čísel. Vlastní vektory obvykle uspořádáme podle velikosti vlast. čísel.

**Reference:** [Francis] Paul J. Francis, Beverley J. Wills http://arxiv.org/abs/astro-ph/9905079 + code ref.

## 2.2 Faktorová analýza

jde o rozklad kovarianční matice $\Sigma$ na několik ($m$) společných faktorů a zbylé "specifické" faktory

$E(X) = \mu$

$X - \mu = LF + \epsilon$

$E(F) = 0, Cov(F) = I$ (ortogonální faktory); $E(\epsilon) = 0, Cov(\epsilon) = \Psi$ (diagonální)

pak $Cov(X) = LL' + \Psi$

faktory $F$ jsou určeny až na ortogonální rotaci, "loading" L určíme jako $L = Cov(X, F)$

faktorizace může vycházet z PCA - $L = \sqrt(\lambda)e$, kdy zahrneme jen $m$ nejvýznamnějších vlastních vektorů

**Faktorová analýza** je termín používaný i pro plány experimentů