Splines, BLUPS, MDL and Knots Selection - ECNU Seminar Notes

Bu Zhou

2011,6,26

1 Linear mixed models

Linear mixed model is
$y = X\beta + Zu + \epsilon$
where
$E\begin{bmatrix} u \\ \epsilon \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $Cov\begin{bmatrix} u \\ \epsilon \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}$

2 Estimation for linear mixed models

suppose:
$u \sim N(0, \sigma_u^2 I_m)$
$\epsilon \sim N(0, \sigma_\epsilon^2 I_n)$
$cov(u, \epsilon') = 0$
$X_{n\times p}$ known matrix
$Z_{n\times m}$ known matrix
$u_{m\times 1}$ vector of random effects
$y_{n\times 1}$ data vector
$\beta_{p\times 1}$ vector of unknown fixed effects parameters
we have:
$y \sim N(X\beta, \sigma^2 V)$, $V = \lambda ZZ' + I_n$, $\lambda = \sigma_u^2/\sigma_\epsilon^2$ and $y|u \sim N(X\beta + Zu, \sigma_\epsilon^2 I_n)$ (*0)

penalized least squares criterion: $||y - X\beta - Zu||^2 + \frac{1}{\lambda}||u||^2$(*), $\lambda > 0$ is a tuning parameter.

For given λ, minimizing (*) with respect to β and u leads to so-called mixed model equations

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + \frac{1}{\lambda} I_m \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

then:
$\hat{\delta} = \begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix} = (M'M + \frac{1}{\lambda} D)^{-1} M'y$; $M = (XZ)$,$D_{(p+m)\times(p+m)} = diag(0, \cdots, 0, 1, \cdots, 1)$

the conditional ML(see(*0)) estimator of σ_ϵ^2 is $\hat{\sigma_\epsilon^2} = n^{-1}||y - M\hat{\delta}||^2 = n^{-1} y'(I - H)^2 y$

fitted values are $\hat{y} = Hy$; $H = M(M'M + \frac{1}{\lambda} D)^{-1} M'$

3 Penalized Splines as BLUPs

(Brumback BA , Ruppert D , Wand MP (1999) Comment on Shively, Kohn & Wood . Journal of the American Statistical Association)

$y_i = f(x_i) + \sigma_\epsilon \epsilon_i, 1 \leq i \leq n$,

scatterplot data: $(x_i, y_i), 1 \leq i \leq n$

ϵ_i: independent $N(0,1)$ variates, $cov(\epsilon) = I$

$\kappa_1, \cdots, \kappa_K$: set of distinct numbers inside the range of the x_i's

$x_+ = max(0, x)$

spline model: $f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x^p + \sigma_u \Sigma_{k=1}^K u_k (x_i - \kappa_k)_+^p$

$u = \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix} \sim N(0, I)$, is independent of $\epsilon = \begin{bmatrix} \epsilon_1 & \cdots & \epsilon_n \end{bmatrix}^T$

$y = X\beta + \sigma_u Z u + \sigma_\epsilon \epsilon$, $\begin{bmatrix} u \\ \epsilon \end{bmatrix} \sim N(0, I)$

where

$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ and $Z = \begin{bmatrix} (x_1 - \kappa_1)_+ & \cdots & (x_1 - \kappa_K)_+ \\ \vdots & \ddots & \vdots \\ (x_n - \kappa_1)_+ & \cdots & (x_n - \kappa_K)_+ \end{bmatrix}$

Estimation for given σ_u and σ_ϵ(replaced by estimators $\hat{\sigma_u}$ and $\hat{\sigma_\epsilon}$), the best linear unbiased predictor (BLUP) of y is

$\hat{f} = X\hat{\beta} + \sigma_u Z \hat{u}$

where $\hat{\beta} = \{X^T(\sigma_u^2 ZZ^T + \sigma_\epsilon^2 I)^{-1} X\}^{-1} X^T (\sigma_u^2 ZZ^T + \sigma_\epsilon^2 I)^{-1} y$ and $\hat{u} = (\sigma_u^2 Z^T Z + \sigma_\epsilon^2 I)^{-1} Z^T (y - X\hat{\beta})$

can be treated as an estimator of $f = [f(x_1), \cdots f(x_n)]^T$(extends to $f(x)$ for arbitrary x is straightforward)

$\hat{f}$ can be rewritten as $\hat{f} = C(C^T C + \lambda D)^{-1} C^T y$

where $C = [XZ]$, $D = diag(0, 0, 1, 1, \cdots, 1)$ and $\lambda = \sigma_\epsilon^2 / \sigma_u^2$.

4 Minimum Description Length in the view of Bayesianism

[Applying MDL To Learning Best Model Granularity, by Q.Gao, M.Li, and P.M.B.Vitanyi, Artificial Intelligence, 2000]

4.1 Introduction:

Drawback of Person-Neyman testing:

Rejection of the zero hypothesis does not imply the acceptation of alternative hypothesis. Does not establish the relative likelihood between competing hypotheses.(All hypothesis different from the zero hypothesis must be taken together to form the alternative hypothesis, we can not even use the same data to test the alternative hypothesis or a subset of it)

Bayesianism:

$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D)=\Sigma_i P(D|H_i)P(H_i)}$, select hypothesis/model with the maximum a posterior probability(MAP)

where

Ω: a discrete sample space

$D, H_1, H_2, \cdots$: a countable set of events(subsets) of Ω

$H = \{H_1, H_2, \cdots\}$: hypothesis space

hypotheses H_i are exhaustive(at least one is true), mutually exclusive($H_i \cap H_j = \phi$ for all i, j)

Advantage: Allow to estimate the relative likelihood of different possible hypotheses.

Disadvantage: Prior probability $P(H_i)$, how to initially derive it? May be unknown, uncomputable or conceivably nonexistent.

We know where to go next(Bayes's updating rule), but where shall we start(prior)?

The answer: Find a single probability distribution to use as the prior distribution in each different case, with approximately the same result as if we had used the real distribution.

Surprisingly, this solution turns out to be possible up to some mild restrictions.

Universal prior in Bayes' rule: algorithmic universal probability $m(x|y) = 2^{-K(x|y)}$,where $K(x|y)$: prefix Kolmogrov complexity of x given y.

Problem: Cannot be directly used since Kolmogrov complexity $K(x|y)$ is non-computable, and so is the algorithmic universal probability $m(x|y)$. Approximation is needed in the real world applications.

A "good" computable approximation to $m(x)$ $\Rightarrow$MDL:

4.2 MDL in one page:

From Bayes' formula, we must choose the hypothesis H that maximizes the posterior $P(H|D)$, taking the negative logarithm on both side,

$-logP(H|D) = -logP(D|H) - logP(H) + logP(D)$

$logP(D)$ is a constant and can be ignored because we just want to optimize the left-hand side of the equation over H.

The problem is minimizing $-logP(D|H) - logP(H) =$

- (the log universal probability of the model + the log of the probability of the data given the model)

Ideal MDL: $K(H) + K(D|H)$ (use universal prior)

Real MDL: $-logP(D|H) - logP(H)$, $P(D|H)$ must be computable. (Applied statistical version of MDL, use Shannon-Frano code as the approximation of the non-computable Kolmogrov complexity)

The Shannon-Frano code assigns code words of $length \stackrel{+}{=} -logP(.)$ to elements randomly drawn according to a probability $P(.)$.

$-logP(H)$: the length, in bits, of the description of the theory, $= K(H)$, provided that $-logP(H) \stackrel{+}{=} -logm(H)$

$-logP(D|H)$: the length, in bits, of data when encoded with the help of the theory, $= K(D|H)$ provided that $-logP(D|H) \stackrel{+}{=} -logm(D|H)$

For "typical" outcomes, $K(D|H) \stackrel{+}{=} -logP(D|H)$ means that the classic Shannon-Frano code length reaches the prefix Kolmogorov complexity on these data samples. (Under the assumption that the data sample is typical for the contemplated hypotheses, the ideal MDL principal and the applied statistical one coincide, and moreover, both are valid for a set of data samples of Lebesgue measure one.)

4.3 Examples

Example 1. (Hypothesis Testing)

H: some model $H(\theta)$ with a set of parameter $\theta = \{\theta_1, \cdots, \theta_k\}$ of precision c.

number k may vary and influence the description complexity of $H(\theta)$.

For example, if we want to determine the distribution of the length of beans, then H is a normal distribution $N(\mu, \sigma)$ with parameter median μ and variation σ. So essentially we have to determine the correct hypothesis described by identifying the type of distribution(normal) and the correct parameter vector (μ, σ).

In such cased, we minimize $-logP(D|\theta) - logP(\theta)$.

Example 2. (Fitting of a 'best' polynomial on n given sample points in the 2-dimensional plane.)

For each fixed k, $k = 0, \cdots, n-1$, let f_k be the best polynomial of degree k, fitting on points $(x_i, y_i)(1 \leq i \leq n)$, which minimizes the error

$error(f_k) = \Sigma_{i=1}^{n}(f_k(x_i) - y_i)^2$.

Assume each coefficient takes c bits, so f_k is encoded in $c(k+1)$ bits.

Assume $Pr(y_1, \cdots, y_n|f, x_1, \cdots, x_n) = \prod exp\{-O[(f(x_i)-y_i)^2]\}$.(measurement with Gaussian/normal errors), so $-log(Pr(y_1, \cdots, y_n|f, x_1, \cdots, x_n)) = c'error(f)$ for some computable c'.

The MDL principle tells us to choose $f = f_m$ with $m \in \{0, \cdots, n-1\}$, which minimizes $c(m+1) + c'error(f_m)$.

5 Model selection by the MDL principle

data: $\{x_i, y_i\}_{i=1}^{n}, y_i = f(x_i) + \epsilon_i, \epsilon_i \sim iidN(0, \sigma^2)$

f can be well approximated by r-order regression spline with m knots: $f(x) \approx b_0 + b_1 x + \cdots + b_r x^r + \Sigma_{j=1}^{m}\beta_j(x - k_j)_+^r$

k_j: location of the jth knot

$\{b_0, \cdots, b_r, \beta_1, \cdots, \beta_m\}$: set of the coefficients

$(a)_+ = max(0, a)$

$(a)_+^c = ((a)_+)^c$

assume $min(x_i) < k_1 < \cdots < k_m < max(x_i)$ and $\{k_1, \cdots, k_m\}$ is subset of $\{x_1, \cdots, x_m\}$

estimation of f can be transformed into a model selection problem with each plausible model θ completely specified by $\theta = \{r, m, k, b, \beta\}$

note that different θ's may have different dimensions (number of parameters)

Let

$x = (x_1, \cdots, x_n)^T$, $y = (y_1, \cdots, y_n)^T$ and $M = (XZ) = (1, x, \cdots, x^{\hat{r}}, (x - \hat{k_1}1_{n\times 1})_+^{\hat{r}}, \cdots, (x - \hat{k}_{\hat{m}}1)_+^{\hat{r}})$

5.1 regression spline

the natural estimates(maximum likelihood estimates conditional on $\hat{r}$, $\hat{m}$ and $\hat{k}$) of b and β are given by

$(\hat{b}^T, \hat{\beta}^T)^T = (M^T M)^{-1} M^T y$

we estimate θ and hence f by minimizing the

$$MDL(\hat{\theta}) = \begin{array}{c} log\hat{m} + \Sigma_{j=1}^{\hat{m}} log\hat{l_j} + \frac{1}{2}\Sigma_{j=1}^{\hat{m}+1} log\hat{l_j} + \frac{n}{2} log\{\frac{RSS(\hat{\theta})}{n}\}, \hat{r} = 0; \\ log\hat{m} + \Sigma_{j=1}^{\hat{m}} log\hat{l_j} + \frac{\hat{m}+\hat{r}+1}{2} logn + \frac{n}{2} log\{\frac{RSS(\hat{\theta})}{n}\}, hatr \geq 1. \end{array}$$

Example 3. (Derivation of code length expression for choosing knots in regression spline smoothing)

$L(y) = L$(fitted model)$+L$(data given the fitted model)$= L(\hat{\theta}) + L(y|\hat{\theta})$.

1. $L(\hat{\theta}) = L(\hat{r}) + L(\hat{m}) + L(\hat{k}|\hat{m}) + L(\hat{b}, \hat{\beta}|\hat{r}, \hat{m}, \hat{k})$

code length for $\hat{r}$ and $\hat{m}(L(\hat{r}) + L(\hat{m}))$:

use $r \in \{0, 1, 2, 3\} \Longrightarrow L(\hat{r}) = -log_2 \frac{1}{4} = log_2 4 = 2$ bits, a constant that will be ignored.

$\hat{m}$ is an integer, $L(\hat{m}) = L^*(\hat{m}) = log_2 c + log_2 \hat{m} + log_2 log_2 \hat{m} + \cdots$, where $c \approx 2.865$, $L^*(\hat{m}) \approx log_2 \hat{m}$when $\hat{m}$ is reasonably large.

Thus $L(\hat{r}) + L(\hat{m}) \approx log_2 \hat{m}$.

code length for $\hat{k}$ given $\hat{m}(L(\hat{k}|\hat{m}))$:

note that $\{\hat{k_1}, \ldots, \hat{k_{\hat{m}}}\}$ is restricted to be a subset of $\{x_1, \cdots, x_n\}$.

$L(\hat{k}|\hat{m}) = L(\hat{l_1}, \cdots, \hat{l_{\hat{m}}}|\hat{m}) = \Sigma_{j=1}^{\hat{m}} L^*(\hat{l_j}) \approx \Sigma_{j=1}^{\hat{m}} log_2 \hat{l_j}$.

where $\hat{l_j}$(integers) is the jth successive "index difference", the number of x_i's which satisfy $\hat{k_{j-1}} \leq x_i < \hat{k_j}, j = 1, \cdots, \hat{m}; \hat{k_0} = min(x_i), \hat{k_{\hat{m}+1}} = max(x_i)$.

($\hat{l_1}, \cdots, \hat{l_{\hat{m}}}$ implies complete knowledge of $\hat{k}$.)

code length for $\{\hat{b}, \hat{\beta}\}$ given $\{\hat{r}, \hat{m}, \hat{k}\}(L(\hat{b}, \hat{\beta}|\hat{r}, \hat{m}, \hat{k}))$:

each of $\{\hat{b}, \hat{\beta}\}$- $\hat{b}_i$ or $\hat{\beta}_i$, is (conditional) maximum likelihood estimate from n data points when $\hat{r} \geq 1$, so it can be effectively encoded with $\frac{1}{2} log_2 n$ bits.

$L(\hat{b_0}) = \cdots = L(\hat{b_{\hat{r}}}) = L(\hat{\beta}_1) = \cdots = L(\hat{\beta_{\hat{m}}}) = \frac{1}{2} log_2 n$, so

$L(\hat{b}, \hat{\beta}|\hat{r}, \hat{m}, \hat{k}) = \frac{\hat{r}+\hat{m}+1}{2} log_2 n$, when $\hat{r} \geq 1$.

$L(\hat{b}, \hat{\beta}|\hat{r}, \hat{m}, \hat{k}) = \Sigma_{j=1}^{\hat{m}} L(\text{"}j\text{th estimated height"}) = \frac{1}{2}\Sigma_{j=1}^{\hat{m}+1} log_2 \hat{l_j}$, when $\hat{r} = 0$.

code length for y given $\hat{\theta} = \{\hat{r}, \hat{m}, \hat{k}, \hat{b}, \hat{\beta}\}(L(y|\hat{\theta}))$:

$L(y|\hat{\theta}) = \frac{n}{2} log_2\{\frac{RSS(\hat{\theta})}{n}\} + C$, where C is a negligible term and $RSS(\hat{\theta}) = \Sigma\{y_i - \hat{f}(x_i)\}^2$.

5.2 smoothing spline/penalized spline

$(\hat{b}^T, \hat{\beta}^T)^T = (M^T M + \frac{1}{\lambda} D)^{-1} M^T y$

consider a set of normal models of form $y|b_\eta \sim N(X_\eta \beta_\eta + Z_\eta b_\eta, \sigma^2 I_n)$

η:index of the set of candidate models

X_η: $n \times r_\eta$ matrix

Z_η: $n \times m_\eta$ matrix

β_η: $r_\eta \times 1$ parameter vector

b_η: $m_\eta \times 1$ parameter vector

Note that estimates β_η, b_η and $\hat{\sigma}^2_\eta$ depends on the parameter $\lambda \in [0, \infty]$

we specify a model by giving the pair $\gamma = (\eta, \lambda)$

assume response data are modelled with a set of density functions $f(y; \gamma, \theta)$, where parameter vector θ varies within a specified parameter space,

the NML(normalized maximum likelihood) function is defined by

$\hat{f}(y; \gamma) = \frac{f(y; \gamma, \hat{\theta})}{C(\gamma)}$,

where

$\hat{\theta} = \hat{\theta}(y)$ is the ML estimator of θ and

$C(\gamma) = \int f(x; \gamma, \hat{\theta}(x)) dx$ is the normalizing constant.

integral is taken over the sample space

$\hat{f}(y; \gamma)$ defines a density function provided that $C(\gamma)$ is bounded.

shortest code length: $-log\hat{f}(y; \gamma) = -logf(y; \gamma, \hat{\theta}) + logC(\gamma)$

$f(y; \gamma, \hat{\theta}) = (2\pi)^{-\frac{n}{2}} |\hat{\Sigma}|^{-\frac{1}{2}} exp\{-\frac{1}{2}(y - (X\hat{\beta} + Z\hat{b}))^T \hat{\Sigma}^{-1} (y - (X\hat{\beta} + Z\hat{b}))\}$

where $\hat{\Sigma} = \hat{\sigma^2} I$,

$\hat{\sigma^2} = n^{-1}||y - \hat{y}||^2 = n^{-1}||y - (X\hat{\beta} + Z\hat{b})||^2 = n^{-1}||y - M^T\hat{\theta}||^2 = n^{-1} y^T ||I - H||^2 y$, $H = (M^T M + \frac{1}{\lambda} D)^{-1} M^T$.

then we have $f(y; \gamma, \hat{\theta}) = (2\pi)^{-\frac{n}{2}} |n^{-1}||y - \hat{y}||^2 I_n|^{-\frac{1}{2}} exp\{-\frac{1}{2} trace(I_n)\} = (2\pi)^{-\frac{n}{2}} \{n^{-1}||y - \hat{y}||^2\}^{-\frac{n}{2}} exp\{-\frac{1}{2}n\}$

$-log_2 f(y; \gamma, \hat{\theta}) = \frac{n}{2} log_2(2\pi) + \frac{n}{2} log_2\{n^{-1}||y - \hat{y}||^2\} + log_2\{exp(-\frac{1}{2}n)\} = \frac{n}{2} log_2\{\frac{RSS(\hat{\theta})}{n}\} + C$,

where $RSS(\hat{\theta}) = n^{-1}||y - \hat{y}||^2 = \Sigma_{i=1}^n (y_i - \hat{y_i})^2$ is the residual sum of squares.

$C = \frac{n}{2} log_2(2\pi) + log_2\{exp(-\frac{1}{2}n)\}$ is a negligible term.

5.3 Knots deletion algorithm

Problem: finding the global minimizer of $MDL(\hat{\theta})$ is difficult(Due to the complexity of $\hat{\theta}$)- a global search is infeasible even if n is only of moderate size.

Disadvantage: miss the global minimizer

Advantage: guarantees to find a local minimizer

Algorithm:

1. fixed a value for $\hat{r}$(spline order, 0,1,2,3), placing(every s(3~5) sorted values of x_i's/equally spaced sample quantiles of $x_1 \cdots x_n$) a relatively large number(K) of initial knots, (overfitted model) compute $MDL(\hat{\theta})$

2. for each knot in all knots, remove this knot, recompute $MDL(\hat{\theta})$; choose/delete/record the knot s.t. $MDL(\hat{\theta})1$-$MDL(\hat{\theta})2$=max (greedy strategy)

3. goto 1 until all initial knots are removed

4. choose the one has the smallest $MDL(\hat{\theta})$ value with respect to K ($K+$ 1models in total) as the best fitted model for that fixed value of $\hat{r}$

5. change the value of $\hat{r}$, goto 1 until all candidate values(0,1,2,3) are computed

(5+.change the value of λ, goto 1 until all candidate values are computed)

6. chose the model has smallest $MDL(\hat{\theta})$ with respect to $\hat{r}$(and λ) as the final model

6 References:

6.1 Books:

[Semiparametric Regression by David Ruppert, M. P. Wand, R. J. Carroll, 2003]

[Longitudinal Data Analysis, Edited by Garrett Fitzmaurice, Marie Davidian, Geert Verbeke & Geert Molenberghs, 2008]

[Stochastic Complexity in Statistical Inquiry, by Jorma Rissanen, World Scientific, 1989]

[Information and Complexity in Statistical Modeling, by Jorma Rissanen, Springer, 2007]

[The Minimum Description Length Principle, by Peter Grunwald, MIT Press, 2007]

6.2 Papers:

[Applying MDL To Learning Best Model Granularity, by Q.Gao, M.Li, and P.M.B.Vitanyi, Artificial Intelligence, 2000]

[Comment on Shively, Kohn and Wood, by Babette A. Brumback , David Ruppert , M. P. Wand, Journal of the American Statistical Association, 1999]

[Regression Spline Smoothing Using The Minimum Description Length Principle, by Thomas C. M. Lee, Statistics & Probability Letters, 2000]

[MDL Knot Selection For Penalized Splines, by Antti Liski and Erkki P. Liski, WITMSE, 2008]

[Model Selection In Linear Mixed Models Using MDL Criterion With An Application To Spline Smoothing, by Erkki P. Liski and Antti Liski, WITMSE, 2008]

[http://www.mdl-research.org/]