Parameter estimation

Introduction

- Linear random effects model

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + a_i + e_{ij}, \quad i = 1, \cdots, m, j = 1, \cdots, n_i, \tag{1}$$

  where $a_i \sim N(0, \sigma_a^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$.
- Two different parameters
  1. Level-1 model parameter: $\theta = (\boldsymbol{\beta}, \sigma_e^2)$
  2. Level-2 model parameter (or tuning parameter): $\lambda = \sigma_e^2/\sigma_a^2$
- The tuning parameter determines the level of shrinkage in the final prediction.
- We can treat a_i as missing data and use EM algorithm to compute the MLE of $\boldsymbol{\beta}, \sigma_e^2$ and λ simultaneously.
- However, such a joint estimation may not be a good idea.

Joint estimation

- Let $L(\theta, \lambda) = f_m(y; \theta, \lambda)$ be the likelihood function of (θ, λ).
- To estimate the parameters, we often use the following procedure:
  1. Compute the profile likelihood for λ:
  $$L_p(\lambda) = L(\hat{\theta}_\lambda, \lambda) \tag{2}$$
  where $\hat{\theta}_\lambda = \hat{\theta}(\lambda)$ is the maximizer of $L(\theta, \lambda)$ with respect to θ only.
  2. Find the maximizer $\hat{\lambda}$ of $L_p(\lambda)$ and obtain $\hat{\hat{\theta}} = \hat{\theta}(\hat{\lambda})$.
- However, the profile likelihood in (2) is not a true likelihood. Note that
$$\int L_p(\lambda) dy \neq 1$$
while we have
$$\int L(\theta, \lambda) dy = 1.$$

Remark

- For accurate estimation of λ, we may consider

$$f(y;\lambda) = \frac{f(y;\hat{\theta}_\lambda,\lambda)}{\int f(y;\hat{\theta}_\lambda,\lambda)dy}.$$

- Now, taking log of the above equality, we obtain the marginal log-likelihood

$$\ell_m(\lambda) = \ell_p(\lambda) - \log K(\lambda)$$

  where $K(\lambda) = \int f(y;\hat{\theta}_\lambda,\lambda)dy$.
- Note that $K(\lambda)$ contains information about λ.
- The maximizer of $\ell_m(\lambda)$ is different from the maximizer of $\ell_p(\lambda)$.

Direct ML estimation

- We wish to consider the marginal log-likelihood

$$
\begin{aligned}
\ell(\theta \mid \lambda) &= \sum_{i=1}^{m} \log \int \exp\left\{ \ell_1(\theta, a_i; \mathbf{y}_i) - \frac{1}{2\sigma_a^2} a_i^2 \right\} da_i \\
&= \sum_{i=1}^{m} \log \int \exp\left\{ Q_\lambda(a_i, \theta) \right\} da_i
\end{aligned}
$$

- Writing

$$
\hat{a}_i^* = \arg\max_{a_i} Q_\lambda(a_i, \theta),
$$

we may approximate

$$
\begin{aligned}
Q_\lambda(a_i, \theta) &\cong Q_\lambda\left(\hat{a}_i^*, \theta\right) + 0.5 \ddot{Q}_\lambda\left(\hat{a}_i^*, \theta\right)\left(a_i - \hat{a}_i^*\right)^2 \\
&:= Q_\lambda\left(\hat{a}_i^*, \theta\right) - 0.5 \{V_i^*(\theta)\}^{-1}\left(a_i - \hat{a}_i^*\right)^2
\end{aligned}
$$

where $V_i^*(\theta) = -1/\ddot{Q}_\lambda\left(\hat{a}_i^*, \theta\right)$.

- We use the density function of the normal distribution to get

$$\int \exp\left[-0.5\{V_i^*(\theta)\}^{-1}(a_i - \hat{a}_i^*)^2\right] da_i = \sqrt{2\pi}\{V_i^*(\theta)\}^{1/2}.$$

- Thus,

$$\begin{aligned} \ell(\theta \mid \lambda) &\cong \sum_{i=1}^{m} Q_\lambda(\hat{a}_i^*, \theta) + \frac{1}{2}\sum_{i=1}^{m} \log\{V_i^*(\theta)\} + C \\ &= \sum_{i=1}^{m} Q_\lambda(\hat{a}_i^*, \theta) - \frac{1}{2}\sum_{i=1}^{m} \log\{-\ddot{Q}_\lambda(\hat{a}_i^*, \theta)\} + C \end{aligned}$$

Example (Normal random effects model)

- For model random effects model in (1), we can express

$$\mathbf{y}_i \sim N\left(X_i \boldsymbol{\beta}, V_i \sigma_e^2\right)$$

where

$$V_i = \lambda^{-1} \boldsymbol{J}_{n_i} + \boldsymbol{I}_{n_i}$$

where $\lambda = \sigma_e^2 / \sigma_a^2$.

- Thus, since λ is known, we know $V_i = V_i(\lambda)$.
- GLS estimator

$$\hat{\boldsymbol{\beta}}_\lambda = \left(\sum_i X_i' V_i^{-1} X_i \right)^{-1} \sum_i X_i' V_i^{-1} \mathbf{y}_i. \tag{3}$$

Tuning parameter selection: How to find the right model?

- Wish to balance the trade-off in the model selection by finding the best λ^* that minimizes the predictive risk.
- How to find a good model?
  1. Sample split approach:
     1. Estimate the predictive risk directly by 10-fold cross validation (for each λ).
     2. Choose λ^* with the smallest 10-fold CV.
  2. Marginal likelihood approach

Sample Split approach

Idea

1. Split the sample into two parts: training sample and test sample
2. Use the training sample to estimate θ for each λ.
3. Use the test sample to evaluate the performance of $\hat{\theta}_\lambda$ by computing the empirical risk function in terms of λ
4. Choose the optimal value of λ minimizing the empirical risk as the final choice.

To make the best use of the data, we can compute the average of the empirical risk by K-fold cross validation.

Marginal likelihood approach

- Recall that the observed likelihood is a function of θ and λ.
- We can treat θ as a nuisance parameter and integrate out over θ:

$$L_m(\lambda) = \int L(\theta, \lambda) d\theta \tag{4}$$

where $L(\theta, \lambda)$ is the likelihood function using the density of the marginal distribution of $\mathbf{y}$. That is,

$$L(\theta, \lambda) = \prod_{i=1}^{K} \frac{1}{\sqrt{2\pi |V_i(\lambda)\sigma_e^2|}} \exp\left\{ -\frac{1}{2\sigma_e^2} (\mathbf{y}_i - \mathbf{x}_i' \boldsymbol{\beta})' \{V_i(\lambda)\}^{-1} (\mathbf{y}_i - \mathbf{x}_i' \boldsymbol{\beta}) \right\}$$

and $V_i(\lambda)$ is a function of λ.

- The actual computation for $L_m(\lambda)$ in (4) may involve Laplace approximation. (next page)

Computing the marginal likelihood using Laplace approximation

- We wish to compute

$$L_m(\lambda) = \int L(\theta, \lambda) d\theta = \int \exp\{\ell(\theta, \lambda)\} d\theta.$$

- Apply the second order Taylor expansion to get

$$\ell(\theta, \lambda) \cong \ell(\hat{\theta}_\lambda, \lambda) - \frac{1}{2} I_{11}(\hat{\theta}_\lambda, \lambda)(\hat{\theta}_\lambda - \theta)^2,$$

where

$$\hat{\theta}_\lambda = \arg\max_\theta \ell(\theta, \lambda)$$

and

$$I_{11}(\theta, \lambda) = -\frac{\partial^2}{\partial \theta^2} \ell(\theta, \lambda).$$

- Thus, we obtain

$$L_m(\lambda) \cong \int L(\hat{\theta}_\lambda, \lambda) \exp\left\{-\frac{1}{2} I_{11}(\hat{\theta}_\lambda, \lambda)(\hat{\theta}_\lambda - \theta)^2\right\} d\theta.$$

- Now, using

$$\int \exp\left\{-\frac{1}{2} I_{11}(\hat{\theta}_\lambda, \lambda)(\hat{\theta}_\lambda - \theta)^2\right\} d\theta = (2\pi)^{p/2} \left| I_{11}(\hat{\theta}_\lambda, \lambda)\right|^{-1/2},$$

we have the following approximation for $\ell_m(\lambda) = \log L_m(\lambda)$:

$$\ell_m(\lambda) \cong \ell(\hat{\theta}_\lambda, \lambda) - \frac{1}{2} \log \left| I_{11}(\hat{\theta}_\lambda, \lambda)\right| + C. \tag{5}$$

- The approximation in (5) is also called the modified profile likelihood as the second term is a modification term for the profile log-likelihood term $\ell(\hat{\theta}_\lambda, \lambda)$.

Remark

- Modified profile likelihood in (5) consists of two terms.
  1. Profile log-likelihood: $\ell_p(\lambda) = \ell(\hat{\theta}_\lambda, \lambda)$
  2. Penalty term: the "undeserved" information on the nuisance parameter θ.
- Small values of λ means less smoothing, which increases the profile log-likelihood term but its penalty term also increases.
- Thus, including the penalty term prevents over-fitting.
- The largest value of λ in $\ell_m(\lambda)$ will be selected.
- Closely related to BIC of Schwarz (1978).

Return to Random Effects Model

- Linear model expression

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}$$

with

$$\boldsymbol{u} \sim N(\mathbf{0}, V\sigma_e^2)$$

- Thus, $V = V(\lambda)$.
- The overall likelihood is

$$\log L(\theta, \lambda) = -\frac{1}{2}\log\left|V\sigma_e^2\right| - \frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})'(V\sigma_e^2)^{-1}(\mathbf{y} - X\boldsymbol{\beta})\,.$$

- Given λ, the MLE of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_\lambda = \left(X'V_\lambda^{-1}X\right)^{-1}X'V_\lambda^{-1}\mathbf{y}.$$

Also, the MLE of σ_e^2 can be obtained as a function of λ.

Return to Random Effects Model

- The profile likelihood of λ is

$$\log L_p(\lambda) = -\frac{1}{2}\log\left|V_\lambda \hat{\sigma}_e^2\right| - \frac{1}{2}(\mathbf{y} - X\hat{\boldsymbol{\beta}}_\lambda)'(V_\lambda \hat{\sigma}_e^2)^{-1}(\mathbf{y} - X\hat{\boldsymbol{\beta}}_\lambda).$$

- The modified profile likelihood is

$$\log L_m(\lambda) = \log L_p(\lambda) - \frac{1}{2}\log\left|X'(V_\lambda \hat{\sigma}_e^2)^{-1}X\right|.$$

- The maximizer of the modified profile likelihood matches exactly with the so-called restricted maximum likelihood estimator, which is derived using the marginal distribution of the error term $\mathbf{y} - X\hat{\boldsymbol{\beta}}_\lambda$.
- First proposed by Patterson and Thompson (1971) and discussed by Harville (1977).

REFERENCES

Harville, D. (1977), 'Maximum likelihood approaches to variance component estimation', *Journal of the American Statistical Association* **72**, 320–340.

Patterson, H. D. and R. Thompson (1971), 'Recovery of inter-block information when block sises are unequal', *Biometrika* **58**, 545–554.

Schwarz, E. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**, 461–464.