Statistical Methods for Handling Incomplete Data
Chapter 3.3: EM algorithm (Part 2)

Jae-Kwang Kim

Latent variable models

- Mixture models
- Random effects models

Example 3.8 (Mixture model)

- Latent variable:

$$Z_i \sim Bernoulli\left(\pi\right).$$

- Model specification for Y_i:

$$\begin{aligned} Y_i \mid (z_i = 1) &\sim N\left(\mu_1, \sigma_1^2\right) \\ Y_i \mid (z_i = 0) &\sim N\left(\mu_2, \sigma_2^2\right) \end{aligned}$$

- Parameter of interest: $\theta = \left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi\right)$
- Observed likelihood

$$L_{\rm obs}\left(\theta\right) = \prod_{i=1}^{n} \left\{\pi\phi\left(y_i \mid \mu_1, \sigma_1^2\right) + (1-\pi)\phi\left(y_i \mid \mu_2, \sigma_2^2\right)\right\}$$

where

$$\phi\left(y \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right].$$

Example 3.8 (Cont'd)

Complete-sample likelihood

$$L_{\text{com}}(\theta) = \prod_{i=1}^{n} \text{pdf}(y_i, z_i \mid \theta)$$

where

$$\text{pdf}(y, z \mid \theta) = \left[\phi\left(y \mid \mu_1, \sigma_1^2\right)\right]^{z} \left[\phi\left(y \mid \mu_2, \sigma_2^2\right)\right]^{1-z} \pi^{z} (1-\pi)^{1-z}.$$

Thus,

$$\begin{aligned} \ln L_{com}(\theta) &= \sum_{i=1}^{n} \left[z_i \ln \phi\left(y_i \mid \mu_1, \sigma_1^2\right) + (1-z_i) \ln \phi\left(y_i \mid \mu_2, \sigma_2^2\right)\right] \\ &\quad + \sum_{i=1}^{n} \left\{z_i \ln(\pi) + (1-z_i) \ln(1-\pi)\right\} \end{aligned}$$

Example 3.8 (Cont'd)

[E-step]

$$
\begin{aligned}
Q\left(\theta \mid \theta^{(t)}\right) &= \sum_{i=1}^{n}\left[w_i^{(t)} \ln \phi\left(y_i \mid \mu_1, \sigma_1^2\right) + (1 - w_i^{(t)}) \ln \phi\left(y_i \mid \mu_2, \sigma_2^2\right)\right] \\
&\quad + \sum_{i=1}^{n}\left\{w_i^{(t)} \ln(\pi) + \left(1 - w_i^{(t)}\right) \ln(1 - \pi)\right\}
\end{aligned}
$$

where $w_i^{(t)} = E\left(z_i \mid y_i, \theta^{(t)}\right)$ with

$$
E\left(z_i \mid y_i, \theta\right) = \frac{\pi\phi\left(y_i \mid \mu_1, \sigma_1^2\right)}{\pi\phi\left(y_i \mid \mu_1, \sigma_1^2\right) + (1 - \pi)\phi\left(y_i \mid \mu_2, \sigma_2^2\right)}
$$

[M-step]

$$\frac{\partial}{\partial \theta} Q\left(\theta \mid \theta^{(t)}\right) = 0.$$

Thus,

$$\begin{aligned}
\mu_j^{(t+1)} &= \sum_{i=1}^{n} w_{ij}^{(t)} y_i \Big/ \sum_{i=1}^{n} w_{ij}^{(t)}, \\
\sigma_j^{2(t+1)} &= \sum_{i=1}^{n} w_{ij}^{(t)} \left(y_i - \mu_j^{(t+1)}\right)^2 \Big/ \sum_{i=1}^{n} w_{ij}^{(t)}, \\
\pi^{(t+1)} &= \sum_{i=1}^{n} w_i^{(t)} / n,
\end{aligned}$$

for $j = 1, 2,$ where $w_{i1}^{(t)} = w_i^{(t)}$ and $w_{i2}^{(t)} = 1 - w_i^{(t)}$.

Extensions

- Let $\mathbf{Y} = (Y_1, \cdots, Y_p)$ be a p-dimensional random vector with density

$$f(\mathbf{y}; \pi, \theta) = \sum_{k=1}^{K} \pi_k f(\mathbf{y} \mid z = k, \theta_k) \tag{1}$$

  where $\pi_k = P(z = k)$ satisfies $0 < \pi_1 < \pi_2 < \cdots < \pi_K < 1$ and $\sum_{k=1}^{K} \pi_k = 1$.
- If we assume Gaussian model for $f(\mathbf{y} \mid z = k, \theta_k)$, then model (1) is often called the Gaussian Mixture Model (GMM).
- Here, K is a hyperparameter (or tuning parameter) that determines the level of model complexity.

Parameter Estimation

Estimation of (π, θ) for given K:

- [E-step] Using the current parameter values $(\pi^{(t)}, \theta^{(t)})$, compute

$$p_{ik}^{(t)} = \frac{f(\mathbf{y}_i \mid z_i = k, \theta_k^{(t)})\pi_k^{(t)}}{\sum_{k=1}^{K} f(\mathbf{y}_i \mid z_i = k, \theta_k^{(t)})\pi_k^{(t)}}.$$

- [M-step] Update the parameters by maximizing

$$Q(\pi, \theta \mid \pi^{(t)}, \theta^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} p_{ik}^{(t)} \left\{ \log(\pi_k) + \log f(\mathbf{y}_i \mid z_i = k, \theta_k) \right\}$$

with respect to π and θ.

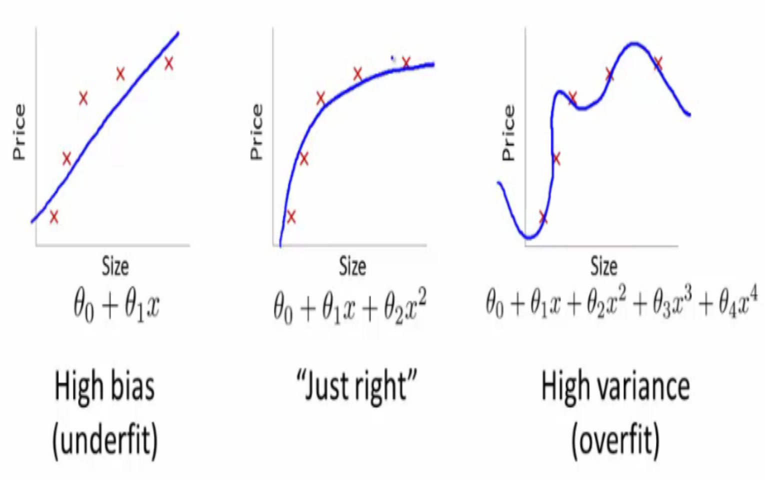
Choice of K

- Goal: Let $\hat{y}_K(x_i)$ be the predictor of y at $x = x_i$ using the training sample with GMM model indexed by tuning parameter K. We are interested in determining K that minimizes the predictive risk:

$$R(K) = E\left\{ \frac{1}{|S_{new}|} \sum_{i \in S_{new}} (y_i - \hat{y}_i(K))^2 \right\},$$

  where the expectation is for the observations for future prediction.
- Bias-variance trade-off
  1. Under-fitting: Bias $\uparrow$, Variance $\downarrow$
  2. Over-fitting: Bias $\downarrow$, Variance $\uparrow$

Overfit vs Underfit



How to find the right model?

- Wish to balance the trade-off in the model selection by finding the best K^* that minimizes the predictive risk.
- How to find a good model?
  1. Direct method:
     1. Estimate the predictive risk directly by 10-fold cross validation (for each K).
     2. Choose K^* with the smallest 10-fold CV.
  2. Use some information criteria: AIC, BIC etc.

2. Random effect model

James-Stein Theorem

James and Stein (1961)

Suppose that

$$x_i \sim N(\mu_i, \sigma^2)$$

independently for $i = 1, 2, \ldots, n$, with $n \geq 4$ and a known σ^2. Then

$$\sum_{i=1}^{n} E\left\{(\hat{\mu}_i^{JS} - \mu_i)^2\right\} < \sum_{i=1}^{n} E\left\{(\hat{\mu}_i^{MLE} - \mu_i)^2\right\},$$

for all choices of μ_i, where

$$\hat{\mu}_i^{JS} = \bar{x} + \hat{B}\left(x_i - \bar{x}\right)$$

and $\hat{B} = 1 - \frac{(n-3)\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Baseball player example (Efron and Hastie, 2016)

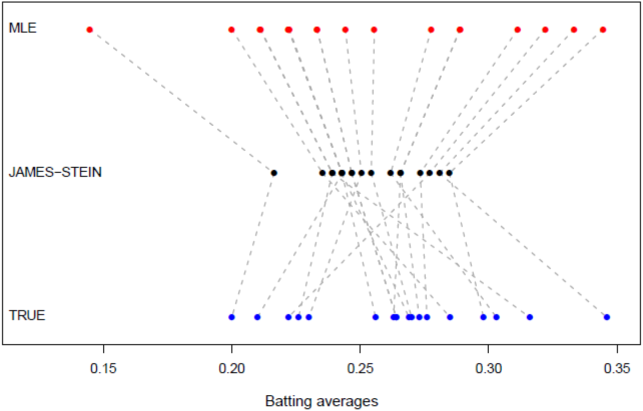


Figure 7.1 Eighteen baseball players; top line MLE, middle James–Stein, bottom true values. Only 13 points are visible, since there are ties.

Remark

- The James-Stein estimator is a shrinkage estimator. It is a weighted average of the MLEs for two different models.

$$\hat{\mu}_i^{JS} = (1 - \hat{B})\hat{\mu}_i^{MLE,r} + \hat{B}\hat{\mu}_i^{MLE,f}$$

  where $\hat{\mu}_i^{MLE,r} = \bar{x}$ is the MLE under the reduced model $\mu_1 = \cdots = \mu_N$ and $\hat{\mu}_i^{MLE,f} = x_i$ is the MLE under the full model.
- The James-Stein estimator has a Bayesian interpretation.

$$\begin{aligned} x_i \mid \mu_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &\sim N(\xi, \tau^2) \end{aligned}$$

  lead to

$$\mu_i \mid x_i \sim N[\xi + B(x_i - \xi), B\sigma^2] \tag{2}$$

  where $B = \tau^2/(\tau^2 + \sigma^2)$.

Justification for (2)

- Posterior is proportional to likelihood times prior

$$
\begin{aligned}
p(\mu_i \mid x_i) &\propto f(x_i \mid \mu_i)\pi(\mu_i) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu_i)^2\right\} \cdot \exp\left\{-\frac{1}{2\tau^2}(\mu_i - \zeta)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)\mu_i^2 + \left(\frac{x_i}{\sigma^2} + \frac{\xi}{\tau^2}\right)\mu_i\right\}
\end{aligned}
$$

- The above density can be viewed as a density function of μ:

$$
p(\mu_i \mid x_i) \propto \exp\left\{-\frac{1}{2V(\mu_i \mid x_i)}(\mu_i - E(\mu_i \mid x_i))^2\right\}
$$

where

$$
\begin{aligned}
E(\mu_i \mid x_i) &= \frac{x_i/\sigma^2 + \xi/\tau^2}{1/\sigma^2 + 1/\tau^2} \\
V(\mu_i \mid x_i) &= \frac{1}{1/\sigma^2 + 1/\tau^2}
\end{aligned}
$$

Random Effect Model (Example 3.3)

- Consider random effect model

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + a_i + e_{ij}, \quad i = 1, \cdots, m, j = 1, \cdots, n_i, \tag{3}$$

where $a_i \sim N(0, \sigma_a^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$.

- The two error terms are independent of each other. $\mathbf{x}_{ij}$ are fixed.
- Let $\mathbf{y}_i = (y_{i1}, \cdots, y_{in_i})'$ be observed, but a_i is never observed (i.e. latent).
- Random effects model is useful in describing the clustered structure of the data.
- For example, i= subject, j= repetition

Marginal model expression

- We can express (3) as

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta} + \boldsymbol{u}_i, \tag{4}$$

  where $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{in_i})'$, $\boldsymbol{u}_i = (u_{i1}, \cdots, u_{in_i})'$ and $u_{ij} = a_i + e_{ij}$.
- Note that

$$Cov\left(u_{ij}, u_{ik}\right) = \begin{cases} \sigma_a^2 + \sigma_e^2 & \text{if } j = k \\ \sigma_a^2 & \text{otherwise.} \end{cases}$$

- Thus, we have

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

  where

$$V_i = (\sigma_a^2/\sigma_e^2)\boldsymbol{J}_{n_i} + \boldsymbol{I}_{n_i}$$

  and $\boldsymbol{J}_i$ is a $n_i \times n_i$ matrix of 1's.
- Model (4) is called marginal model while (3) is called conditional model.

Two-level models: general form

- Conditional model
  - Level 1 model

$$\mathbf{y}_i \mid a_i \sim f_1(\mathbf{y}_i \mid \mathbf{x}_i, a_i; \theta_1)$$

  - Level 2 model

$$a_i \sim f_2(a_i; \theta_2)$$

- Marginal model

$$\mathbf{y}_i \sim \int f_1(\mathbf{y}_i \mid \mathbf{x}_i, a_i; \theta_1) f_2(a_i; \theta_2) da_i.$$

The marginal model is a mixture model which take the average of f_1 over the distribution of the nuisance parameter a_i. Using Bayesian framework, f_2 plays the role of the prior distribution of a_i.

Marginal distribution under model (3)

- The marginal model is obtained by integrating out over the latent variable a_i:

$$f_m(\mathbf{y}_i \mid \mathbf{x}_i; \theta) \propto \int \exp\left\{-\frac{1}{2\sigma_e^2}\sum_{j=1}^{n_i}\left(y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta} - a_i\right)^2 - \frac{1}{2\sigma_a^2}a_i^2\right\} da_i. \tag{5}$$

- Thus, the observed likelihood function derived from the marginal density can be written as

$$L_{\text{obs}}(\theta) = \prod_{i=1}^{K} f_m(\mathbf{y}_i \mid \mathbf{x}_i; \theta) \propto \prod_{i=1}^{K} \int \exp\left\{-\frac{1}{2\sigma_e^2} Q_\lambda(a_i, \boldsymbol{\beta})\right\} da_i, \tag{6}$$

where

$$Q_\lambda(a_i, \boldsymbol{\beta}) = \sum_{j=1}^{n_i}\left(y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta} - a_i\right)^2 + \lambda a_i^2 \tag{7}$$

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

Remark 1

- The first term of (7) can be written as

$$\sum_{j=1}^{n_i}\left(y_{ij}-\mathbf{x}_{ij}'\boldsymbol{\beta}-a_i\right)^2 = \sum_{j=1}^{n_i}\left\{y_{ij}-\mathbf{x}_{ij}'\boldsymbol{\beta}-(\bar{y}_i-\bar{\mathbf{x}}_i'\boldsymbol{\beta})\right\}^2 + n_i\{(\bar{y}_i-\bar{\mathbf{x}}_i'\boldsymbol{\beta})-a_i\}^2.$$

- Thus, we can express (7) as

$$Q_\lambda(a_i,\boldsymbol{\beta}) = Q^{(1)}(\boldsymbol{\beta}) + Q_\lambda^{(2)}(a_i \mid \boldsymbol{\beta}), \tag{8}$$

where

$$Q^{(1)}(\boldsymbol{\beta}) = \sum_{j=1}^{n_i}\{(y_{ij}-\bar{y}_i)-(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)'\boldsymbol{\beta}\}^2$$

$$Q_\lambda^{(2)}(a_i \mid \boldsymbol{\beta}) = n_i(\bar{y}_i-\bar{\mathbf{x}}_i'\boldsymbol{\beta}-a_i)^2+\lambda {a_i}^2.$$

- The optimal value of a_i minimizing $Q_\lambda(a_i, \boldsymbol{\beta})$ can be obtaining by minimizing $Q_\lambda^{(2)}(a_i \mid \boldsymbol{\beta})$ with respect to a_i. The solution is

$$\hat{a}_i^* = \frac{n_i}{n_i + \lambda}(\bar{y}_i - \bar{\mathbf{x}}_i' \boldsymbol{\beta}). \tag{9}$$

- The optimization using $Q_\lambda(a_i, \boldsymbol{\beta})$ takes the form of the penalized regression problem where parameter $\lambda = \sigma_e^2/\sigma_a^2$ serves the role of tuning parameter in the shrinkage estimation.
- The tuning parameter represents a trade-off between fidelity to the data and "smoothness" of the solution.
  - If $\lambda \to 0$, then $\hat{a}_i^* = \bar{y}_i - \bar{\mathbf{x}}_i'\boldsymbol{\beta}$ and

$$\hat{y}_{ij}^* = \bar{y}_i + (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'\, \boldsymbol{\beta}$$

  - If $\lambda \to \infty$, then $\hat{a}_i^* = 0$ and

$$\hat{y}_{ij}^* = \mathbf{x}_{ij}'\boldsymbol{\beta}$$

Parameter estimation

- Two main approaches
  1. Direct ML: maximize the observed likelihood using the marginal density function
  2. EM algorithm: treat a_i as a latent variable
- We can estimate (θ_1, θ_2) jointly, or separately.
- For the separate estimation, $\lambda = \theta_2$ plays the role of tuning parameter and may use cross-validation after sample splits. (Use the training sample to estimate θ_1 for given λ and use the validation sample to estimate λ.)

EM algorithm for joint estimation

- Define

$$Q(\theta \mid \theta^{(t)}) = \sum_i E\left\{\log f_1(\mathbf{y}_i \mid a_i; \boldsymbol{\beta}, \sigma_e^2) + \log f_2(a_i; \sigma_a^2) \mid \mathbf{y}_i; \theta^{(t)}\right\}$$

where

$$\begin{aligned} f_1(\mathbf{y}_i \mid a_i; \boldsymbol{\beta}, \sigma_e^2) &= (2\pi\sigma_e^2)^{-n_i/2} \exp\left\{-\frac{1}{2\sigma_e^2}\sum_j (y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta} - a_i)^2\right\} \\ f_2(a_i; \sigma_a^2) &= (2\pi\sigma_a^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma_a^2} a_i^2\right\}. \end{aligned}$$

- EM algorithm find the MLE by an iterative algorithm:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(t)}) \tag{10}$$

until convergence.

- The solution in (10) is often obtained by solving

$$E\{S_{\text{com}}(\theta) \mid \mathbf{y}; \theta^{(t)}\} = 0$$

where $S_{\text{com}}(\theta)$ is the score function of θ treating a_i as if observed.

Complete-sample Score functions

$$
\begin{aligned}
S_{\mathrm{com},1}(\theta) &= \sum_i \sum_j \left(y_{ij} - \mathbf{x}_{ij}' \boldsymbol{\beta} - a_i \right) \mathbf{x}_{ij} / \sigma_e^2 \\
S_{\mathrm{com},2}(\theta) &= \frac{1}{2\sigma_a^4} \sum_i \left(a_i^2 - \sigma_a^2 \right) \\
S_{\mathrm{com},3}(\theta) &= \frac{1}{2\sigma_e^4} \sum_i \sum_j \left\{ \left(y_{ij} - \mathbf{x}_{ij}' \boldsymbol{\beta} - a_i \right)^2 - \sigma_e^2 \right\}.
\end{aligned}
$$

- EM algorithm:
  - E-step: Compute the conditional expectation of the score functions given the observed data:
    $$E\{S_{\mathrm{com}}(\theta) \mid \mathbf{y}; \hat{\theta}^{(t)}\}.$$
    When both f_1 and f_2 are normal, then the above conditional distribution is also normal
    $$a_i \mid \boldsymbol{y}_i \sim N\left(\tau_i\left(\bar{y}_i - \bar{\mathbf{x}}_i'\boldsymbol{\beta}\right), \sigma_a^2(1-\tau_i)\right), \tag{11}$$
    where $\tau_i = n_i/(n_i+\lambda)$ and $\lambda = \sigma_e^2/\sigma_a^2$.
  - M-step: Update the parameter by solving
    $$E\{S_{\mathrm{com}}(\theta) \mid \mathbf{y}; \hat{\theta}^{(t)}\} = 0$$
    for θ, where the conditional expectation is computed from the E-step.
- If the normality does not hold either in f_1 or in f_2, then (11) is not necessarily normal. In this case, E-step may involve Monte Carlo approximation.

Remark 2 (advanced topic)

- Now, let's evaluate the marginal distribution in (5).
- In (5), we wish to evaluate the integral:

$$\int \exp\left\{-\frac{1}{2\sigma_e^2}\sum_{j=1}^{n_i}\left(y_{ij}-\mathbf{x}_{ij}'\beta-a_i\right)^2-\frac{1}{2\sigma_a^2}a_i^2\right\}da_i.$$

  That is, we wish to get rid of nuisance parameter a_i by taking the average of f_1 using the prior distribution f_2.

- Note that the conditional distribution $a_i \mid (\mathbf{x}_i, \mathbf{y}_i)$ is also normal with mean $\hat{a}_i^* = E(a_i \mid \mathbf{x}_i, \mathbf{y}_i)$ in (9) and variance V_i^*, where

$$\begin{aligned} V_i^* &= V(\hat{a}_i^* - a_i) \\ &= V\left\{\frac{n_i}{n_i+\lambda}(a_i+\bar{e}_i)-a_i\right\} \\ &= \frac{1}{n_i+\lambda}\sigma_e^2 \end{aligned}$$

- Now, the marginal density is computed by the joint density divided by the conditional density:

$$f(\mathbf{y}_i \mid \mathbf{x}_i) = \frac{f_1(\mathbf{y}_i \mid \mathbf{x}_i, a_i) f_2(a_i)}{f(a_i \mid \mathbf{x}_i, \mathbf{y}_i)} \tag{12}$$

- Taking logarithm on (12), we have

$$\begin{aligned}
& \log f(\mathbf{y}_i \mid \mathbf{x}_i) \\
= \; & \log f_1(\mathbf{y}_i \mid \mathbf{x}_i, a_i) + \log f_2(a_i) - \log f(a_i \mid \mathbf{x}_i, \mathbf{y}_i) \\
= \; & -\frac{1}{2\sigma_e^2} \left\{ \sum_{j=1}^{n_i} \left(y_{ij} - \mathbf{x}_{ij}' \boldsymbol{\beta} - a_i \right)^2 + \lambda a_i^2 - \frac{\sigma_e^2}{V_i^*} \left(a_i - \hat{a}_i^* \right)^2 \right\} + C
\end{aligned}$$

This term should be free of a_i.

- If either f_1 or f_2 is not Gaussian, the computation is not exact. But, we may use normal approximation in the denominator of (12).
- That is,

$$f(\mathbf{y}_i \mid \mathbf{x}_i) \cong \frac{f_1(\mathbf{y}_i \mid \mathbf{x}_i, a_i) f_2(a_i)}{f(a_i \mid \mathbf{x}_i, \mathbf{y}_i)} \tag{13}$$

and

$$f(a_i \mid \mathbf{x}_i, \mathbf{y}_i) = \frac{1}{\sqrt{2\pi V_i^*}} \exp\left\{-\frac{1}{2V_i^*}(a_i - \hat{a}_i^*)^2\right\}$$

- If we insert $a_i = \hat{a}_i^*$ into (13), we obtain

$$f(\mathbf{y}_i \mid \mathbf{x}_i) \cong \frac{f_1(\mathbf{y}_i \mid \mathbf{x}_i, \hat{a}_i^*) f_2(\hat{a}_i^*)}{f(\hat{a}_i^* \mid \mathbf{x}_i, \mathbf{y}_i)} = f_1(\mathbf{y}_i \mid \mathbf{x}_i, \hat{a}_i^*) f_2(\hat{a}_i^*) \cdot \sqrt{2\pi V_i^*}.$$

This is essentially Laplace approximation of

$$f(\mathbf{y}_i \mid \mathbf{x}_i) = \int f_1(\mathbf{y}_i \mid \mathbf{x}_i, a_i) f_2(a_i) da_i.$$

- The (approximate) marginal density function can be used to compute the observed likelihood.

REFERENCES

Efron, B. and T. Hastie (2016), *Computer Age Statistical Inference*, Cambridge University Press.

James, W. and C. Stein (1961), Estimation with quadratic loss, *in* 'Proceedings of the 4-th Berkeley Symposium on Mathematical Statistics and Probability', pp. 361–379.