Example Class 5

Department of Statistics and Actuarial Science,
The University of Hong Kong

October 25, 2022

Bayesian Model Averaging (BMA)

- Assume that we have K candidate models, $M_1, \ldots, M_K$ to fit the data D.
- $P(M_k)$ is the prior probability that M_k is the true model.
- For example, you can assign an equal weight to all candidate models, i.e., $P(M_k) = 1/K$ for $k = 1, \ldots, K$.
- Each model M_k has its own parameter $\boldsymbol{\theta}_k$, and let $\pi(\boldsymbol{\theta_k}|M_k)$ be the prior of $\boldsymbol{\theta}_k$.
- The posterior probability of M_k has the form,

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{j=1}^{K} P(D|M_j)P(M_j)}$$

$$P(D|M_k) = \int \pi(\boldsymbol{\theta}_k|M_k) f(D|\boldsymbol{\theta}_k, M_k) d\boldsymbol{\theta}_k$$

BMA Estimator

- For the model parameter $\boldsymbol{\theta}$, the BMA estimator is given by

$$\bar{\boldsymbol{\theta}} = \sum_{k=1}^{K} \hat{\boldsymbol{\theta}}_k P(M_k|D)$$

- Here each $\hat{\boldsymbol{\theta}}_k$ is the posteior mean of $\boldsymbol{\theta}_k$,

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_k &= \int \boldsymbol{\theta}_k f(\boldsymbol{\theta}_k|D, M_k) d\boldsymbol{\theta}_k \\
&= \int \boldsymbol{\theta}_k \frac{\pi(\boldsymbol{\theta}_k|M_k) f(D|\boldsymbol{\theta}_k, M_k)}{\int \pi(\boldsymbol{\theta}_k|M_k) f(D|\boldsymbol{\theta}_k, M_k) d\boldsymbol{\theta}_k} d\boldsymbol{\theta}_k
\end{aligned}$$

- By assigning the posterior mean $\hat{\boldsymbol{\theta}}_k$ a weight of $P(M_k|D)$, BMA automatically lean toward the best fitting model, and thus $\bar{\boldsymbol{\theta}}$ will be close to the best parameter estimate

- If T is the quantity of interest,

$$f(T|D) = \sum_{k=1}^{K} f(T|D, M_k) P(M_k|D)$$

Why BMA?

- Frequentist
  - Model selection
  - Regularization

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\min_{\boldsymbol{\beta}} \quad (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

- Bayesian
  - BMA
  - Marginalization

$$f(T|D) = \sum_{k=1}^{K} f(T|D, M_k)P(M_k|D)$$

$$P(D|M_k) = \int \pi(\boldsymbol{\theta}_k|M_k)f(D|\boldsymbol{\theta}_k, M_k)d\boldsymbol{\theta}_k$$

Why BMA?

- Averaging over all of the models provides better predictive ability, as measured by a logarithmic scoring rule, than using any single model M_j

$$-E\left[\log\left\{\sum_{k=1}^{K} P(T|D, M_k)P(M_k|D)\right\}\right] \leq -E[\log P(T|D, M_j)]$$

- In decision theory, a score function, or scoring rule, measures the accuracy of probabilistic predictions.

BMA Linear Regression

- Simple linear regression

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \sigma^2 \boldsymbol{I}$$

- With $q+1$ predictors (including the intercept term), overall you have $K = 2^{(q+1)}$ possible models $\{M_1, \ldots, M_K\}$
- Simplified case including one predictor without intercept

$$y_i = \beta_\gamma x_{i\gamma} + \epsilon_i$$

BMA Example

- Linear regression with only one predictor

$$y_i = \beta_\gamma x_{i\gamma} + \epsilon_i,\ \ \epsilon_i \sim N(0, \sigma^2)$$

- Overall q predictors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q$
- The j-th Model M_j includes the j-th predictor $\boldsymbol{X}_j$
- Prior

$$\begin{aligned}
\boldsymbol{Y}|M_j, \beta_j, \boldsymbol{X}_j, \sigma^2 &\sim N(\boldsymbol{X}_j\beta_j, \sigma^2\boldsymbol{I}) \\
\beta_j|\sigma^2, \mu_0, \lambda_0 &\sim N(\mu_0, \sigma^2/\lambda_0) \\
\sigma^2|a_0, b_0 &\sim \text{InverseGamma}(a_0, b_0) \\
P(M_j) &= 1/q
\end{aligned}$$

Normal-Inverse-Gamma Prior

- The normal-inverse-gamma distribution (or Gaussian-inverse-gamma distribution) is a four-parameter family of bivariate continuous probability distributions.
- It is the conjugate prior of a normal distribution with unknown mean and variance.
- (x, σ^2) has a normal-inverse-gamma distribution with parameter $(\mu, \lambda, \alpha, \beta)$ if

$$x|\sigma^2, \mu, \lambda \sim N(\mu, \sigma^2/\lambda)$$
$$\sigma^2|\alpha, \beta \sim IG(\alpha, \beta)$$

$$f(x, \sigma^2|\mu, \lambda, \alpha, \beta) = \frac{\sqrt{\lambda}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\sigma^2\right)^{-\alpha-1} e^{-\frac{2\beta+\lambda(x-\mu)^2}{2\sigma^2}}$$

BMA Example

$$
\begin{aligned}
f(D|\beta_k, \sigma^2, M_k) &\propto \left(\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{\sum_{i=1}^n (y_i - \beta_k x_{ik})^2}{2\sigma^2}} \\
f(\beta_k, \sigma^2|D, M_k) &\propto \left(\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{\sum_{i=1}^n (y_i - \beta_k x_{ik})^2}{2\sigma^2}} \left(\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{\lambda_0(\beta_k - \mu_0)^2}{2\sigma^2}} \left(\sigma^2\right)^{-a_0-1} e^{-\frac{b_0}{\sigma^2}} \\
&\sim \text{NIG}\left(\frac{\lambda_0\mu_0 + \sum_{i=1}^n y_i x_{ik}}{\lambda_0 + \sum_{i=1}^n x_{ik}^2}, \lambda_0 + \sum_{i=1}^n x_{ik}^2, a_0 + \frac{n}{2}, b_{\text{new}}\right) \\
b_{\text{new}} &= b_0 + \frac{1}{2}\sum_{i=1}^n y_i^2 + \frac{\lambda_0\mu_0^2 \sum_{i=1}^n x_{ik}^2 - \left(\sum_{i=1}^n y_i x_{ik}\right)^2 - 2\lambda_0\mu_0 \sum_{i=1}^n y_i x_{ik}}{2(\lambda_0 + \sum_{i=1}^n x_{ik}^2)}
\end{aligned}
$$

BMA Example

$$
\begin{aligned}
P(D|M_k) &= \int \pi(\beta_k, \sigma^2 | M_k) f(D|\beta_k, \sigma^2, M_k) d(\beta_k, \sigma^2) \\
&= \int \frac{\sqrt{\lambda_0}}{\sigma\sqrt{2\pi}} \frac{b_0^{a_0}}{\Gamma(a_0)} \left(\sigma^2\right)^{-a_0-1} e^{-\frac{2b_0+\lambda_0(\beta_k-\mu_0)^2}{2\sigma^2}} \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{\frac{\sum_{i=1}^n (y_i - x_{ik}\beta_k)^2}{2\sigma^2}} d(\beta_k, \sigma^2) \\
&= \sqrt{\lambda_0}\,(2\pi)^{-\frac{n+1}{2}} \frac{b_0^{a_0}}{\Gamma(a_0)} \int \left(\sigma^2\right)^{-\frac{n+1}{2}-a_0-1} e^{-\frac{\sum_{i=1}^n (y_i-\beta_k x_{ik})^2}{2\sigma^2}} e^{-\frac{2b_0+\lambda_0(\beta_k-\mu_0)^2}{2\sigma^2}} d(\beta_k, \sigma^2) \\
&= \frac{\sqrt{\lambda_0}}{\sqrt{\lambda_0 + \sum_{i=1}^n x_{ik}^2}} (2\pi)^{-\frac{n}{2}} \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_0+n)}{b_{\text{new}}^{a_0+\frac{n}{2}}}
\end{aligned}
$$

Or you can calculate $P(D|M_k)$ via the Monte Carlo method for integration.

BMA Example

- Five covariates:
  - $(X_1, X_2) \sim N_2\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} 1 & 0.8 \\ 0.8 & 1 \end{array}\right)\right)$
  - $X_3 \sim N(0, 1)$
  - $X_4 \sim \text{Bernoulli}(0.5) - 0.5$
  - $X_5 = 2 \cdot X_1$
- True model: $y = \beta_{\text{true}} X_1 + \epsilon,\ \beta_{\text{true}} = 1,\ \epsilon \sim N(0, 1)$
- Prior: $P(M_k) = 1/5,\ \mu_0 = \beta_{\text{true}} = 1,\ \lambda_0 = 0.1,\ a_0 = b_0 = 0.1$
- Posteior Model Probability $P(M_k|D)$ for different sample sizes n

BMA Example

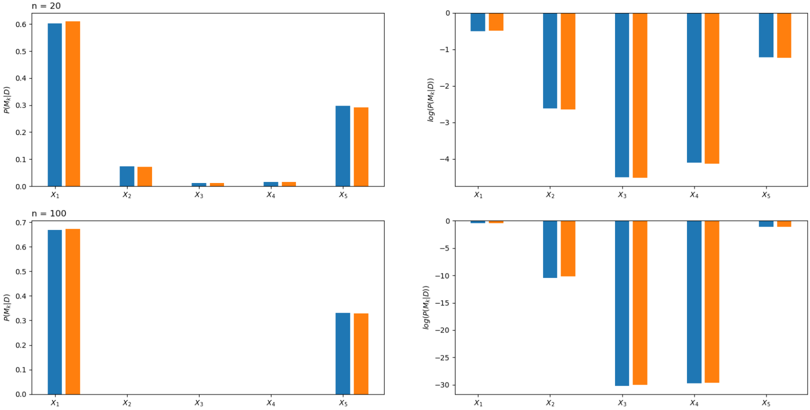

Figure 1: Barplots of $P(M_k|D)$ for $n = 20$ (left) and 100 (right).

Multivariate Linear Regression

- Multivariate Linear Regression

$$
\begin{aligned}
y_{i,1} &= \boldsymbol{x}_i^T \boldsymbol{\beta}_1 + \epsilon_{i,1}, \\
&\dots \\
y_{i,m} &= \boldsymbol{x}_i^T \boldsymbol{\beta}_m + \epsilon_{i,m}.
\end{aligned}
$$

- The sets of errors $\boldsymbol{\epsilon}_i = \{\epsilon_{i,1}, \dots, \epsilon_{i,m}\}$ are correlated.
- $\boldsymbol{y}_i^T = \boldsymbol{x}_i^T \boldsymbol{B} + \boldsymbol{\epsilon}_i$; $\boldsymbol{y}_i = \{y_{i,1}, \dots, y_{i,m}\}$; $\boldsymbol{B} = [\beta_{i,j}]_{p \times m}$
- We assume that $\boldsymbol{\epsilon}_i$ is jointly normal, i.e., $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$
- Regression problem in matrix form:

$$
\underset{(n \times m)}{\boldsymbol{Y}} = \underset{(n \times p)}{\boldsymbol{X}} \underset{(p \times m)}{\boldsymbol{B}} + \underset{(n \times m)}{\boldsymbol{E}},
$$

Bayesian Multivariate Linear Regression

- Model: $\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}, \ \epsilon_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$
- Observations: $\boldsymbol{Y}, \boldsymbol{X}$
- Parameters of interest: $\boldsymbol{B}, \boldsymbol{\Sigma}_\epsilon$
- Likelihood:

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{B}, \boldsymbol{\Sigma}_\epsilon) \propto |\boldsymbol{\Sigma}_\epsilon|^{-\frac{n}{2}} e^{-\frac{1}{2}\left\{\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{x}_i^T\boldsymbol{B})^T \boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i^T\boldsymbol{B})\right\}}$$
$$\propto |\boldsymbol{\Sigma}_\epsilon|^{-\frac{n}{2}} e^{-\frac{1}{2}\mathrm{tr}((\boldsymbol{Y}-\boldsymbol{XB})^T(\boldsymbol{Y}-\boldsymbol{XB})\boldsymbol{\Sigma}_\epsilon^{-1})}$$

- Priors:

$$\pi(\boldsymbol{B}, \boldsymbol{\Sigma}_\epsilon) = \pi(\boldsymbol{B}|\boldsymbol{\Sigma}_\epsilon)\pi(\boldsymbol{\Sigma}_\epsilon),$$
$$\pi(\boldsymbol{B}|\boldsymbol{\Sigma}_\epsilon) \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_\epsilon/\lambda_0)$$
$$\pi(\boldsymbol{\Sigma}_\epsilon) \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}_0, \nu_0)$$

Inverse Wishart Distribution

- The inverse Wishart distribution ($\mathcal{W}^{-1}$) is used as the conjugate prior for the covariance matrix of a multivariate normal distribution.
- For $\boldsymbol{\Sigma}_\epsilon \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$, its pdf is,

$$f_p(\boldsymbol{\Sigma}_\epsilon | \boldsymbol{\Psi}, \nu) = \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{\nu p/2}\Gamma_p(\nu/2)} |\boldsymbol{\Sigma}_\epsilon|^{-(\nu+p+1)/2} e^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}_\epsilon^{-1})}$$

  where $\boldsymbol{\Sigma}_\epsilon$ and $\boldsymbol{\Psi}$ are $p \times p$ positive definite matrices, $|\cdot|$ is the determinant, and Γ_p is the multivariate gamma function.
- If $\boldsymbol{\Sigma}_\epsilon \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$, its inverse $\boldsymbol{\Sigma}_\epsilon^{-1}$ has a Wishart distribution $\mathcal{W}(\boldsymbol{\Psi}^{-1}, \nu)$.

Normal-Inverse-Wishart (NIW) Distribution

- The normal-inverse-Wishart distribution is a multivariate four-parameter family of continuous probability distributions.
- We say $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim NIW(\boldsymbol{\mu}_0, \lambda, \boldsymbol{\Psi}, \nu)$ if

$$
\begin{aligned}
f(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{\mu}_0, \lambda, \boldsymbol{\Psi}, \nu) = N(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\lambda)\mathcal{W}^{-1}(\boldsymbol{\Sigma}|\boldsymbol{Psi}, \nu) \\
\boldsymbol{\mu}|\boldsymbol{\mu}_0, \lambda, \boldsymbol{\Psi} \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\lambda) \\
\boldsymbol{\Sigma}|\boldsymbol{\Psi}, \nu \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)
\end{aligned}
$$

- It is the conjugate prior of a multivariate normal distribution with unknown mean and covariance matrix.

$$\pi(\boldsymbol{B}, \boldsymbol{\Sigma}_\epsilon | \boldsymbol{Y}, \boldsymbol{X}) \tag{1}$$

$$\propto \boldsymbol{\Sigma}_\epsilon|^{-\frac{n}{2}} e^{-\frac{1}{2}\mathrm{tr}((\boldsymbol{Y}-\boldsymbol{XB})^T(\boldsymbol{Y}-\boldsymbol{XB})\boldsymbol{\Sigma}_\epsilon^{-1})} \times \qquad \text{likelihood} \tag{2}$$

$$\boldsymbol{\Sigma}_\epsilon|^{-\frac{p}{2}} e^{-\frac{\lambda_0}{2}\mathrm{tr}((\boldsymbol{B}-\boldsymbol{B}_0)^T(\boldsymbol{B}-\boldsymbol{B}_0)\boldsymbol{\Sigma}_\epsilon^{-1})} \times \qquad \text{prior for B, } B_0 \ is \ \mu_0 \tag{3}$$

$$\boldsymbol{\Sigma}_\epsilon|^{-\frac{\nu_0+m+1}{2}} e^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}_0\boldsymbol{\Sigma}_\epsilon^{-1})} \qquad \text{prior for } \boldsymbol{\Sigma}_\epsilon \tag{4}$$

$$\propto |\boldsymbol{\Sigma}_\epsilon|^{-\frac{p+m}{2}} e^{-\frac{1}{2}tr\left((\boldsymbol{B}-\boldsymbol{B}_n)^T(\lambda_0\boldsymbol{I}+\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{B}-\boldsymbol{B}_n)\boldsymbol{\Sigma}_\epsilon^{-1}\right)} \times \qquad \text{MatrixNormal} \tag{5}$$

$$\boldsymbol{\Sigma}_\epsilon|^{-\frac{\nu_0+n+1}{2}} e^{-\frac{1}{2}tr\left((\boldsymbol{\Psi}_0+(\boldsymbol{Y}-\boldsymbol{XB}_n)^T(\boldsymbol{Y}-\boldsymbol{XB}_n)+\lambda_0(\boldsymbol{B}_0-\boldsymbol{B}_n)^T(\boldsymbol{B}_0-\boldsymbol{B}_n))\boldsymbol{\Sigma}_\epsilon^{-1}\right)} \qquad \text{inverse-Wishart} \tag{6}$$

where $\boldsymbol{B}_n = (\lambda_0\boldsymbol{I} + \boldsymbol{X}^T\boldsymbol{X})^{-1}(\lambda_0\boldsymbol{B}_0 + \boldsymbol{X}^T\boldsymbol{Y})$

$\text{Psi\_new} = \boldsymbol{\Psi}_0 + (\boldsymbol{Y} - \boldsymbol{XB}_n)^T(\boldsymbol{Y} - \boldsymbol{XB}_n) + \lambda_0(\boldsymbol{B}_0 - \boldsymbol{B}_n)^T(\boldsymbol{B}_0 - \boldsymbol{B}_n)$

$\text{MatrixNormal}(\boldsymbol{X}|\boldsymbol{X}_0, \boldsymbol{U}, \boldsymbol{V}) \propto |\boldsymbol{U}|^{-d_V/2}|\boldsymbol{V}|^{-d_U/2}e^{-\frac{1}{2}tr\left(\boldsymbol{V}^{-1}(\boldsymbol{X}-\boldsymbol{X}_0)^T\boldsymbol{U}^{-1}(\boldsymbol{X}-\boldsymbol{X}_0)\right)}$

$V \ shape : (d_V, d_V) \quad U \ shape : (d_U, d_U)$

Compare formula (5),(6) with the definition of MatrixNormal and $\mathcal{W}^{-1}$ we get

$\pi(\boldsymbol{B}, \boldsymbol{\Sigma}_\epsilon|\boldsymbol{Y}, \boldsymbol{X}) \propto \mathcal{W}^{-1}(\nu_0 + n, \text{Psi\_new}) \cdot \text{MatrixNormal}(\boldsymbol{B}_n, (\lambda_0\boldsymbol{I} + \boldsymbol{X}^T\boldsymbol{X})^{-1}, \boldsymbol{\Sigma}_\epsilon)$

BMLR (Improper Prior)

- So

$$\boldsymbol{\Sigma}_\epsilon | \boldsymbol{Y}, \boldsymbol{X} \sim \mathcal{W}^{-1}(\nu_0 + n, \text{Psi\_new})$$

$$\boldsymbol{B} | \boldsymbol{Y}, \boldsymbol{X} \sim \text{MatrixNormal}(\boldsymbol{B}_n, (\lambda_0 \boldsymbol{I} + \boldsymbol{X}^T \boldsymbol{X})^{-1}, \boldsymbol{\Sigma}_\epsilon)$$

  In the code, we will draw samples for $\boldsymbol{\Sigma}_\epsilon$ and $\boldsymbol{B}$ from this two posterior distribution, and get the estimated value of $\boldsymbol{\Sigma}_\epsilon$ and $\boldsymbol{B}$.

- For simplicity, we can consider an improper prior for $(\boldsymbol{B}, \boldsymbol{\Sigma}_\epsilon)$

$$\pi(\boldsymbol{B}, \boldsymbol{\Sigma}_\epsilon) \propto \boldsymbol{\Sigma}_\epsilon^{-\frac{m+p+1}{2}}$$

$$\pi(\boldsymbol{B}, \boldsymbol{\Sigma}_\epsilon | \boldsymbol{Y}, \boldsymbol{X}) \propto |\boldsymbol{\Sigma}_\epsilon|^{-\frac{n}{2}} e^{-\frac{1}{2}\text{tr}((\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}) \boldsymbol{\Sigma}_\epsilon^{-1})} |\boldsymbol{\Sigma}_\epsilon|^{-\frac{m+p+1}{2}}$$

BMLR (Example)

- Model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}, \quad \boldsymbol{B} = \begin{bmatrix} 1 & 2 \\ 2 & 0.5 \end{bmatrix}$$

$$\boldsymbol{E} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon), \;\; \boldsymbol{\Sigma}_\epsilon = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$$

- Prior:

$$\boldsymbol{\mu}_0 = \mathbf{0}, \;\; \lambda_0 = 0.01,$$
$$\boldsymbol{\Psi}_0 = \mathbf{0}, \;\; \nu_0 = 0.01.$$

BMLR Example

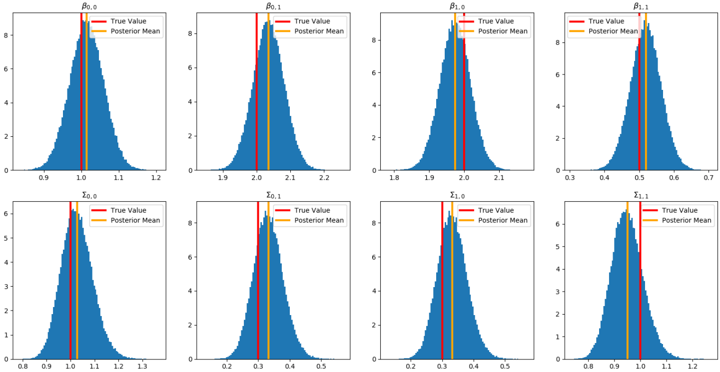

Figure 2: Histograms of posterior samples of $\boldsymbol{B}$ (upper) and $\boldsymbol{\Sigma}_\epsilon$ (lower).