Chapter 3 Bayesian Methods for Estimating Structural Equation Models

Notations and concepts related to the Bayesian approach:

- M — an arbitrary SEM with a vector of unknown parameters $\boldsymbol{\theta}$.
- $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$ — the observed data set
- $p(\boldsymbol{\theta}|M)$ (or $p(\theta)$) — the prior density function of $\boldsymbol{\theta}$.
- $p(\mathbf{Y}, \boldsymbol{\theta}|M)$ — the probability density function of the joint distribution of $\mathbf{Y}$ and $\boldsymbol{\theta}$ under M.
- $p(\boldsymbol{\theta}|\mathbf{Y}, M)$ — the density function of the posterior distribution of $\boldsymbol{\theta}$ under M. This function fully describes the behavior of $\boldsymbol{\theta}$ under the given data $\mathbf{Y}$.
- $p(\mathbf{Y}|\boldsymbol{\theta}, M)$ — the likelihood function under M.

Based on a well-known identity in probability, we have

$$p(\mathbf{Y}, \boldsymbol{\theta}|M) = p(\mathbf{Y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{Y}, M)p(\mathbf{Y}|M).$$

As $p(\mathbf{Y}|M)$ does not depend on $\boldsymbol{\theta}$, and can be regarded as a constant with fixed $\mathbf{Y}$, we have

$$p(\boldsymbol{\theta}|\mathbf{Y}, M) \propto p(\mathbf{Y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}), \quad \text{or}$$
$$\log p(\boldsymbol{\theta}|\mathbf{Y}, M) = \log p(\mathbf{Y}|\boldsymbol{\theta}, M) + \log p(\boldsymbol{\theta}) + \text{constant}.$$

Key steps in the Bayesian estimation:

1. Specify prior distributions.
2. Derive posterior distributions.
3. Sample from posterior distributions using MCMC algorithms.
4. Conduct sensitivity analysis.

1. Specify prior distributions.

The most commonly used prior distributions are informative priors (i.e., $p(\theta)$ and $p(\theta|y)$ have the same form).

We consider the following model:

$$\begin{aligned}\mathbf{y}_i &= \mathbf{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \\ \boldsymbol{\eta}_i &= \mathbf{B}\mathbf{d}_i + \mathbf{\Pi}\boldsymbol{\eta}_i + \mathbf{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\delta}_i = \mathbf{\Lambda}_\omega \mathbf{v}_i + \boldsymbol{\delta}_i,\end{aligned}$$

where $\boldsymbol{\omega}_i = (\boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T$, $\mathbf{\Lambda}$, $\mathbf{B}$, $\mathbf{\Pi}$, $\mathbf{\Gamma}$ are parameter matrices of unknown regression coefficients, $\mathbf{\Lambda}_\omega = (\mathbf{B}, \mathbf{\Pi}, \mathbf{\Gamma})$, and $\mathbf{v}_i = (\mathbf{d}_i^T, \boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T$. The distributions of $\boldsymbol{\xi}_i$, $\boldsymbol{\epsilon}_i$, and $\boldsymbol{\delta}_i$ are $N[\mathbf{0}, \mathbf{\Phi}]$, $N[\mathbf{0}, \mathbf{\Psi}_\epsilon]$, and $N[\mathbf{0}, \mathbf{\Psi}_\delta]$.

Let $\mathbf{\Lambda}_k^T$ be the kth row of $\mathbf{\Lambda}$, $\psi_{\epsilon k}$ be the kth diagonal element of $\mathbf{\Psi}_\epsilon$, $\mathbf{\Lambda}_{\omega k}^T$ be the kth row of $\mathbf{\Lambda}_\omega$, and $\psi_{\delta k}$ be the kth diagonal element of $\mathbf{\Psi}_\delta$. Then, the conjugate prior distributions are as follows:

$$\psi_{\epsilon k} \overset{D}{=} IG[\alpha_{0\epsilon k}, \beta_{0\epsilon k}] \text{ or } \psi_{\epsilon k}^{-1} \overset{D}{=} Gamma[\alpha_{0\epsilon k}, \beta_{0\epsilon k}],$$

$$[\mathbf{\Lambda}_k | \psi_{\epsilon k}] \overset{D}{=} N\left[\mathbf{\Lambda}_{0k}, \psi_{\epsilon k}\mathbf{H}_{0yk}\right],$$

$$\mathbf{\Phi} \overset{D}{=} IW_{q_2}[\mathbf{R}_0^{-1}, \rho_0], \text{ or equivalently } \mathbf{\Phi}^{-1} \overset{D}{=} W_{q_2}[\mathbf{R}_0, \rho_0],$$

$$\psi_{\delta k} \overset{D}{=} IG[\alpha_{0\delta k}, \beta_{0\delta k}] \text{ or } \psi_{\delta k}^{-1} \overset{D}{=} Gamma[\alpha_{0\delta k}, \beta_{0\delta k}],$$

$$[\mathbf{\Lambda}_{\omega k} | \psi_{\delta k}] \overset{D}{=} N[\mathbf{\Lambda}_{0\omega k}, \psi_{\delta k}\mathbf{H}_{0\omega k}].$$

Appendix 1: The Gamma, Inverted Gamma, Wishart, and Inverted Wishart distributions:

1. *Gamma distribution:* $\theta \overset{D}{=} \textit{Gamma}[\alpha, \beta]$

$$p(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta},$$
$$E(\theta) = \alpha/\beta, \quad Var(\theta) = \alpha/\beta^2.$$

2. *Inverted Gamma distribution:* $\theta \overset{D}{=} \textit{Inverted Gamma}[\alpha, \beta]$

$$p(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\beta/\theta},$$
$$E(\theta) = \frac{\beta}{\alpha - 1}, \quad Var(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-1)}.$$

3. *Relation between Gamma and inverted Gamma distributions*

If $\theta \overset{D}{=} \textit{Inverted Gamma}[\alpha, \beta]$, then $\theta^{-1} \overset{D}{=} \textit{Gamma}[\alpha, \beta]$.

4. *Wishart distribution:* $\mathbf{W} \overset{D}{=} Wishart_q[\mathbf{R}_0, \rho_0]$ $\mathbf{R}_0 \in \mathbb{R}^{q \times q}$

$$p(\mathbf{W}) = \left[2^{\rho_0 q/2} \pi^{q(q-1)/4} \prod_{i=1}^{q} \Gamma\left(\frac{\rho_0 + 1 - i}{2}\right)\right]^{-1} \times |\mathbf{R}_0|^{-\rho_0/2} \times |\mathbf{W}|^{(\rho_0 - q - 1)/2} \times \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{R}_0^{-1}\mathbf{W})\right\},$$

$$E(\mathbf{W}) = \rho_0 \mathbf{R}_0.$$

5. *Inverted Wishart distribution:* $\mathbf{W} \overset{D}{=} IW_q[\mathbf{R}_0^{-1}, \rho_0]$

$$p(\mathbf{W}) = \left[2^{\rho_0 q/2} \pi^{q(q-1)/4} \prod_{i=1}^{q} \Gamma\left(\frac{\rho_0 + 1 - i}{2}\right)\right]^{-1} \times |\mathbf{R}_0|^{-\rho_0/2} \times |\mathbf{W}|^{-(\rho_0 + q + 1)/2} \times \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{R}_0^{-1}\mathbf{W}^{-1})\right\},$$

$$E(\mathbf{W}) = \frac{\mathbf{R}_0^{-1}}{\rho_0 - q - 1}.$$

6. *Relation between Wishart and inverted Wishart distributions*

$$\text{If } \mathbf{W} \overset{D}{=} IW[\mathbf{R}_0^{-1}, \rho_0], \quad \text{then } \mathbf{W}^{-1} \overset{D}{=} W[\mathbf{R}_0, \rho_0].$$

The choice of hyperparameters:

1. If we have good prior information about $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_\omega$, then choose $\mathbf{\Lambda}_{0k}$ and $\mathbf{\Lambda}_{0\omega k}$ close to the true values of $\mathbf{\Lambda}_k$ and $\mathbf{\Lambda}_{\omega k}$, along with small $\mathbf{H}_{0yk}$ and $\mathbf{H}_{0\omega k}$ (e.g., $0.5\mathbf{I}$ or $\mathbf{I}$, where $\mathbf{I}$ is the identity matrix).
2. If we have no good information about $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_\omega$, then choose $\mathbf{\Lambda}_{0k}$ and $\mathbf{\Lambda}_{0\omega k}$ at ad hoc values (e.g., zero elements), along with large $\mathbf{H}_{0yk}$ and $\mathbf{H}_{0\omega k}$ (e.g., $10\mathbf{I}$ or $100\mathbf{I}$).
3. For $\alpha_{0\epsilon k}$ and $\beta_{0\epsilon k}$ (similarly $\alpha_{0\delta k}$ and $\beta_{0\delta k}$), we may assign their values according to the rationale: If $\psi_{\epsilon k} \overset{D}{=} \textit{Inverted Gamma}(\alpha_{0\epsilon k}, \beta_{0\epsilon k})$, then

$$E(\psi_{\epsilon k}) = \beta_{0\epsilon k}/(\alpha_{0\epsilon k} - 1),$$
$$Var(\psi_{\epsilon k}) = \beta_{0\epsilon k}^2/\{(\alpha_{0\epsilon k} - 1)^2(\alpha_{0\epsilon k} - 2)\}.$$

4. For $\mathbf{R}_0$ and ρ_0, since $E(\mathbf{\Phi}) = \mathbf{R}_0^{-1}/(\rho_0 - q_2 - 1)$, we may choose $\mathbf{R}_0^{-1}$ and ρ_0 such that

$$\mathbf{R}_0^{-1} = (\rho_0 - q_2 - 1)\mathbf{\Phi}_0.$$

2. Derive posterior distributions.

$$\begin{cases} y = \Lambda\omega + \varepsilon \\ \eta = Bd + \pi\eta + \Gamma\xi + \delta \end{cases}$$

$$\omega_i = \begin{bmatrix} \eta_i \\ \xi_i \end{bmatrix}$$

$$Cov(\omega_i) = \begin{bmatrix} Var\,\eta_i & Cov(\eta_i, \xi_i) \\ Cov(\xi_i, \eta_i) & Var\,\xi_i \end{bmatrix}.$$

(1) *Conditional Distribution* $[\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\theta}]$

For $i = 1, \cdots, n$, as $\boldsymbol{\omega}_i$ are conditionally independent and $\mathbf{y}_i$ are also conditionally independent, then

$$p(\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta}).$$

It implies that the conditional distributions of $\boldsymbol{\omega}_i$ given $(\mathbf{y}_i, \boldsymbol{\theta})$ are mutually independent for different i, and

$$p(\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta}) = \frac{p(\boldsymbol{\omega}_i, \mathbf{y}_i|\boldsymbol{\theta})}{p(\mathbf{y}_i|\boldsymbol{\theta})} \propto p(\boldsymbol{\omega}_i|\boldsymbol{\theta}) p(\mathbf{y}_i|\boldsymbol{\omega}_i, \boldsymbol{\theta}).$$

Let $\boldsymbol{\Pi}_0 = \mathbf{I} - \boldsymbol{\Pi}$ and the covariance matrix of $\boldsymbol{\omega}_i$ be

$$\boldsymbol{\Sigma}_\omega = \begin{bmatrix} \boldsymbol{\Pi}_0^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \boldsymbol{\Psi}_\delta)\boldsymbol{\Pi}_0^{-T} & \boldsymbol{\Pi}_0^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}\boldsymbol{\Gamma}^T\boldsymbol{\Pi}_0^{-T} & \boldsymbol{\Phi} \end{bmatrix}.$$

It can be shown that

$$[\boldsymbol{\omega}_i|\boldsymbol{\theta}] \stackrel{D}{=} N\left[\begin{pmatrix}\boldsymbol{\Pi}_0^{-1}\mathbf{B}\mathbf{d}_i\\ \mathbf{0}\end{pmatrix}, \boldsymbol{\Sigma}_\omega\right],$$

and

$$[\mathbf{y}_i|\boldsymbol{\omega}_i, \boldsymbol{\theta}] \stackrel{D}{=} N[\boldsymbol{\Lambda}\boldsymbol{\omega}_i, \boldsymbol{\Psi}_\epsilon].$$

Let $\boldsymbol{\mu}_\omega = \begin{pmatrix}\boldsymbol{\Pi}_0^{-1}\mathbf{B}\mathbf{d}_i\\ \mathbf{0}\end{pmatrix}$. Then, we have

$$\begin{aligned}
& p(\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta}) \propto p(\mathbf{y}_i|\boldsymbol{\omega}_i, \boldsymbol{\theta})p(\boldsymbol{\omega}_i|\boldsymbol{\theta}) \\
\propto \quad & \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\Lambda}\boldsymbol{\omega}_i)^T\boldsymbol{\Psi}_\epsilon^{-1}(\mathbf{y}_i - \boldsymbol{\Lambda}\boldsymbol{\omega}_i)\right\} \times \\
& \exp\left\{-\frac{1}{2}(\boldsymbol{\omega}_i - \boldsymbol{\mu}_\omega)^T\boldsymbol{\Sigma}_\omega^{-1}(\boldsymbol{\omega}_i - \boldsymbol{\mu}_\omega)\right\} \\
\propto \quad & \exp\left\{-\frac{1}{2}\left[\mathbf{y}_i^T\boldsymbol{\Psi}_\epsilon^{-1}\mathbf{y}_i - 2\boldsymbol{\omega}_i^T\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_\epsilon^{-1}\mathbf{y}_i + \boldsymbol{\omega}_i^T(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_\epsilon^{-1}\boldsymbol{\Lambda})\boldsymbol{\omega}_i\right.\right. \\
& \left.\left. + \boldsymbol{\omega}_i^T\boldsymbol{\Sigma}_\omega^{-1}\boldsymbol{\omega}_i - 2\boldsymbol{\omega}_i^T\boldsymbol{\Sigma}_\omega^{-1}\boldsymbol{\mu}_\omega\right]\right\}
\end{aligned}$$

$$
\begin{aligned}
\propto \quad & \exp\Big\{ -\frac{1}{2}\big[-2\boldsymbol{\omega}_i^T(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_\epsilon^{-1}\mathbf{y}_i + \boldsymbol{\Sigma}_\omega^{-1}\boldsymbol{\mu}_\omega) + \\
& \boldsymbol{\omega}_i^T\big(\boldsymbol{\Sigma}_\omega^{-1} + \boldsymbol{\Lambda}^T\boldsymbol{\Psi}_\epsilon^{-1}\boldsymbol{\Lambda}\big)\boldsymbol{\omega}_i\big]\Big\} \\
\propto \quad & \exp\Big\{ -\frac{1}{2}\big[\boldsymbol{\omega}_i - \boldsymbol{\Sigma}^{*-1}(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_\epsilon^{-1}\mathbf{y}_i + \boldsymbol{\Sigma}_\omega^{-1}\boldsymbol{\mu}_\omega)\big]^T\boldsymbol{\Sigma}^* \\
& \qquad\qquad \big[\boldsymbol{\omega}_i - \boldsymbol{\Sigma}^{*-1}(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_\epsilon^{-1}\mathbf{y}_i + \boldsymbol{\Sigma}_\omega^{-1}\boldsymbol{\mu}_\omega)\big]\Big\}.
\end{aligned}
$$

Thus,

$$
[\boldsymbol{\omega}_i|\mathbf{y}_i,\boldsymbol{\theta}] \overset{D}{=} N\left[\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_\epsilon^{-1}\mathbf{y}_i + \boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_\omega^{-1}\begin{pmatrix}\boldsymbol{\Pi}_0^{-1}\mathbf{B}\mathbf{d}_i \\ \mathbf{0}\end{pmatrix}, \boldsymbol{\Sigma}^{*-1}\right],
$$

where $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_\omega^{-1} + \boldsymbol{\Lambda}^T\boldsymbol{\Psi}_\epsilon^{-1}\boldsymbol{\Lambda}$. The conditional distribution $[\boldsymbol{\omega}_i|\mathbf{y}_i,\boldsymbol{\theta}]$ is a normal distribution.

(2) *Conditional Distribution* $[\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Omega}]$

Note that

$$p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Omega}) \propto p(\boldsymbol{\theta})p(\mathbf{Y}, \boldsymbol{\Omega}|\boldsymbol{\theta}).$$

Let $\boldsymbol{\theta}_y$ be the unknown parameters in $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}_\epsilon$ associated with the measurement equation, and $\boldsymbol{\theta}_\omega$ be the unknown parameters in $\mathbf{B}$, $\boldsymbol{\Pi}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}_\delta$ associated with the structural equation.

Assumed that the prior distribution of $\boldsymbol{\theta}_y$ is independent of the prior distribution of $\boldsymbol{\theta}_\omega$, that is, $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_y)p(\boldsymbol{\theta}_\omega)$. Note also that

$$p(\mathbf{Y}|\boldsymbol{\Omega}, \boldsymbol{\theta}) = p(\mathbf{Y}|\boldsymbol{\Omega}, \boldsymbol{\theta}_y), \quad p(\boldsymbol{\Omega}|\boldsymbol{\theta}) = p(\boldsymbol{\Omega}|\boldsymbol{\theta}_\omega)$$

and

$$p(\boldsymbol{\theta}_y|\mathbf{Y}, \boldsymbol{\Omega}) \propto p(\mathbf{Y}|\boldsymbol{\Omega}, \boldsymbol{\theta}_y)p(\boldsymbol{\theta}_y), \quad p(\boldsymbol{\theta}_\omega|\mathbf{Y}, \boldsymbol{\Omega}) \propto p(\boldsymbol{\Omega}|\boldsymbol{\theta}_\omega)p(\boldsymbol{\theta}_\omega),$$

Hence, these conditional densities can be treated separately.

(a) $p(\boldsymbol{\theta}_y|\mathbf{Y},\boldsymbol{\Omega})$ or $p(\boldsymbol{\Lambda},\psi_{\epsilon k}|\mathbf{Y},\boldsymbol{\Omega})$:

For notation simplicity, we temporarily assume that all the elements of $\boldsymbol{\Lambda}_k$ are unknown. Let $\nu_k = \psi_{\epsilon k}^{-1}$, we have

$$p(\nu_k) \propto \nu_k^{\alpha_{0\epsilon k}-1} \exp(-\beta_{0\epsilon k}\nu_k),$$
$$p(\boldsymbol{\Lambda}_k|\nu_k) \propto \nu_k^{q/2} \exp\{-\frac{1}{2}(\boldsymbol{\Lambda}_k - \boldsymbol{\Lambda}_{k0})^T \mathbf{H}_{0yk}^{-1}(\boldsymbol{\Lambda}_k - \boldsymbol{\Lambda}_{k0})\nu_k\},$$

and the likelihood of $\mathbf{Y}$ is

$$p(\mathbf{Y}|\boldsymbol{\theta}_y,\boldsymbol{\Omega}) \propto |\boldsymbol{\Psi}_\epsilon|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\Lambda}\boldsymbol{\omega}_i)^T \boldsymbol{\Psi}_\epsilon^{-1}(\mathbf{y}_i - \boldsymbol{\Lambda}\boldsymbol{\omega}_i)\right\}.$$

Let $\mathbf{Y}_k^T$ be the kth row of $\mathbf{Y}$, y_{ik} be the ith component of $\mathbf{Y}_k^T$, $\mathbf{A}_k^* = (\mathbf{\Omega}\mathbf{\Omega}^T)^{-1}\mathbf{\Omega}\mathbf{Y}_k$, and $b_k = \mathbf{Y}_k^T\mathbf{Y}_k - \mathbf{Y}_k^T\mathbf{\Omega}^T(\mathbf{\Omega}\mathbf{\Omega}^T)^{-1}\mathbf{\Omega}\mathbf{Y}_k = \mathbf{Y}_k^T\mathbf{Y}_k - \mathbf{A}_k^{*T}(\mathbf{\Omega}\mathbf{\Omega}^T)\mathbf{A}_k^*$. The exponential term in $p(\mathbf{Y}|\boldsymbol{\theta}_y, \mathbf{\Omega})$ can be expressed as

$$-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{\Lambda}\boldsymbol{\omega}_i)^T\mathbf{\Psi}_\epsilon^{-1}(\mathbf{y}_i - \mathbf{\Lambda}\boldsymbol{\omega}_i) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{p}\nu_k(y_{ik} - \mathbf{\Lambda}_k^T\boldsymbol{\omega}_i)^2$$

$$= -\frac{1}{2}\sum_{k=1}^{p}\left\{\nu_k\Big[\sum_{i=1}^{n}y_{ik}^2 - 2\mathbf{\Lambda}_k^T\sum_{i=1}^{n}y_{ik}\boldsymbol{\omega}_i + tr(\mathbf{\Lambda}_k\mathbf{\Lambda}_k^T\sum_{i=1}^{n}\boldsymbol{\omega}_i\boldsymbol{\omega}_i^T)\Big]\right\}$$

$$= -\frac{1}{2}\sum_{k=1}^{p}\Big\{\nu_k\big[\mathbf{Y}_k^T\mathbf{Y}_k - 2\mathbf{\Lambda}_k^T\mathbf{\Omega}\mathbf{Y}_k + \mathbf{\Lambda}_k^T(\mathbf{\Omega}\mathbf{\Omega}^T)\mathbf{\Lambda}_k\big]\Big\}$$

$$= -\frac{1}{2}\sum_{k=1}^{p}\Big\{\nu_k\big[\mathbf{Y}_k^T\mathbf{Y}_k - \mathbf{Y}_k^T\mathbf{\Omega}^T(\mathbf{\Omega}\mathbf{\Omega}^T)^{-1}\mathbf{\Omega}\mathbf{Y}_k\big] +$$

$$\nu_k\big[\mathbf{\Lambda}_k - (\mathbf{\Omega}\mathbf{\Omega}^T)^{-1}\mathbf{\Omega}\mathbf{Y}_k\big]^T(\mathbf{\Omega}\mathbf{\Omega}^T)\big[\mathbf{\Lambda}_k - (\mathbf{\Omega}\mathbf{\Omega}^T)^{-1}\mathbf{\Omega}\mathbf{Y}_k\big]$$

$$= -\frac{1}{2}\sum_{k=1}^{p}\Big\{\nu_k\big[b_k + (\mathbf{\Lambda}_k - \mathbf{A}_k^*)^T(\mathbf{\Omega}\mathbf{\Omega}^T)(\mathbf{\Lambda}_k - \mathbf{A}_k^*)\big]\Big\}.$$

Thus,

$$
\begin{aligned}
p(\boldsymbol{\theta}_y|\mathbf{Y},\boldsymbol{\Omega}) &= p(\boldsymbol{\Psi}_\epsilon)p(\boldsymbol{\Lambda}|\boldsymbol{\Psi}_\epsilon)p(\mathbf{Y}|\boldsymbol{\theta}_y,\boldsymbol{\Omega}) \\
&\propto \prod_{k=1}^{p}\Bigg\{\nu_k^{n/2+q/2+\alpha_{0\epsilon k}-1}\exp\Big[-\frac{1}{2}\nu_k\big\{(\boldsymbol{\Lambda}_k-\mathbf{A}_k^*)^T(\boldsymbol{\Omega}\boldsymbol{\Omega}^T)(\boldsymbol{\Lambda}_k-\mathbf{A}_k^*) \\
&\quad +(\boldsymbol{\Lambda}_k-\boldsymbol{\Lambda}_{0k})^T\mathbf{H}_{0yk}^{-1}(\boldsymbol{\Lambda}_k-\boldsymbol{\Lambda}_{0k})\big\}-\nu_k(\beta_{0\epsilon k}+b_k/2)\Big]\Bigg\} \\
&= \prod_{k=1}^{p} p(\boldsymbol{\Lambda}_k,\nu_k|\mathbf{Y},\boldsymbol{\Omega}).
\end{aligned}
$$

To derive $p(\boldsymbol{\theta}_y|\mathbf{Y},\boldsymbol{\Omega})$, it is sufficient to derive $p(\boldsymbol{\Lambda}_k,\nu_k|\mathbf{Y},\boldsymbol{\Omega})$. Let $\mathbf{A}_k = (\mathbf{H}_{0yk}^{-1}+\boldsymbol{\Omega}\boldsymbol{\Omega}^T)^{-1}$ and $\mathbf{a}_k = \mathbf{A}_k(\mathbf{H}_{0yk}^{-1}\boldsymbol{\Lambda}_{0k}+\boldsymbol{\Omega}\mathbf{Y}_k)$, it follows that

$$
\begin{aligned}
&(\boldsymbol{\Lambda}_k-\mathbf{A}_k^*)^T(\boldsymbol{\Omega}\boldsymbol{\Omega}^T)(\boldsymbol{\Lambda}_k-\mathbf{A}_k^*)+(\boldsymbol{\Lambda}_k-\boldsymbol{\Lambda}_{0k})^T\mathbf{H}_{0yk}^{-1}(\boldsymbol{\Lambda}_k-\boldsymbol{\Lambda}_{0k}) \\
&= (\boldsymbol{\Lambda}_k-\mathbf{a}_k)^T\mathbf{A}_k^{-1}(\boldsymbol{\Lambda}_k-\mathbf{a}_k)-\mathbf{a}_k^T\mathbf{A}_k^{-1}\mathbf{a}_k \\
&\qquad +\mathbf{A}_k^{*T}\boldsymbol{\Omega}\boldsymbol{\Omega}^T\mathbf{A}_k^*+\boldsymbol{\Lambda}_{0k}^T\mathbf{H}_{0yk}^{-1}\boldsymbol{\Lambda}_{0k}.
\end{aligned}
$$

Hence,

$$p(\mathbf{\Lambda}_k, \nu_k|\mathbf{Y}, \mathbf{\Omega}) = p(\nu_k|\mathbf{Y}, \mathbf{\Omega})p(\mathbf{\Lambda}_k|\mathbf{Y}, \mathbf{\Omega}, \nu_k)$$

$$\propto \left\{\nu_k^{n/2+\alpha_{0\epsilon k}-1} \exp(-\beta_{\epsilon k}\nu_k)\right\}\left\{\nu_k^{q/2} \exp\left[-\frac{1}{2}(\mathbf{\Lambda}_k - \mathbf{a}_k)^T \mathbf{A}_k^{-1}(\mathbf{\Lambda}_k - \mathbf{a}_k)\nu_k\right]\right\}$$

where $\beta_{\epsilon k} = \beta_{0\epsilon k} + \frac{1}{2}(\mathbf{Y}_k^T\mathbf{Y}_k - \mathbf{a}_k^T\mathbf{A}_k^{-1}\mathbf{a}_k + \mathbf{\Lambda}_{0k}^T\mathbf{H}_{0yk}^{-1}\mathbf{\Lambda}_{0k})$.

Finally, we can conclude that $p(\mathbf{\Lambda}_k, \nu_k|\mathbf{Y}, \mathbf{\Omega})$ is the following normal-gamma distribution:

$$[\nu_k|\mathbf{Y}, \mathbf{\Omega}] \stackrel{D}{=} Gamma[n/2 + \alpha_{0\epsilon k}, \beta_{\epsilon k}], \quad \text{and}$$

$$[\mathbf{\Lambda}_k|\mathbf{Y}, \mathbf{\Omega}, , \nu_k] \stackrel{D}{=} N[\mathbf{a}_k, \nu_k^{-1}\mathbf{A}_k],$$

If some elements of $\mathbf{\Lambda}_k$ are fixed, we identify the positions of the fixed elements via an index matrix $\mathbf{L}$ with the following elements:

$$l_{kj} = \begin{cases} 0, & \text{if } \lambda_{kj} \text{ is fixed,} \\ 1, & \text{if } \lambda_{kj} \text{ is free;} \end{cases} \quad \text{for} \quad j = 1, \cdots, q \quad \text{and} \quad k = 1, \cdots, p.$$

Let $\mathbf{\Lambda}_k^*$ be a vector of unknown parameters in $\mathbf{\Lambda}_k$, $\mathbf{Y}_k$ be the submatrix of $\mathbf{Y}$ such that all the rows corresponding to $l_{kj} = 0$ are deleted; and let $\mathbf{Y}_k^{*T} = (y_{1k}^*, \cdots, y_{nk}^*)$ with

$$y_{ik}^* = y_{ik} - \sum_{j=1}^{q} \lambda_{kj} y_{ij} (1 - l_{kj}).$$

where y_{ij} is the jth element of $\mathbf{y}_i$. Then,

$$[\nu_k | \mathbf{Y}, \mathbf{\Omega}] \stackrel{D}{=} \ Gamma[n/2 + \alpha_{0\epsilon k}, \beta_{\epsilon k}], \ [\mathbf{\Lambda}_k^* | \mathbf{Y}, \mathbf{\Omega}, \nu_k] \stackrel{D}{=} N[\mathbf{a}_k, \nu_k \mathbf{A}_k],$$

where $\mathbf{A}_k = (\mathbf{H}_{0yk}^{-1} + \mathbf{Y}_k \mathbf{Y}_k^T)^{-1}$, $\mathbf{a}_k = \mathbf{A}_k (\mathbf{H}_{0yk}^{-1} \mathbf{\Lambda}_{0yk} + \mathbf{\Omega} \mathbf{Y}_k^*)$, and $\beta_{\epsilon k} = \beta_{0\epsilon k} + \frac{1}{2} (\mathbf{Y}_k^{*T} \mathbf{Y}_k^* - \mathbf{a}_k^T \mathbf{A}_k^{-1} \mathbf{a}_k + \mathbf{\Lambda}_{0k}^T \mathbf{H}_{0yk}^{-1} \mathbf{\Lambda}_{0k})$.

(b) $p(\boldsymbol{\theta}_\omega|\mathbf{Y}, \boldsymbol{\Omega})$

Note that

$$p(\boldsymbol{\theta}_\omega|\mathbf{Y}, \boldsymbol{\Omega}) \propto p(\boldsymbol{\Omega}|\boldsymbol{\theta}_\omega)p(\boldsymbol{\theta}_\omega).$$

Let $\boldsymbol{\Omega}_1 = (\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_n)$ and $\boldsymbol{\Omega}_2 = (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n)$. Since the distribution of $\boldsymbol{\xi}_i$ only involves $\boldsymbol{\Phi}$, $p(\boldsymbol{\Omega}_2|\boldsymbol{\theta}_\omega) = p(\boldsymbol{\Omega}_2|\boldsymbol{\Phi})$. Under the assumption that the prior distribution of $\boldsymbol{\Phi}$ is independent of the prior distributions of $\mathbf{B}, \boldsymbol{\Pi}, \boldsymbol{\Gamma}$, and $\boldsymbol{\Psi}_\delta$, we have

$$p(\boldsymbol{\Omega}|\boldsymbol{\theta}_\omega)p(\boldsymbol{\theta}_\omega) = [p(\boldsymbol{\Omega}_1|\boldsymbol{\Omega}_2, \mathbf{B}, \boldsymbol{\Pi}, \boldsymbol{\Gamma}, \boldsymbol{\Psi}_\delta)p(\mathbf{B}, \boldsymbol{\Pi}, \boldsymbol{\Gamma}, \boldsymbol{\Psi}_\delta)][p(\boldsymbol{\Omega}_2|\boldsymbol{\Phi})p(\boldsymbol{\Phi})].$$

Hence, the conditional distributions of $(\mathbf{B}, \boldsymbol{\Pi}, \boldsymbol{\Gamma}, \boldsymbol{\Psi}_\delta)$ and $\boldsymbol{\Phi}$ can be treated separately.

Consider a conjugate type prior distribution for $\boldsymbol{\Phi}$ with $\boldsymbol{\Phi} \stackrel{D}{=} IW_{q_2}[\mathbf{R}_0^{-1}, \rho_0]$ or $\boldsymbol{\Phi}^{-1} \stackrel{D}{=} W_{q_2}[\mathbf{R}_0, \rho_0]$, with hyperparameters ρ_0 and $\mathbf{R}_0^{-1}$ or $\mathbf{R}_0$.

To derive $p(\mathbf{\Phi}|\mathbf{\Omega}_2)$, we first note that it is proportional to $p(\mathbf{\Phi})p(\mathbf{\Omega}_2|\mathbf{\Phi})$. As $\boldsymbol{\xi}_i$ are independent, we have

$$p(\mathbf{\Phi}|\mathbf{\Omega}_2) \propto p(\mathbf{\Phi}) \prod_{i=1}^{n} p(\boldsymbol{\xi}_i|\boldsymbol{\theta}).$$

Moreover, since the distribution of $\boldsymbol{\xi}_i$ given $\mathbf{\Phi}$ is $N(\mathbf{0}, \mathbf{\Phi})$, we have

$$p(\mathbf{\Phi}|\mathbf{\Omega}_2) \propto \left[|\mathbf{\Phi}|^{-(\rho_0+q_2+1)/2} \exp\left\{ -\frac{1}{2}\mathrm{tr}[\mathbf{R}_0^{-1}\mathbf{\Phi}^{-1}] \right\} \right] \times$$

invert Wisher dist.

$$\left[|\mathbf{\Phi}|^{-n/2} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{\xi}_i^T \mathbf{\Phi}^{-1} \boldsymbol{\xi}_i \right\} \right]$$

$$= |\mathbf{\Phi}|^{-(n+\rho_0+q_2+1)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}[\mathbf{\Phi}^{-1}(\mathbf{\Omega}_2\mathbf{\Omega}_2^T + \mathbf{R}_0^{-1})] \right\}.$$

Since the right hand side is proportional to the density function of an inverted Wishart distribution (Zellner, 1971), it follows that the conditional distribution of $\mathbf{\Phi}$ given $\mathbf{\Omega}_2$ is given by

$$[\mathbf{\Phi}|\mathbf{\Omega}_2] \stackrel{D}{=} IW_{q_2}[(\mathbf{\Omega}_2\mathbf{\Omega}_2^T + \mathbf{R}_0^{-1}), n+\rho_0]$$

Recall that

$$\boldsymbol{\eta}_i = \boldsymbol{\Lambda}_\omega \mathbf{v}_i + \boldsymbol{\delta}_i,$$

where $\boldsymbol{\Lambda}_\omega = (\mathbf{B}, \boldsymbol{\Pi}, \boldsymbol{\Gamma})$ with general elements $\lambda_{\omega kj}$ for $k = 1, \cdots, q_1$, and $\mathbf{v}_i = (\mathbf{d}_i^T, \boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T = (\mathbf{d}_i^T, \boldsymbol{\omega}_i^T)^T$ be an $(r_2 + q_1 + q_2) \times 1$ vector. The model $\boldsymbol{\eta}_i = \boldsymbol{\Lambda}_\omega \mathbf{v}_i + \boldsymbol{\delta}_i$ is similar to $\mathbf{y}_i = \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i$ considered before. Hence, the derivation for the conditional distributions corresponding to $\boldsymbol{\theta}_\omega$ is similar to that corresponding to $\boldsymbol{\theta}_y$.

Let $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_n)$, $\mathbf{L}_\omega$ be the index matrix with general elements $l_{\omega kj}$ that similarly defined as $\mathbf{L}$ to indicate the fixed known parameters in $\boldsymbol{\Lambda}_\omega$; $\psi_{\delta k}$ be the kth diagonal element of $\boldsymbol{\Psi}_\delta$ and $\boldsymbol{\Lambda}_{\omega k}^T$ be the row vector that contains the unknown parameters in the kth row of $\boldsymbol{\Lambda}_\omega$. Let $\mathbf{V}_k$ be the submatrix of $\mathbf{V}$ such that all the rows corresponding to $l_{\omega kj} = 0$ are deleted; and let $\boldsymbol{\Xi}_k^T = (\eta_{1k}^*, \cdots, \eta_{nk}^*)$ where

$$\eta_{ik}^* = \eta_{ik} - \sum_{j=1}^{r_2+q} \lambda_{\omega kj} v_{ij} (1 - l_{\omega kj}).$$

It can be shown that

$$[\psi_{\delta k}^{-1}|\boldsymbol{\Omega}] \stackrel{D}{=} \ Gamma[n/2 + \alpha_{0\delta k}, \beta_{\delta k}],$$
$$[\boldsymbol{\Lambda}_{\omega k}|\boldsymbol{\Omega}, \psi_{\delta k}^{-1}] \stackrel{D}{=} N[\mathbf{a}_{\omega k}, \psi_{\delta k}\mathbf{A}_{\omega k}],$$

where $\mathbf{A}_{\omega k} = (\mathbf{H}_{0\omega k}^{-1} + \mathbf{V}_k\mathbf{V}_k^T)^{-1}$, $\mathbf{a}_{\omega k} = \mathbf{A}_{\omega k}(\mathbf{H}_{0\omega k}^{-1}\boldsymbol{\Lambda}_{0\omega k} + \mathbf{V}_k\boldsymbol{\Xi}_k)$, and

$$\beta_{\delta k} = \beta_{0\delta k} + \frac{1}{2}(\boldsymbol{\Xi}_k^T\boldsymbol{\Xi}_k - \mathbf{a}_{\omega k}^T\mathbf{A}_{\omega k}^{-1}\mathbf{a}_{\omega k} + \boldsymbol{\Lambda}_{0\omega k}^T\mathbf{H}_{0\omega k}^{-1}\boldsymbol{\Lambda}_{0\omega k}).$$

Again, the conditional distribution

$$[\boldsymbol{\theta}_\omega|\boldsymbol{\Omega}] = \prod_{k=1}^{q_1}[\psi_{\delta k}^{-1}|\boldsymbol{\Omega}][\boldsymbol{\Lambda}_{\omega k}|\boldsymbol{\Omega}, \psi_{\delta k}^{-1}]$$

is a normal-Gamma distribution.

3. Sample from posterior distribution using MCMC methods.

(1) Data augmentation:

Instead of working on the intractable posterior density $p(\boldsymbol{\theta}|\mathbf{Y})$, we work on $p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y})$, where $\boldsymbol{\Omega}$ is the set of latent variables in the model.

Given the latent Ω to form complete data (Y, Ω)

(2) The Gibbs sampler:

The Gibbs sampler (Geman and Geman, 1984) is used to simulate observations from $p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y})$ by drawing observations iteratively from their full conditional densities $p(\boldsymbol{\theta}|\boldsymbol{\Omega}, \mathbf{Y})$ and $p(\boldsymbol{\Omega}|\boldsymbol{\theta}, \mathbf{Y})$.

(3) Check convergence:

The burn-in iterations for achieving convergence of the Gibbs sampler can be determined by plots of the simulated sequences of the individual parameters. At convergence, parallel sequences generated with different starting values should mix well together. EPSR values can also be used to check convergence of MCMC algorithms.

The 'estimated potential scale reduction (EPSR)' value is used to monitor the convergence of MCMC algorithm. As suggested by Gelmen (1996), convergence is achieved when the EPSR values are all less than 1.2. The computation of the EPSR values is based on several simulation sequences generated independently from different starting points. The EPSR approach monitors each parameter of interest separately.

Let n be the length of the each sequence. For each parameter estimate, say θ, let θ_{jk} $(j = 1, \cdots, n; k = 1, \cdots, K)$ be the observations from K parallel sequences of length n. The between- and within-sequence variances are computed as

* Introduced in PML: Intro.

$$B = \frac{n}{K-1}\sum_{k=1}^{K}(\theta_{\cdot k} - \theta_{\cdot\cdot})^2, \quad \theta_{\cdot k} = n^{-1}\sum_{j=1}^{n}\theta_{jk}, \ \theta_{\cdot\cdot} = K^{-1}\sum_{k=1}^{K}\theta_k,$$

$$W = \frac{1}{K}\sum_{k=1}^{K}s_k^2, \qquad s_k^2 = (n-1)^{-1}\sum_{j=1}^{n}(\theta_{jk} - \theta_{\cdot k})^2.$$

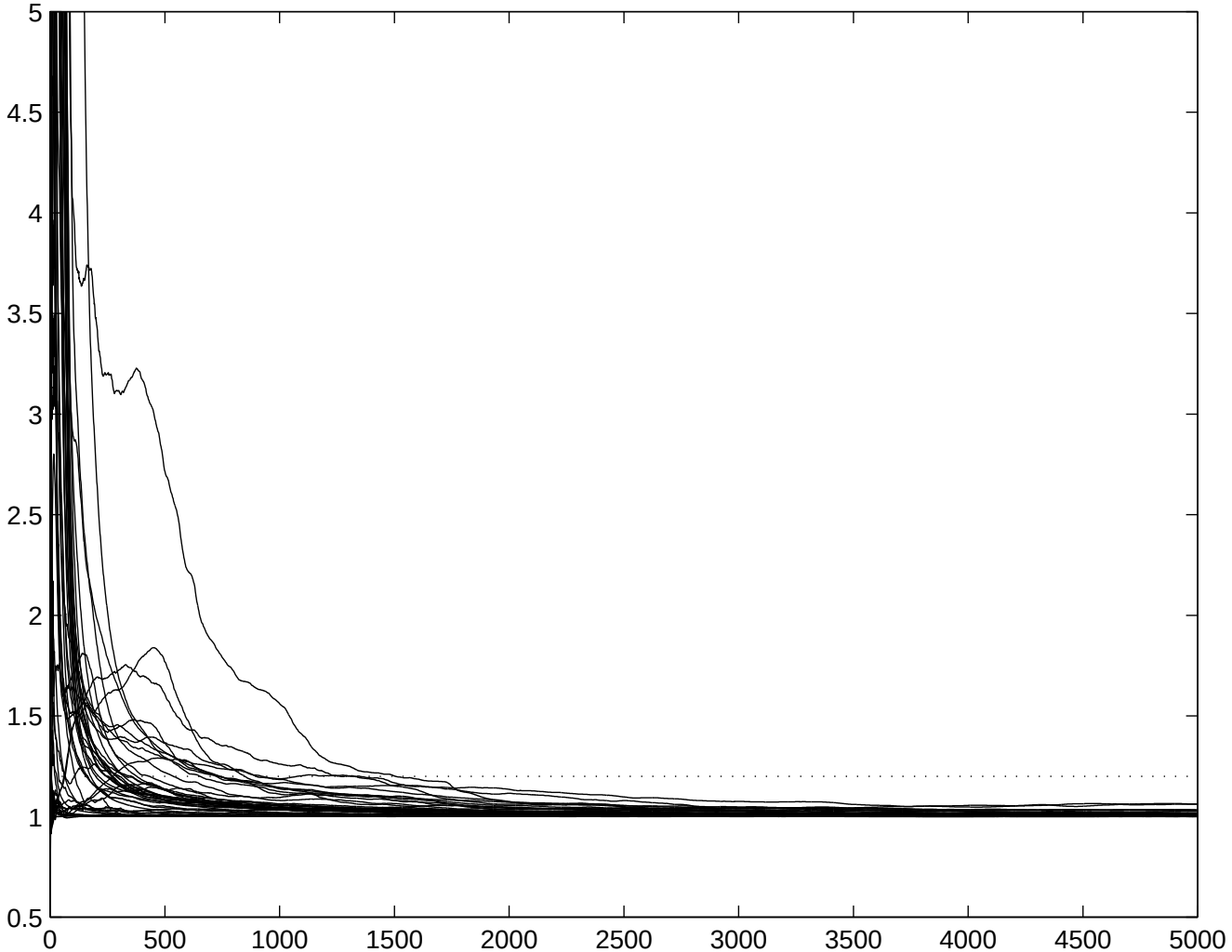
The estimate of $\text{var}(\theta|\mathbf{Y})$, the marginal posterior variance of the estimate, is then obtained by a weighted average of B and W as follows:

$$\widehat{\text{var}}(\theta) = \frac{n-1}{n} W + \frac{1}{n} B.$$

The EPSR is defined as

$$\hat{R}^{1/2} = \left[\widehat{\text{var}}(\theta)/W\right]^{1/2}.$$

As the algorithm converges, $\hat{R}^{1/2}$ should be close to 1.0. In monitoring convergence, all EPSR values for all parameters are computed. Convergence is achieved if all the EPSR values are less than 1.2.


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000

(4) Statistical inference:

Statistical inference of the model can be conducted on the basis of a simulated sample of observations from $p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y})$, namely, $\{(\boldsymbol{\theta}^{(t)}, \boldsymbol{\Omega}^{(t)}) : \ t = 1, \cdots, T^*\}$. The Bayesian estimate of $\boldsymbol{\theta}$ and its standard error estimate can be obtained as follows:

$$\hat{\boldsymbol{\theta}} = T^{*-1} \sum_{t=1}^{T^*} \boldsymbol{\theta}^{(t)}, \qquad \hat{\omega}_i = T^{*-1} \sum_{t=1}^{T^*} \omega_i^{(t)}.$$

$$Var(\hat{\boldsymbol{\theta}}|\mathbf{Y}) = (T^* - 1)^{-1} \sum_{t=1}^{T^*} (\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}})^T,$$

Other statistical inference on $\boldsymbol{\theta}$ can be carried out based on the simulated sample, $\{\boldsymbol{\theta}^{(t)} : \ t = 1, \ldots, T^*\}$. For instance, the 2.5% and 97.5% quantiles of the sampled distribution of an individual parameter can give a 95% posterior credible interval and convey skewness in its marginal posterior density.

⇒ Significant or not.

4. Sensitivity analysis.

A sensitivity analysis is usually necessary in the Bayesian analysis. The main objective is to check whether Bayesian results are sensitive to given prior inputs. It can be conducted by disturbing prior inputs in certain ranges.

Based on our experience, Bayesian results are

(1) relatively robust to the prior inputs for regression type parameters;

(2) relatively sensitive to the prior inputs for variance and covariance parameters.

Developing robust Bayesian methods and less sensitive prior distributions has attracted continuous attention in the statistical literature.