Bayesian reduced rank regression in econometrics

Caleb Jin

5-9-2018

Contents

Chapter 1

Foreword

I am Caleb Jin. After I read this paper, **Bayesian reduced rank regression in econometrics**(Geweke, 1996), I write down the nodes of the key idea and R code to realize it.

Chapter 2

Bayesian Reduced Rank Regression

2.1 Introduction

In contemporary sociey, a big amount of data are generated and collected more easily and routinely in various academic and industrial areas, such as engineering, politics, B2C e-ccommerce, genomics, etc. Many problems could be cast into statistical problems under the framework of multivariate linear regression model, which is characterized by that both reponse variables and predictors are high dimensionality.

We assume n independent observations of the response $\mathbf{y}_i \in \mathcal{R}^q$ with predictor vector $\mathbf{x}_i \in \mathcal{R}^p$, $i = 1, 2, \ldots, n$. Cosider multivariate linear regression model as follows:

$$\mathbf{Y} = \mathbf{XC} + \mathbf{E}, \tag{2.1}$$

where $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)^\top \in \mathcal{R}^{n\times q}$ is the response matrix, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^\top \in \mathcal{R}^{n\times p}$ is the design matrix, $\mathbf{C} \in \mathcal{R}^{p\times q}$ is the coefficient matrix, and $\mathbf{E} = (\mathbf{ue}_1, \mathbf{ue}_2, \ldots, \mathbf{ue}_n)^\top \in \mathcal{R}^{n\times q}$ is the disturbance matrix with $\mathbf{ue}_i$'s $\overset{iid}{\sim} \mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma}_e)$. We assume $\mathbf{\Sigma}_e = \sigma^2\mathbf{I}_q$. Therefore, we have $\mathbf{Y} \sim \mathcal{MN}(\mathbf{XC}, \mathbf{\Sigma}_e, \mathbf{I}_n)$.

2.2 Bayesian Rank Reduced Regression model

We consider to decompose the coefficient matrix into two part as follows:

$$\mathbf{C} = \mathbf{AB}^\top, \tag{2.2}$$

where $\mathbf{A} \in \mathcal{R}^{p\times r}$, $\mathbf{B} \in \mathcal{R}^{r\times q}$ and known $r \leq \min(p,q)$. However, this decomposition is not unique, because with a $r \times r$ nonsingular matrix $\mathbf{Q}, \mathbf{C} = \mathbf{A}\mathbf{B}^\top = \mathbf{A}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{B}^\top = \tilde{\mathbf{A}}\tilde{\mathbf{B}}^\top$. In order to indentify it, we further decompose $\mathbf{A}$. $\mathbf{A}^\top = [\mathbf{I}_r, \mathbf{A}^{*\top}]$. The author assumes that $p(\mathbf{A},\mathbf{B}) \propto \exp\left(-\frac{\tau^2}{2} trace\{\mathbf{A}^\top\mathbf{A} + \mathbf{B}^\top\mathbf{B}\}\right)$ and $\sigma^2 \sim \mathcal{IG}(\frac{a}{2}, \frac{b}{2})$.

2.2.1 Posterior Distribution

Let $\tilde{\mathbf{a}}_k$ and $\tilde{\mathbf{b}}_k$ denote the kth column of $\mathbf{A}$ and $\mathbf{B}$, respectively. Let $\mathbf{a}_j^\top$ and $\mathbf{b}_l^\top$ denote the yth row of $\mathbf{A}$ and lth row of $\mathbf{B}$, respectively.

$$
\begin{aligned}
& p(\mathbf{Y}|\mathbf{A},\mathbf{B},\mathbf{\Sigma}_e) \\
\propto\ & |\mathbf{\Sigma}_e|^{-\frac{q}{2}} \exp\left(-\frac{1}{2} trace\{(\mathbf{Y}-\mathbf{X}\mathbf{A}\mathbf{B}^\top)(\mathbf{Y}-\mathbf{X}\mathbf{A}\mathbf{B}^\top)^\top\mathbf{\Sigma}_e^{-1}\}\right) \\
=\ & (\sigma^2)^{-\frac{nq}{2}} \exp\left(-\frac{1}{2\sigma^2} trace\{(\mathbf{Y}-\mathbf{X}\mathbf{A}\mathbf{B}^\top)(\mathbf{Y}-\mathbf{X}\mathbf{A}\mathbf{B}^\top)^\top\}\right) \\
=\ & (\sigma^2)^{-\frac{nq}{2}} \exp\left(-\frac{1}{2\sigma^2} trace\{(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top)(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top)^\top\}\right) \\
\times\ & \exp\left(-\frac{1}{2\sigma^2} trace\{(\tilde{\mathbf{x}}_j\mathbf{a}_j^\top\mathbf{B}^\top\mathbf{B}\mathbf{a}_j\tilde{\mathbf{x}}_j^\top)\}\right) \\
\times\ & \exp\left(\frac{1}{\sigma^2} trace\{(\tilde{\mathbf{x}}_j\mathbf{a}_j^\top\mathbf{B}^\top(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top)^\top)\}\right) \\
=\ & (\sigma^2)^{-\frac{nq}{2}} \exp\left(-\frac{1}{2\sigma^2} trace\{(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top)(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top)^\top\}\right) \\
\times\ & \exp\left(-\frac{1}{2\sigma^2}\mathbf{a}_j^\top\mathbf{B}^\top\mathbf{B}\mathbf{a}_j\tilde{\mathbf{x}}_j^\top\tilde{\mathbf{x}}_j\right) \times \exp\left((-2)\frac{-1}{2\sigma^2}\mathbf{a}_j^\top\mathbf{B}^\top(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top)^\top\tilde{\mathbf{x}}_j\right).
\end{aligned}
$$

$$
\begin{aligned}
& p(\mathbf{a}_j|\mathbf{A}_{(j)},\mathbf{B},\boldsymbol{\Sigma}_e,\mathbf{Y}) \\
\propto\ & p(\mathbf{A}|\mathbf{B},\boldsymbol{\Sigma}_e,\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{A},\mathbf{B},\boldsymbol{\Sigma}_e)p(\mathbf{A},\mathbf{B}) \\
\propto\ & p(\mathbf{Y}|\mathbf{A},\mathbf{B},\boldsymbol{\Sigma}_e)\exp\left(-\frac{\tau^2}{2}trace\{\mathbf{A}^\top\mathbf{A}+\mathbf{B}^\top\mathbf{B}\}\right) \\
\propto\ & p(\mathbf{Y}|\mathbf{A},\mathbf{B},\boldsymbol{\Sigma}_e)\exp\left(-\frac{\tau^2}{2}\sum_{j=1}^{r}\mathbf{a}_j^\top\mathbf{a}_j\right) \\
\propto\ & \exp\left(-\frac{1}{2\sigma^2}\mathbf{a}_j^\top\mathbf{B}^\top\mathbf{B}\mathbf{a}_j\tilde{\mathbf{x}}_j^\top\tilde{\mathbf{x}}_j\right)\times\exp\left((-2)\frac{-1}{2\sigma^2}\mathbf{a}_j^\top\mathbf{B}^\top(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top)^\top\tilde{\mathbf{x}}_j\right) \\
\times\ & \exp\left(-\frac{\tau^2}{2}\mathbf{a}_j^\top\mathbf{a}_j\right) \\
=\ & \exp\left\{-\frac{1}{2}\left(\mathbf{a}_j^\top(\sigma^{-2}\mathbf{B}^\top\mathbf{B}\tilde{\mathbf{x}}_j^\top\tilde{\mathbf{x}}_j+\mathbf{I}_r\tau^2)\mathbf{a}_j-2\mathbf{a}_j^\top\mathbf{B}^\top(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top)^\top\tilde{\mathbf{x}}_j\sigma^{-2}\right)\right\} \\
=\ & \exp\left\{-\frac{1}{2}\left(\mathbf{a}_j^\top{\boldsymbol{\Sigma}_j^A}^{-1}\mathbf{a}_j-2\mathbf{a}_j^\top{\boldsymbol{\Sigma}_j^A}^{-1}\boldsymbol{\Sigma}_j^A\mathbf{B}^\top(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top)^\top\tilde{\mathbf{x}}_j\sigma^{-2}\right)\right\} \\
=\ & \exp\left\{-\frac{1}{2}\left(\mathbf{a}_j^\top{\boldsymbol{\Sigma}_j^A}^{-1}\mathbf{a}_j-2\mathbf{a}_j^\top{\boldsymbol{\Sigma}_j^A}^{-1}\boldsymbol{\mu}_j^A\right)\right\} \\
\propto\ & \exp\left\{-\frac{1}{2}\left((\mathbf{a}_j-\boldsymbol{\mu}_j^A)^\top{\boldsymbol{\Sigma}_j^A}^{-1}(\mathbf{a}_j-\boldsymbol{\mu}_j^A)\right)\right\},
\end{aligned}
$$

where $\boldsymbol{\Sigma}_j^A=(\mathbf{B}^\top\mathbf{B}\tilde{\mathbf{x}}_j^\top\tilde{\mathbf{x}}_j\sigma^{-2}+\mathbf{I}_r\tau^2)^{-1}$ and $\boldsymbol{\mu}_j^A=\boldsymbol{\Sigma}_j^A\mathbf{B}^\top(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^\top)^\top\tilde{\mathbf{x}}_j\sigma^{-2}$.

Hence, $\mathbf{a}_j|\mathbf{A}_{(j)},\mathbf{B},\boldsymbol{\Sigma}_e,\mathbf{Y}\sim\mathcal{N}_r(\boldsymbol{\mu}_j^A,\boldsymbol{\Sigma}_j^A)$.

$$
\begin{aligned}
& p(\mathbf{Y}|\mathbf{A},\mathbf{B},\boldsymbol{\Sigma}_e) \\
\propto\ & |\boldsymbol{\Sigma}_e|^{-\frac{q}{2}}\exp\left(-\frac{1}{2}trace\{(\mathbf{Y}-\mathbf{X}\mathbf{A}\mathbf{B}^\top)^\top\boldsymbol{\Sigma}_e^{-1}(\mathbf{Y}-\mathbf{X}\mathbf{A}\mathbf{B}^\top)\}\right) \\
=\ & (\sigma^2)^{-\frac{nq}{2}}\exp\left(-\frac{1}{2\sigma^2}trace\{(\mathbf{Y}-\mathbf{X}\mathbf{A}\mathbf{B}^\top)^\top(\mathbf{Y}-\mathbf{X}\mathbf{A}\mathbf{B}^\top)\}\right) \\
=\ & (\sigma^2)^{-\frac{nq}{2}}\exp\left(-\frac{1}{2\sigma^2}trace\{(\mathbf{Y}_{(\tilde{l})}-\mathbf{X}\mathbf{A}(\mathbf{B}^\top)_{(\tilde{l})})^\top(\mathbf{Y}_{(\tilde{l})}-\mathbf{X}\mathbf{A}(\mathbf{B}^\top)_{(\tilde{l})})\}\right) \\
\times\ & \exp\left(-\frac{1}{2\sigma^2}\left[\mathbf{b}_l^\top(\mathbf{X}\mathbf{A})^\top\mathbf{X}\mathbf{A}\mathbf{b}_l-2\mathbf{b}_l^\top(\mathbf{X}\mathbf{A})^\top\tilde{\mathbf{y}}_l+\tilde{\mathbf{y}}_l^\top\tilde{\mathbf{y}}_l\right]\right).
\end{aligned}
$$

$$
\begin{aligned}
& p(\mathbf{b}_l|\mathbf{A},(\mathbf{B}^\top)_{(\tilde{l})},\mathbf{Y},\mathbf{\Sigma}_e) \\
\propto\ & p(\mathbf{B}|\mathbf{A},\mathbf{Y},\mathbf{\Sigma}_e) \propto p(\mathbf{Y}|\mathbf{A},\mathbf{B},\mathbf{\Sigma}_e)p(\mathbf{A},\mathbf{B}) \\
\propto\ & p(\mathbf{Y}|\mathbf{A},\mathbf{B},\mathbf{\Sigma}_e)\exp\left(-\frac{\tau^2}{2}trace\{\mathbf{A}^\top\mathbf{A}+\mathbf{B}^\top\mathbf{B}\}\right) \\
\propto\ & p(\mathbf{Y}|\mathbf{A},\mathbf{B},\mathbf{\Sigma}_e)\exp\left(-\frac{\tau^2}{2}\sum_{l=1}^{r}\mathbf{b}_l^\top\mathbf{b}_l\right) \\
\propto\ & \exp\left(-\frac{1}{2\sigma^2}\left[\mathbf{b}_l^\top(\mathbf{XA})^\top\mathbf{XA}\mathbf{b}_l-2\mathbf{b}_l^\top(\mathbf{XA})^\top\tilde{\mathbf{y}}_l+\tilde{\mathbf{y}}_l^\top\tilde{\mathbf{y}}_l\right]\right)\exp\left(-\frac{\tau^2}{2}\mathbf{b}_l^\top\mathbf{b}_l\right) \\
=\ & \exp\left(-\frac{1}{2}\left[\mathbf{b}_l^\top(\mathbf{XA})^\top\mathbf{XA}\mathbf{b}_l\sigma^{-2}-2\mathbf{b}_l^\top(\mathbf{XA})^\top\tilde{\mathbf{y}}_l\sigma^{-2}+\tilde{\mathbf{y}}_l^\top\tilde{\mathbf{y}}_l\sigma^{-2}\right]\right)\exp\left(-\frac{\tau^2}{2}\mathbf{b}_l^\top\mathbf{b}_l\right) \\
=\ & \exp\left\{-\frac{1}{2}\left(\mathbf{b}_l^\top((\mathbf{XA})^\top\mathbf{XA}\sigma^{-2}+\mathbf{I}_r\tau^2)\mathbf{b}_l-2\mathbf{b}_l^\top(\mathbf{XA})^\top\tilde{\mathbf{y}}_l\sigma^{-2}\right)\right\} \\
=\ & \exp\left\{-\frac{1}{2}\left(\mathbf{b}_l^\top\mathbf{\Sigma}_j^{B^{-1}}\mathbf{b}_l-2\mathbf{b}_l^\top\mathbf{\Sigma}_j^{B^{-1}}\mathbf{\Sigma}_j^B(\mathbf{XA})^\top\tilde{\mathbf{y}}_l\sigma^{-2}\right)\right\} \\
=\ & \exp\left\{-\frac{1}{2}\left(\mathbf{b}_l^\top\mathbf{\Sigma}_j^{B^{-1}}\mathbf{b}_l-2\mathbf{b}_l^\top\mathbf{\Sigma}_j^{B^{-1}}\boldsymbol{\mu}_j^B\right)\right\} \\
\propto\ & \exp\left\{-\frac{1}{2}\left((\mathbf{b}_l-\boldsymbol{\mu}_j^B)^\top\mathbf{\Sigma}_j^{B^{-1}}(\mathbf{b}_l-\boldsymbol{\mu}_j^B)\right)\right\},
\end{aligned}
$$

where $\mathbf{\Sigma}_j^B=((\mathbf{XA})^\top\mathbf{XA}\sigma^{-2}+\mathbf{I}_r\tau^2)^{-1}$ and $\boldsymbol{\mu}_j^B=\mathbf{\Sigma}_j^B(\mathbf{XA})^\top\tilde{\mathbf{y}}_l\sigma^{-2}$.

Hence, $\mathbf{b}_l|\mathbf{A},(\mathbf{B}^\top)_{(\tilde{l})},\mathbf{Y},\mathbf{\Sigma}_e\sim\mathcal{N}_r(\boldsymbol{\mu}_j^B,\mathbf{\Sigma}_j^B)$.

We know that the element in kth row and kth column of $\mathbf{\Sigma}_e$ is $\sigma^2, k=1,2,\ldots,q$.

$$
\begin{aligned}
& p(\sigma^2|\mathbf{A},\mathbf{B},\mathbf{Y},(\mathbf{\Sigma}_e)_{(k)(\tilde{k})}) \\
\propto\ & p(\mathbf{\Sigma}_e|\mathbf{A},\mathbf{B},\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{A},\mathbf{B},\mathbf{\Sigma}_e)p(\mathbf{\Sigma}_e) \propto p(\mathbf{Y}|\mathbf{A},\mathbf{B},\mathbf{\Sigma}_e)p(\sigma^2) \\
=\ & (\sigma^2)^{-\frac{nq}{2}}\exp\left(-\frac{1}{2\sigma^2}trace\{(\mathbf{Y}-\mathbf{XAB}^\top)(\mathbf{Y}-\mathbf{XAB}^\top)^\top\}\right) \\
\times\ & (\sigma^2)^{-\frac{a}{2}-1}\exp\left(-\frac{b}{2\sigma^2}\right) \\
=\ & (\sigma^2)^{-\frac{nq+a}{2}-1}\exp\left(-\frac{1}{2\sigma^2}(trace\{(\mathbf{Y}-\mathbf{XAB}^\top)(\mathbf{Y}-\mathbf{XAB}^\top)^\top\}+b)\right).
\end{aligned}
$$

Hence, $\sigma^2|\mathbf{A},\mathbf{B},\mathbf{Y},(\mathbf{\Sigma}_e)_{(k)(\tilde{k})}\sim\mathcal{IG}\left(\frac{nq+a}{2},\frac{1}{2}(trace\{(\mathbf{Y}-\mathbf{XAB}^\top)(\mathbf{Y}-\mathbf{XAB}^\top)^\top\}+b)\right)$.

2.2.2 Gibbs Sampling

The algorithm is easy to construct. Begin with initial values $\mathbf{A}^{(0)},\mathbf{\Sigma}^{(0)}$, then for t = 1,2,…

- Step (1) Given $\mathbf{A}_{(j)}^{(t-1)}, \mathbf{B}^{\top(t-1)}, {\sigma^2}^{(t-1)}$, draw $\mathbf{a}_j^{(t)}$ from $\mathcal{N}_r(\boldsymbol{\mu}_j^A, \boldsymbol{\Sigma}_j^A), j = r+1, r+2, \ldots, p$;
- Step (2) Given $\mathbf{A}^{(t)}, \mathbf{B}_{(\tilde{l})}^{\top(t-1)}, {\sigma^2}^{(t-1)}$, draw $\mathbf{b}_l^{(t)}$ from $\mathcal{N}_r(\boldsymbol{\mu}_j^B, \boldsymbol{\Sigma}_j^B), l = 1, 2, \ldots, q$;
- Step (3) Given $\mathbf{B}^{(t)}, \mathbf{A}^{(t)}$, draw ${\sigma^2}^{(t)}$ from $\mathcal{IG}\,(a^*, b^*)$.
- Step (4) Repeat step (1) to step (3) until convergence.

Note that, the step (1) samples the $\mathbf{a}_j$ from $j = r+1$, because first part of $\mathbf{A}$ is $\mathbf{I}_r$, which is known. Hence, step (1) samples the $\mathbf{A}^*$ and then insert $\mathbf{I}_r$ together to construct $\mathbf{A}$.

2.3 Model Selection

To this point we have proceeded as if r were known. In most applications this will not be true and so analysis to this point is conditional.When r is unknown, the analysis may be carried out for several alternative values of r. Under the Bayesian framework, model selection can be implemented by using the model posterior probability conditioning the data,

$$p(M_r|y) = \frac{p(y|M_r)p(M_r)}{\sum_{r\in\mathcal{M}} p(y|M_r)p(M_r)} = \frac{p(y|M_r)}{\sum_{r\in\mathcal{M}} p(y|M_r)},$$

where $\mathcal{M} = 1, 2, \ldots, min(p, q)$, and $p(M_r) = \frac{1}{card(\mathcal{M})}$.

Hence, as the prior of rank(model) is flat, $p(M_r|y)$ is only determined by marginal likelihood ($p(y|M_r)$). The model selection problem here is then converted to find out the rank which maximizes the marginal likelihood ($p(y|M_r)$). In order to calculate the $p(y|M_r)$, I used Laplace and Gelfand and Dey (GD) methods, and then compare with DIC.

2.3.1 Laplace

Let $\boldsymbol{\theta} = (\mathbf{A}, \mathbf{B}, \sigma^2)$

$$\begin{aligned} p(\mathbf{Y}|M_r) &= \int\int\int p(\mathbf{Y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta})d\mathbf{A}d\mathbf{B}d\sigma^2 \\ &\approx p(\mathbf{Y}|\hat{\boldsymbol{\theta}}, M)p(\hat{\boldsymbol{\theta}}|M)|(nq)^{-1}\hat{\boldsymbol{\Sigma}}_M|^{1/2}(2\pi)^{(k_M/2)}, \end{aligned}$$

where $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\sigma}^2) = \arg\max p(\mathbf{Y}|\mathbf{A}, \mathbf{B}, \sigma^2, M)p(\mathbf{A}, \mathbf{B}, \sigma^2|M)$.

Hence,

$$\begin{aligned}
& \log p(\mathbf{Y}|M_r) \\
\approx\ & \log p(\mathbf{Y}|\hat{\boldsymbol{\theta}}, M) + \log p(\hat{\boldsymbol{\theta}}|M) - \frac{1}{2}k_M \log nq + \frac{1}{2}|\boldsymbol{\Sigma}_M| + \frac{k_M}{2}\log 2\pi \\
=\ & -\frac{1}{2}\left(-2\log(p(\mathbf{Y}|\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\sigma^2}, M)) + k_M \log nq\right) + C \\
=\ & -\frac{1}{2}(nq\log(2\pi\hat{\sigma^2}) - \frac{1}{2\hat{\sigma^2}} trace\left\{(\mathbf{Y} - \mathbf{X}\hat{\mathbf{C}})^\top(\mathbf{Y} - \mathbf{X}\hat{\mathbf{C}})\right\} \\
& +[r(p+q-r)+1]\log(nq)) + C \\
=\ & -\frac{1}{2}BIC + C \\
=\ & -\frac{1}{2}BIC, \text{as } n \to \infty. \qquad (2.3)
\end{aligned}$$

The problem we are facing when we use Laplace here is that we need to find the mode of $p(\mathbf{A}, \mathbf{B}, \sigma^2|\mathbf{Y})$, in which $(\mathbf{A}, \mathbf{B}, \sigma^2)$ is high-dimensional. To address this problem, Besag proposed an iterated conditional modes (ICM) algorithm. The ICM obtains the local maximum of the joint posterior by iteratively maximizing the full conditionals as follows:

- Begin with initial values $\mathbf{A}^{(0)}, \boldsymbol{\Sigma}^{(0)}$, then for t = 1,2,...
- Given $\mathbf{A}_{(j)}^{(t-1)}, \mathbf{B}^{\top(t-1)}, \sigma^{2(t-1)}$, $\mathbf{a}_j^{(t)} \leftarrow \boldsymbol{\mu}_j^A, j = r+1, r+2, \ldots, p$;
- Given $\mathbf{A}^{(t)}, \mathbf{B}_{(\tilde{l})}^{\top(t-1)}, \sigma^{2(t-1)}$, $\mathbf{b}_l^{(t)} \leftarrow \boldsymbol{\mu}_j^B, l = 1, 2, \ldots, q$;
- Given $\mathbf{B}^{(t)}, \mathbf{A}^{(t)}$, $\sigma^{2(t)} \leftarrow \frac{b^*}{a^*-1}$.

We obtain $\mathbf{C}^{(t)} = \mathbf{A}^{(t)}\mathbf{B}^{\top(t)}$ and estimate $\hat{\mathbf{C}}_{ICM} = T^{-1}\sum_{t=1}^{T}\mathbf{C}^{(t)}$. Put the $\hat{\mathbf{C}}_{ICM}$ in the Eq.(2.3) to calculate $-\frac{1}{2}BIC$. Hence, the Laplace here is actually propotional to BIC.

2.3.2 DIC

Let $\boldsymbol{\theta} = (\mathbf{A}, \mathbf{B}, \sigma^2)$,

$$\begin{aligned}
& D(\boldsymbol{\theta}) \\
=\ & -2\log p(\mathbf{Y}|\boldsymbol{\theta}, M) \\
=\ & nq\log(2\pi\sigma^2) + \frac{1}{\sigma^2} trace\left\{(\mathbf{Y} - \mathbf{X}\mathbf{C})^\top(\mathbf{Y} - \mathbf{X}\mathbf{C})\right\}
\end{aligned}$$

Then DIC can be computed by

$$DIC \approx 2T^{-1}\sum_{t=1}^{T} D(\boldsymbol{\theta}^{(t)}) - D(T^{-1}\sum_{t=1}^{T}\boldsymbol{\theta}^{(t)}),$$

where $\boldsymbol{\theta}^{(t)}$ is a MCMC sample generated from $p(\boldsymbol{\theta}|\mathbf{Y})$.

Note that $DIC = D(\bar{\boldsymbol{\theta}}) + 2P_D$, which is analogous to AIC.

2.3.3 Gelfand & Dey (GD)

Let $\boldsymbol{\theta} = (\mathbf{A}, \mathbf{B}, \sigma^2)$,then the GD estimator is

$$p(\mathbf{Y}|M) \approx \left[T^{-1} \sum_{t=1}^{T} \frac{g(\boldsymbol{\theta}^{(t)})}{p(\mathbf{Y}|\boldsymbol{\theta}^{(t)}) p(\boldsymbol{\theta}^{(t)})} \right]^{-1},$$

where $\boldsymbol{\theta}^{(t)}$ is a MCMC sample generated from $p(\boldsymbol{\theta}|\mathbf{Y})$. Define $g(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\Sigma}})$, where $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\Sigma}}$ are MCMC sample mean and variance, respectively. I use formular as follows:

$$\begin{aligned}
\log p(\mathbf{Y}|M) &\approx \log \left[\frac{1}{T} \sum_{t=1}^{T} \frac{g^{(t)}}{f^{(t)}} \right]^{-1} \\
&= \log \left[\frac{T}{\sum_{t=1}^{T} \frac{g^{(t)}}{f^{(t)}}} \right] \\
&= \log T - \log \left(\sum_{t=1}^{T} \frac{g^{(t)}}{f^{(t)}} \right) \\
&= \log T - \log \left(\sum_{t=1}^{T} \exp(\log g^{(t)} - \log f^{(t)}) \right) \qquad (2.4) \\
&= \log T - \log \left(\sum_{t=1}^{T} \exp c_t \right) \\
&= \log T - \log \left(e^{c_1} e^{\sum_{t=1}^{T} (c_t - c_1)} \right) \\
&= \log T - \left[c_1 + \log \left(\sum_{t=2}^{T} (c_t - c_1) \right) \right],
\end{aligned}$$

where $g^{(t)} = g(\boldsymbol{\theta})$, $f^{(t)} = p(\mathbf{Y}|\boldsymbol{\theta}^{(t)}) p(\boldsymbol{\theta}^{(t)})$, $c_t = \log g^{(t)} - \log f^{(t)}$.

Note that when calculating $\log p(\mathbf{Y}|M)$, there is a computation problem in formula (2.4), $\exp(\log g^{(t)} - \log f^{(t)})$ goes to infinity, due to a very large magnitude of $\log g^{(t)} - \log f^{(t)}$. To solve this problem, I used $\log T - \left[c_1 + \log \left(\sum_{t=2}^{T} (c_t - c_1) \right) \right]$ to calculate $\log p(\mathbf{Y}|M)$.

2.4 Simulation Study

2.4.1 Data Generation

In the simulation study, my goal is to find out the true model or true rank of coefficient matrix($\mathbf{C}$) based on Laplace, DIC and GD method. I set $n = 100, q = 12, p = 7$ and $\sigma^2 = 2$. The coefficient matrix is as follows:

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 1 & 0 & 0 & 0 & -3 & 2 & 0 & 0 & 0 & -1 & 3 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 3 & -3 & 4 & 2 & 2 \\ 1 & 0 & 0 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 1 & 0 & 0 & 0 & -3 & 2 & 0 & 0 & 0 & -1 & 3 \\ 0 & 1 & 0 & 0 & 0 & -3 & 2 & 0 & 0 & 0 & -1 & 3 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 3 & -3 & 4 & 2 & 2 \end{bmatrix}$$

Hence, the true rank of $\mathbf{C}$ is 3.

Generate data based on the model from Eq.(2.1). For the prior, $p(\mathbf{A}, \mathbf{B}) \propto \exp\left(-\frac{\tau^2}{2} trace\{\mathbf{A}^\top \mathbf{A} + \mathbf{B}^\top \mathbf{B}\}\right)$ and $\sigma^2 \sim \mathcal{IG}(\frac{a}{2}, \frac{b}{2})$, where $\tau = 10^{-3}$ and $a = b = 1$.

After running the MCMC simulation, the result based on 1 replication is shown in Table 2.1.

Table 2.1: Model Selection among Laplace, DIC and GD Based on 1 Replication.

Method(log)	1	2	**3**	4	5	6	7
Laplace	-3007	-2535	**-2223**	-2257	-2285	-2310	-2329
DIC	84300	17779	**7302**	7317	7325	7363	7405
GD	-3062	-2631	**-2347**	-2397	-2436	-2470	-2495
MSE	1.032	0.186	**0.0120**	0.014	0.0170	0.0192	0.0217

The MSE in the table is defined as follows:

$$MSE = \frac{trace\{(\hat{\mathbf{C}} - \mathbf{C})^\top (\hat{\mathbf{C}} - \mathbf{C})\}}{pq}.$$

MSE indicates the goodness of fit in the reduced rank regression model. We can observe that the MSE is minimized when the rank is 3. This demonstrates that parameter estimation is good enough. For the Laplace, DIC and GD method, in this case, all of them select the true rank. Given rank($\mathbf{C}$) = 3, the estimated $\mathbf{C}$ is as follows:

1.07	0.17	0.01	2.34	-0.89	-0.12	-0.11	-0.08	-0.07	0.03	1.26	-1.10

-0.02	1.00	0.04	-0.14	-0.17	-2.83	2.10	0.10	0.03	0.03	-1.04	3.17
0.13	-0.03	0.99	0.18	-0.02	0.08	0.00	3.06	-2.85	4.03	2.10	2.06
0.94	0.14	0.03	2.05	-0.78	-0.08	-0.11	-0.03	-0.10	0.07	1.13	-0.95
0.04	0.99	0.04	-0.00	-0.21	-2.79	2.05	0.10	0.01	0.04	-0.95	3.05
0.12	1.05	0.00	0.18	-0.29	-2.93	2.14	-0.02	0.12	-0.12	-0.98	3.03
0.08	-0.01	0.98	0.08	0.01	0.03	0.04	3.03	-2.81	3.98	2.01	2.14

In addition, when selected rank is larger than 3, the MSE is not worse than I expect. In other words, the model based on larger model (larger rank in $\mathbf{C}$), the overfiteed models still have a good estimation for the $\mathbf{C}$. The values of DIC, Laplace and GD are not far away from the value at true rank. This is very reasonable for overfitted model. But when selected rank smaller than true rank, the MSE increase significantly, and the corresponding values of DIC, Laplace and GD are far away from the value at true rank. This means the model miss some important variables. The analysis is based on one replication.

In Table 2.3, the result is based on 1000 replication. I want to see the perfermance of methods for model selection. In Table 2.3, Laplace and GD always select true rank, however, the DIC tends to select larger model, because the average of selected rank in DIC is about 3.5, whereas, that of Laplace and GD are 3. Besides, for the successful probability of selecting true rank, the Laplace and GD, of course, is 100%, but is 61.7% for DIC. This makes sense because Laplace and GD are approximated and exact calculation for marginal likehihood, respectively. They work well in model fitting. The DIC is analogous to AIC, which tends to select larger model and better for short-term prediction.

Table 2.3: Comparison among Laplace, DIC, GD Based on 1000 Replication.

Method	Mean of Selected rank	Selection Probabiliy
Laplace	3	1
DIC	3.525	0.617
GD	3	1

2.4.2 R Code for Bayeian reduced rank regression with DIC used in the simulation study

```
library(mvtnorm)
library(invgamma)
rm(list=ls())
## Data & Model##
row1 <- c(1,0,0,2,-1,0,0,0,0,0,1,-1)
```

```
row2 <- c(0,1,0,0,0,-3,2,0,0,0,-1,3)
row3 <- c(0,0,1,0,0,0,0,3,-3,4,2,2)
C <- rbind(row1,row2,row3,row1,row2,row2,row3)
true.rank <- 3
p <- dim(C)[1]
q <- dim(C)[2]
true.sig2 <- 2
true.SIGe <- diag(true.sig2,q)
n=100
a=b=1
tau2=1e-3
MC.size=5000
burn_in=3000
reptn <- 125
error <- array(NA,dim = c(reptn,p))
DIC <- array(NA,dim = c(reptn,p))
TPR.DIC <- rep(NA,reptn)
TPR.error <- rep(NA,reptn)
hat.rank <- rep(NA,reptn)
v=0
for (i in 1:reptn) {
  t1=Sys.time()
  set.seed(2144+i+125*v)
  X <- matrix(rnorm(n*p,0,1),n,p)
  E <- rmvnorm(n, rep(0,q), true.SIGe, method="chol")
  Y=X%*%C+E
  ## Reduced Rank Regression model ##
  for (r in 1:p) {
    Hat.A <- array(NA,dim = c(p,r,MC.size))
    Hat.B <- array(NA,dim = c(q,r,MC.size))
    Hat.C <- array(NA,dim = c(p,q,MC.size))
    Hat.sig2 <- rep(NA,MC.size)
    # initial values
    hat.C <- coef(lm(Y~X-1))
    if (r==1){
      hat.B <- as.matrix(hat.C[1:r,])
    } else {
      hat.B <- t(hat.C[1:r,])
    }
    if (r==p) {
      hat.A <- diag(1,r)
    } else {
      hat.A <- rbind(diag(1,r),hat.C[(r+1):p,]%*%hat.B%*%solve(crossprod(hat.B)))
    }
```

```
  hat.sig2 <- mean(diag(tcrossprod(Y-X%*%hat.A%*%t(hat.B))))  #true.sig2
  ## MCMC ##
  for (jin in 1:MC.size) {
    # Sampling A
    if (r==p){
      Hat.A[,,jin] <- hat.A
    } else {
      for(j in (r+1):p) {
        Sig.A_j <- solve(crossprod(hat.B)*as.numeric(crossprod(X[,j]))/hat.sig2+diag(tau2,r))
        mu.A_j <- Sig.A_j%*%t(hat.B)%*%t(Y-X[,-j]%*%hat.A[-j,]%*%t(hat.B))%*%X[,j]/hat.sig2
        hat.a_j <- rmvnorm(n=1,mean = mu.A_j,sigma = Sig.A_j)
        hat.A[j,] <- hat.a_j
      }
      Hat.A[,,jin] <- hat.A
    }
    # Sampling B
    Sig.B_l <- solve(crossprod(X%*%hat.A)/hat.sig2+diag(tau2,r))
    for(l in 1:q) {
      mu.B_l <- Sig.B_l%*%t(X%*%hat.A)%*%Y[,l]/hat.sig2
      hat.b_l <- t(rmvnorm(n=1,mean = mu.B_l,sigma = Sig.B_l))
      hat.B[l,] <- t(hat.b_l)
    }
    Hat.B[,,jin] <- hat.B
    Hat.C[,,jin] <- hat.A%*%t(hat.B)
    #Sampling Sig2
    hat.sig2 <- rinvgamma(n = 1,shape = (q*n+a)/2,rate = 0.5*(b+sum(diag(tcrossprod(Y-X%*%hat.A
    Hat.sig2[jin] <- hat.sig2
    #print(jin)
  }
  hat.C <- apply(Hat.C[,,-(1:burn_in)],c(1,2),mean)
  hat.sig2 <- mean(Hat.sig2[-(1:burn_in)])
  # print(C)
  D.theta <- rep(NA,MC.size-burn_in)
  j=0
  for (jin in (burn_in+1):MC.size) {
    j=j+1
    D.theta[j] <- n*q*log(2*pi*Hat.sig2[jin])+Hat.sig2[jin]*sum(diag(crossprod(Y-X%*%Hat.C[,,ji
  }
  D.hat.theta <- n*q*log(2*pi*hat.sig2)+hat.sig2*sum(diag(crossprod(Y-X%*%hat.C)))
  DIC[i,r] <- 2*mean(D.theta)- D.hat.theta
  error[i,r] <- sum((hat.C-C)^2)/(p*q)
}
hat.rank[i] <- which.min(DIC[i,])
TPR.DIC[i] <- hat.rank[i]==true.rank
TPR.error[i] <- which.min(error[i,])==true.rank
```

```
  print(c(i,hat.rank[i]))
  print(Sys.time()-t1)
}

data.frame(hat.rank=mean(hat.rank),TPR.DIC=mean(TPR.DIC),TPR.ERR=mean(TPR.error))
save(list = ls(), file = paste0('DIC-',v,'.RData'))
```

Bibliography

Geweke, J. F. (1996). Bayesian reduced rank regression in econometrics. *Journal of Econometrics*, 75.