Multivariate bayesian variable selection and prediction

Caleb Jin

2-10-2019

Contents

Chapter 1

Foreword

I am Caleb Jin. After I read this paper, **Multivariate bayesian variable selection and prediction** (Brown and Vannucci, 1998), I write down the nodes of the key idea and R code to realize it.

Chapter 2

Multivariate bayesian variable selection and prediction

Consider a multivariate vairable regression model,

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E},$$

where $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)^\top$ is the $n \times q$ response matrix, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is the $n \times p$ design matrix, $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^\top$ is a $p \times q$ unknown coefficient matrix, and $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)^\top$ is an $n \times q$ error matrix with $\mathbf{e}_i \overset{i.i.d}{\sim} \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$. To identify the variables selected among p regressors, (Brown and Vannucci, 1998) introduced a latent binary vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)^\top$ with

$$\xi_j = \begin{cases} 1 & \text{if } \boldsymbol{\beta}_j \text{ is active} \\ 0 & \text{if } \boldsymbol{\beta}_j \text{ is not active} \end{cases}.$$

The conjugate priors are considered as follows:

$$\begin{aligned}
&\pi(\mathbf{B}_{\boldsymbol{\xi}} | \boldsymbol{\Sigma}, \boldsymbol{\xi}) \sim \mathcal{MN}(\mathbf{0}, \mathbf{H}_{\boldsymbol{\xi}}, \boldsymbol{\Sigma}) \\
&\pi(\boldsymbol{\Sigma}) \sim \mathcal{IW}(\mathbf{Q}, \delta) \\
&\pi(\boldsymbol{\xi}) \overset{ind}{\sim} \prod_{j=1}^{p} \mathrm{Ber}(\omega_j),
\end{aligned}$$

where $\mathbf{Q}$, $\mathbf{H}_{\boldsymbol{\xi}}$ and δ are hyperparameters. Then, by Bayesian theorem, the marginal posterior distribution can be obtained by

$$
\begin{aligned}
& m(\boldsymbol{\xi}|\mathbf{Y}) \propto m(\mathbf{Y}|\boldsymbol{\xi})\pi(\boldsymbol{\xi}) \\
= \quad & \iint f(\mathbf{Y}|\mathbf{B}_{\boldsymbol{\xi}}, \boldsymbol{\Sigma})\pi(\mathbf{B}_{\boldsymbol{\xi}}|\boldsymbol{\Sigma}, \boldsymbol{\xi})\pi(\boldsymbol{\Sigma})d\mathbf{B}d\boldsymbol{\Sigma}\pi(\boldsymbol{\xi}) \\
= \quad & \iint (2\pi)^{-\frac{nq}{2}}|\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}tr\{(\mathbf{Y}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})^\top(\mathbf{Y}-\mathbf{X}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}})\}\boldsymbol{\Sigma}^{-1}\right] \times \\
& (2\pi)^{-\frac{|\boldsymbol{\xi}|q}{2}}|\boldsymbol{\Sigma}|^{-\frac{|\boldsymbol{\xi}|}{2}}|\mathbf{H}_{\boldsymbol{\xi}}|^{-\frac{q}{2}} \exp\left[\frac{1}{2}tr\{\mathbf{B}_{\boldsymbol{\xi}}^\top\mathbf{H}_{\boldsymbol{\xi}}^{-1}\mathbf{B}_{\boldsymbol{\xi}}\boldsymbol{\Sigma}^{-1}\}\right] \times \\
& |\mathbf{Q}|^{\frac{\delta}{2}}2^{-\frac{1}{2}\delta q}\boldsymbol{\xi}_q^{-1}(\frac{\delta}{2})|\boldsymbol{\Sigma}|^{-\frac{\delta+q+1}{2}} \exp\{-\frac{1}{2}tr(\mathbf{Q}\boldsymbol{\Sigma}^{-1})\}d\mathbf{B}_{\boldsymbol{\xi}}d\boldsymbol{\Sigma}\pi(\boldsymbol{\xi}) \\
\propto \quad & (2\pi)^{-\frac{|\boldsymbol{\xi}|q}{2}}|\mathbf{H}_{\boldsymbol{\xi}}|^{-\frac{q}{2}} \iint |\boldsymbol{\Sigma}|^{-\frac{1}{2}(n+|\boldsymbol{\xi}|+\delta+q+1)} \exp\{-\frac{1}{2}tr(\mathbf{Q}\boldsymbol{\Sigma}^{-1})\} \\
& \exp\left[-\frac{1}{2}tr\left\{\left(\mathbf{B}_{\boldsymbol{\xi}}^\top(\mathbf{X}_{\boldsymbol{\xi}}^\top\mathbf{X}_{\boldsymbol{\xi}}+\mathbf{H}_{\boldsymbol{\xi}}^{-1})\mathbf{B}_{\boldsymbol{\xi}} - 2\mathbf{B}_{\boldsymbol{\xi}}^\top\mathbf{X}_{\boldsymbol{\xi}}^\top\mathbf{Y}\right)\boldsymbol{\Sigma}^{-1}\right\}\right] \times \\
& \exp\left[-\frac{1}{2}tr\left\{\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\Sigma}^{-1}\right\}\right] d\mathbf{B}_{\boldsymbol{\xi}}d\boldsymbol{\Sigma}\pi(\boldsymbol{\xi}) \\
= \quad & (2\pi)^{-\frac{|\boldsymbol{\xi}|q}{2}}|\mathbf{H}_{\boldsymbol{\xi}}|^{-\frac{q}{2}} \iint |\boldsymbol{\Sigma}|^{-\frac{1}{2}(n+|\boldsymbol{\xi}|+\delta+q+1)} \exp\{-\frac{1}{2}tr[(\mathbf{Y}^\top\mathbf{Y}+\mathbf{Q})\boldsymbol{\Sigma}^{-1}]\} \times \\
& \exp\left[-\frac{1}{2}tr\left\{\left(\mathbf{B}_{\boldsymbol{\xi}}^\top\mathbf{K}_{\boldsymbol{\xi}}\mathbf{B}_{\boldsymbol{\xi}} - 2\mathbf{B}_{\boldsymbol{\xi}}^\top\mathbf{K}_{\boldsymbol{\xi}}\mathbf{K}_{\boldsymbol{\xi}}^{-1}\mathbf{X}_{\boldsymbol{\xi}}^\top\mathbf{Y}\right)\boldsymbol{\Sigma}^{-1}\right\}\right] d\mathbf{B}_{\boldsymbol{\xi}}d\boldsymbol{\Sigma}\pi(\boldsymbol{\xi}) \\
= \quad & (2\pi)^{-\frac{|\boldsymbol{\xi}|q}{2}}|\mathbf{H}_{\boldsymbol{\xi}}|^{-\frac{q}{2}} \iint |\boldsymbol{\Sigma}|^{-\frac{1}{2}(n+|\boldsymbol{\xi}|+\delta+q+1)} \times \\
& \exp\{-\frac{1}{2}tr[(\mathbf{Y}^\top\mathbf{Y} - \tilde{\mathbf{B}}_{\boldsymbol{\xi}}^\top\mathbf{K}_{\boldsymbol{\xi}}\tilde{\mathbf{B}}_{\boldsymbol{\xi}} + \mathbf{Q})\boldsymbol{\Sigma}^{-1}]\} \times \\
& \exp\left[-\frac{1}{2}tr\left\{\left((\mathbf{B}_{\boldsymbol{\xi}}-\tilde{\mathbf{B}}_{\boldsymbol{\xi}})^\top\mathbf{K}_{\boldsymbol{\xi}}(\mathbf{B}_{\boldsymbol{\xi}}-\tilde{\mathbf{B}}_{\boldsymbol{\xi}})\right)\boldsymbol{\Sigma}^{-1}\right\}\right] d\mathbf{B}_{\boldsymbol{\xi}}d\boldsymbol{\Sigma}\pi(\boldsymbol{\xi}) \\
= \quad & (2\pi)^{-\frac{|\boldsymbol{\xi}|q}{2}}|\mathbf{H}_{\boldsymbol{\xi}}|^{-\frac{q}{2}}(2\pi)^{\frac{|\boldsymbol{\xi}|q}{2}}|\mathbf{K}_{\boldsymbol{\xi}}|^{-\frac{q}{2}} \times \\
& \int |\boldsymbol{\Sigma}|^{\frac{|\boldsymbol{\xi}|}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}(n+|\boldsymbol{\xi}|+\delta+q+1)} \exp\{-\frac{1}{2}tr[(\mathbf{Y}^\top\mathbf{Y} - \tilde{\mathbf{B}}_{\boldsymbol{\xi}}^\top\mathbf{K}_{\boldsymbol{\xi}}\tilde{\mathbf{B}}_{\boldsymbol{\xi}} + \mathbf{Q})\boldsymbol{\Sigma}^{-1}]\}d\boldsymbol{\Sigma}\pi(\boldsymbol{\xi}) \\
= \quad & |\mathbf{H}_{\boldsymbol{\xi}}|^{-\frac{q}{2}}|\mathbf{K}_{\boldsymbol{\xi}}|^{-\frac{q}{2}} \int |\boldsymbol{\Sigma}|^{-\frac{1}{2}(n+\delta+q+1)} \times \\
& \exp\{-\frac{1}{2}tr[(\mathbf{Y}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}_{\boldsymbol{\xi}}\mathbf{K}_{\boldsymbol{\xi}}^{-1}\mathbf{X}_{\boldsymbol{\xi}}^\top\mathbf{Y} + \mathbf{Q})\boldsymbol{\Sigma}^{-1}]\}d\boldsymbol{\Sigma}\pi(\boldsymbol{\xi}) \\
= \quad & |\mathbf{H}_{\boldsymbol{\xi}}\mathbf{K}_{\boldsymbol{\xi}}|^{-\frac{q}{2}} \int |\boldsymbol{\Sigma}|^{-\frac{1}{2}(n+\delta+q+1)} \times \\
& \exp\{-\frac{1}{2}tr[(\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{K}_{\boldsymbol{\xi}}^{-1}\mathbf{X}_{\boldsymbol{\xi}}^\top)\mathbf{Y} + \mathbf{Q})\boldsymbol{\Sigma}^{-1}]\}d\boldsymbol{\Sigma}\pi(\boldsymbol{\xi}) \\
\propto \quad & |\mathbf{H}_{\boldsymbol{\xi}}\mathbf{K}_{\boldsymbol{\xi}}|^{-\frac{q}{2}}|\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{X}_{\boldsymbol{\xi}}\mathbf{K}_{\boldsymbol{\xi}}^{-1}\mathbf{X}_{\boldsymbol{\xi}}^\top)\mathbf{Y} + \mathbf{Q}|^{-\frac{n+\delta}{2}}\pi(\boldsymbol{\xi}) \\
\equiv \quad & g(\boldsymbol{\xi}|\mathbf{Y}),
\end{aligned}
$$

where $\mathbf{K}_{\boldsymbol{\xi}} = \mathbf{X}_{\boldsymbol{\xi}}^{\top}\mathbf{X}_{\boldsymbol{\xi}} + \mathbf{H}_{\boldsymbol{\xi}}^{-1}$ and $g(\boldsymbol{\xi}|\mathbf{Y})$ is the proportional form of $\pi(\boldsymbol{\xi}|\mathbf{Y})$. (Brown and Vannucci, 1998) set $\mathbf{Q} = k\mathbf{I}_q$ and $\mathbf{H}_{\boldsymbol{\xi}} = c(\mathbf{X}_{\boldsymbol{\xi}}^{\top}\mathbf{X}_{\boldsymbol{\xi}})^{-1}$ and $\omega_j = \omega$, then $\pi(\boldsymbol{\xi}) = \prod_{j=1}^{p}\omega_j^{\xi_j}(1-\omega_j)^{1-\xi_j} = \omega^{|\boldsymbol{\xi}|}(1-\omega)^{p-|\boldsymbol{\xi}|}$. Apply $\mathbf{Q} = k\mathbf{I}_q$ and $\mathbf{H}_{\boldsymbol{\xi}} = c(\mathbf{X}_{\boldsymbol{\xi}}^{\top}\mathbf{X}_{\boldsymbol{\xi}})^{-1}$ into (2.1), hence we have

$$\begin{aligned}\log g(\boldsymbol{\xi}|\mathbf{Y}) \quad = \quad & -\frac{n+\delta}{2}\log|k\mathbf{I}_q + \mathbf{Y}^{\top}\mathbf{Y} - \frac{c}{c+1}\mathbf{Y}^{\top}\mathbf{X}_{\boldsymbol{\xi}}(\mathbf{X}_{\boldsymbol{\xi}}^{\top}\mathbf{X}_{\boldsymbol{\xi}}^{-1})\mathbf{X}_{\boldsymbol{\xi}}^{\top}\mathbf{Y}| - \\ & \frac{q|\boldsymbol{\xi}|}{2}\log(c+1) + |\boldsymbol{\xi}|\log\omega + (p-|\boldsymbol{\xi}|)\log(1-\omega).\end{aligned}$$

Exhaustive computation of $g(\boldsymbol{\xi}|\mathbf{Y})$ for 2^p values of $\boldsymbol{\xi}$ becomes prohibitive even in a super computer when p is larger than 40. In this circumstances, using MCMC sampling to explore marginal posterior is possible. (Brown and Vannucci, 1998) use Gibbs sampler to generate each $\boldsymbol{\xi}$ value component-wise from full conditional distributions $\pi(\xi_j|\boldsymbol{\xi}_{-j}, \mathbf{Y})$, where $\boldsymbol{\xi}_{-j} = \boldsymbol{\xi} \setminus \xi_j$. It is easy to show that

$$\pi(\xi_j = 1|\boldsymbol{\xi}_{-j}, \mathbf{Y}) = \frac{\theta_j}{\theta_j + 1},$$

where

$$\theta_j = \frac{g(\xi_j = 1, \boldsymbol{\xi}_{-j}|\mathbf{Y})}{g(\xi_j = 0, \boldsymbol{\xi}_{-j}|\mathbf{Y})}.$$

The complete algorithm of (Brown and Vannucci, 1998) is described in Algorithm 3:

Algorithm 3 Brown: MCMC sampling for $\boldsymbol{\xi}$

1. Initialize $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_p)$.
2. **Repeat** for $t = 1, 2, \ldots$ until convergence
   1). $a \leftarrow g(\xi_1^{(t)} = 1, \boldsymbol{\xi}_{-1}^{(t)}|\mathbf{Y})$ and $b \leftarrow g(\xi_1^{(t)} = 0, \boldsymbol{\xi}_{-1}^{(t)}|\mathbf{Y})$.
   2). $\theta_1 \leftarrow \frac{a}{b}$; $p_1 \leftarrow \frac{\theta_1}{\theta_1+1}$.
   3). $\xi^* \sim \mathrm{Ber}(p_1)$; $\xi_1^{(t)} \leftarrow \xi^*$.
   4). **Repeat** for $j = 2, \ldots, p$:
      i) $\kappa = a\xi^* + b(1-\xi^*)$; $\tilde{\boldsymbol{\xi}} \leftarrow \boldsymbol{\xi}^{(t)}$.
      ii) **If** $\xi_j^{(t)} = 1$,
         $\tilde{\xi}_j \leftarrow 0$; $a \leftarrow \kappa$; $b \leftarrow g(\tilde{\boldsymbol{\xi}}|\mathbf{Y})$.
      else
         $\tilde{\xi}_j \leftarrow 1$; $a \leftarrow g(\tilde{\boldsymbol{\xi}}|\mathbf{Y})$; $b \leftarrow \kappa$.
      iii) $\theta_j \leftarrow \frac{a}{b}$; $p_j \leftarrow \frac{\theta_j}{\theta_j+1}$.
      iv) $\xi^* \sim Ber(p_j)$; $\xi_j^{(t)} \leftarrow \xi^*$.
   5). Get $\boldsymbol{\xi}^{(t)}$

For the hyperparameter, we follow the same setting with (Brown and Vannucci, 1998), which are $\delta = 2 + q, \omega = 20/p, k = 2$ and $c = 10$. Note that since our definition of inverse-wishart distribution is different from that of (Brown and Vannucci, 1998), the values of q and k we set here are after adjustment to make them consistent with (Brown and Vannucci, 1998).

Bibliography

Brown, P. J. and Vannucci, M. (1998). Multivariate bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Methodological)*, pages 627–641.