Sparse Bayesian infinite factor models

A. Bhattacharya and D.B. Dunson

Biometrika (2011)

Presented by XianXing Zhang
Duke University

March 30, 2012
Overview of the discussed paper

- Sparse modelling of covariance matrices using latent factor models
  - Multiplicative gamma process shrinkage prior on factor loadings
    - It’s conjugate to the likelihood
    - Increasingly shrinking the loadings towards zero
    - Allowing introduction of infinitely many factors
  - Adaptive Gibbs sampler for automatically truncating the number of loadings
Overview of the discussed paper

- Sparse modelling of covariance matrices using latent factor models
- Multiplicative gamma process shrinkage prior on factor loadings
  - It’s conjugate to the likelihood
  - Increasingly shrinking the loadings towards zero
  - Allowing introduction of infinitely many factors
- Adaptive Gibbs sampler for automatically truncating the number of loadings
Overview of the discussed paper

- Sparse modelling of covariance matrices using latent factor models
- Multiplicative gamma process shrinkage prior on factor loadings
  - It’s conjugate to the likelihood
  - Increasingly shrinking the loadings towards zero
  - Allowing introduction of infinitely many factors
- Adaptive Gibbs sampler for automatically truncating the number of loadings
Background

- Factor models explain the dependence between data by decomposing covariance matrix $\Omega$ as $\Lambda\Lambda^T + \Sigma$ where $\Lambda$ is the factor loading matrix and $\Sigma$ is a diagonal matrix.

- Identifiability:
  - To constrain the loading matrix to be lower triangular with positive diagonal entries (Geweke & Zhou, 1996)
  - For purpose of prediction or inference on the covariance matrix, identifiability of the loadings is not necessary.

- How to do the inference on the number of factors:
  - Calculating marginal likelihood
  - Reversible jump Markov chain Monte Carlo algorithm (Lopes & West (2004))
  - Zeroing a subset of the loading elements using variable selection priors (Lucas et al., 2006; Carvalho et al., 2008)
Factor models explain the dependence between data by decomposing covariance matrix $\Omega$ as $\Lambda \Lambda^T + \Sigma$ where $\Lambda$ is the factor loading matrix and $\Sigma$ is a diagonal matrix.

**Identifiability:**
- To constrain the loading matrix to be lower triangular with positive diagonal entries (Geweke & Zhou, 1996)
- For purpose of prediction or inference on the covariance matrix, identifiability of the loadings is not necessary.

**How to do the inference on the number of factors**
- Calculating marginal likelihood
- Reversible jump Markov chain Monte Carlo algorithm (Lopes & West (2004))
- Zeroing a subset of the loading elements using variable selection priors (Lucas et al., 2006; Carvalho et al., 2008)
Factor models explain the dependence between data by decomposing covariance matrix $\Omega$ as $\Lambda \Lambda^T + \Sigma$ where $\Lambda$ is the factor loading matrix and $\Sigma$ is a diagonal matrix.

Identifiability:
- To constrain the loading matrix to be lower triangular with positive diagonal entries (Geweke & Zhou, 1996)
- For purpose of prediction or inference on the covariance matrix, identifiability of the loadings is not necessary

How to do the inference on the number of factors:
- Calculating marginal likelihood
- Reversible jump Markov chain Monte Carlo algorithm (Lopes & West (2004))
- Zeroing a subset of the loading elements using variable selection priors (Lucas et al., 2006; Carvalho et al., 2008)
Model specification

Infering number of factors

- Define a parameter-expanded loadings matrix space
  \( \Theta_\Lambda = \{ \Lambda = (\lambda_{jh}), j \leq p, h \leq \infty, \max_j \sum_{h=1}^{\infty} \lambda_{jh}^2 < \infty \} \)

- Define the prior on \( \Theta_\Lambda \)

  \[
  \lambda_{jh} | \phi_{jh}, \tau_h \sim N(0, \phi_{jh}^{-1} \tau_h^{-1}), \quad \phi_{jh} \sim Ga(\nu/2, \nu/2),
  \]

  \[
  \tau_h = \prod_{l=1}^{h} \delta_l, \quad \delta_1 \sim Ga(a_1, 1), \quad \delta_l \sim Ga(a_2, 1), \quad l \geq 2
  \]

- The degree of shrinkage increasing across the column index
Model specification

Infering number of factors

- Define a parameter-expanded loadings matrix space
  \( \Theta_{\Lambda} = \{ \Lambda = (\lambda_{jh}), j \leq p, h \leq \infty, \max_j \sum_{h=1}^{\infty} \lambda_{jh}^2 < \infty \} \)

- Define the prior on \( \Theta_{\Lambda} \)
  \[ \lambda_{jh} | \phi_{jh}, \tau_h \sim N(0, \phi_{jh}^{-1} \tau_h^{-1}), \quad \phi_{jh} \sim Ga(\nu/2, \nu/2), \]
  \[ \tau_h = \prod_{l=1}^{h} \delta_l, \quad \delta_1 \sim Ga(a_1, 1), \quad \delta_l \sim Ga(a_2, 1), \quad l \geq 2 \]

- The degree of shrinkage increasing across the column index
Infering number of factors

- Define a parameter-expanded loadings matrix space
  \[ \Theta_\Lambda = \{ \Lambda = (\lambda_{jh}), j \leq p, h \leq \infty, \max_j \sum_{h=1}^{\infty} \lambda_{jh}^2 < \infty \} \]

- Define the prior on \( \Theta_\Lambda \)

  \[
  \lambda_{jh} | \phi_{jh}, \tau_h \sim N(0, \phi_{jh}^{-1} \tau_h^{-1}), \quad \phi_{jh} \sim \text{Ga}(\nu/2, \nu/2), \]

  \[
  \tau_h = \prod_{l=1}^{h} \delta_l, \quad \delta_1 \sim \text{Ga}(a_1, 1), \quad \delta_l \sim \text{Ga}(a_2, 1), \quad l \geq 2
  \]

- The degree of shrinkage increasing across the column index
Properties of the shrinkage prior I

Truncation approximation error

- For computational reasons, we approximate the infinite loading matrix $\Omega$ by a matrix $\Omega_H$ with $H$ non-zero columns.

- Theoretical bounds on the truncation approximation error:

$$\Pr(d_\infty(\Omega - \Omega_H) > \epsilon) < \frac{6p}{\epsilon a_1(1 - 1/a_2)a_2^H}$$

for $\forall \epsilon > 0$ and $a_2 > 2$, $H > \log(6p/\epsilon a_1(1 - 1/a_2))/\log(a_2)$.
Properties of the shrinkage prior I

Truncation approximation error

- For computational reasons, we approximate the infinite loading matrix $\Omega$ by a matrix $\Omega^H$ with $H$ non-zero columns
- Theoretical bounds on the truncation approximation error

$$\text{pr}(d_\infty(\Omega - \Omega^H) > \epsilon) < \frac{6p}{\epsilon a_1 (1 - 1/a_2) a_2^H}$$

for $\forall \epsilon > 0$ and $a_2 > 2$, $H > \log(6p/\epsilon a_1 (1 - 1/a_2))/\log(a_2)$
Properties of the shrinkage prior II

Weak posterior consistency

Fix $\Omega_0 \in \Theta$, for any $\epsilon > 0$, there exists $\epsilon^* > 0$ such that

$$\{\Omega : d_\infty(\Omega_0, \Omega) < \epsilon^*\} \in \{\Omega : K(\Omega_0, \Omega) < \epsilon\},$$

where $K(\Omega_0, \Omega)$ is the KL divergence between $N(y; 0, \Omega_0)$ and $N(y; 0, \Omega)$.
Posterior Computation

- Conjugacy leads to a block Gibbs sampler with a fixed truncation level $k^*$
  - Adaptive Gibbs sampler is designed for inference on $k^*$
    - Trigger adaptations with probability $p(t) = \exp(\alpha_0 + \alpha_1 t)$ at the $t$th iteration
    - Discard the columns in the loadings having all elements within some pre-specified small neighbourhood of zero
    - Draw one from the prior distribution if the number of columns drops to zero
Posterior Computation

- Conjugacy leads to a block Gibbs sampler with a fixed truncation level $k^*$

- Adaptive Gibbs sampler is designed for inference on $k^*$
  - Trigger adaptations with probability $p(t) = \exp(\alpha_0 + \alpha_1 t)$ at the $t$th iteration
  - Discard the columns in the loadings having all elements within some pre-specified small neighbourhood of zero
  - Draw one from the prior distribution if the number of columns drops to zero
Results on estimating covariance matrix

Table 1. Comparative performance in covariance matrix estimation in the simulation study. The average, best and worst case performance across 50 simulation replicates in terms of mean square error ($\times 10^2$), average absolute bias ($\times 10^2$) and maximum absolute bias ($\times 10^2$) are tabulated for the different methods.

<table>
<thead>
<tr>
<th>true $(p, k)$ method</th>
<th>(100, 5)</th>
<th>(500, 10)</th>
<th>(1000, 15)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MGPS</td>
<td>Banding</td>
<td>MAP</td>
</tr>
<tr>
<td>MSE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>0.2</td>
<td>1.3</td>
<td>0.2</td>
</tr>
<tr>
<td>min</td>
<td>0.1</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>max</td>
<td>0.3</td>
<td>1.6</td>
<td>0.3</td>
</tr>
<tr>
<td>average absolute bias</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>1.9</td>
<td>3.1</td>
<td>1.0</td>
</tr>
<tr>
<td>min</td>
<td>1.3</td>
<td>2.5</td>
<td>0.6</td>
</tr>
<tr>
<td>max</td>
<td>2.5</td>
<td>4.9</td>
<td>1.5</td>
</tr>
<tr>
<td>maximum absolute bias</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>50.9</td>
<td>111.0</td>
<td>44.8</td>
</tr>
<tr>
<td>min</td>
<td>38.8</td>
<td>99.8</td>
<td>24.7</td>
</tr>
<tr>
<td>max</td>
<td>74.1</td>
<td>131.0</td>
<td>105.0</td>
</tr>
</tbody>
</table>

MGPS, posterior mean using our proposed multiplicative shrinkage prior; Banding, Banding sample covariance matrix; MAP, approximate maximum a posteriori estimate under our proposed prior; MSE, mean square error.
Predictive performance in factor regression

Posterior predictive distribution: \( f(z_{n+1}|x_{n+1}, y_1, \ldots, y_n) = \int f(z_{n+1}|x_{n+1}, \Omega)\pi(\Omega|y_1, \ldots, y_n)d\Omega \)

<table>
<thead>
<tr>
<th>true ((p, k))</th>
<th>(100, 5)</th>
<th>(500, 10)</th>
<th>(1000, 15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td>MGPS</td>
<td>Lasso</td>
<td>Elastic net</td>
</tr>
<tr>
<td>mspe</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>0.63</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>min</td>
<td>0.32</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>max</td>
<td>0.89</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td>aape</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>0.62</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td>min</td>
<td>0.47</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>max</td>
<td>0.85</td>
<td>0.73</td>
<td>0.72</td>
</tr>
<tr>
<td>mape</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>2.19</td>
<td>2.07</td>
<td>2.07</td>
</tr>
<tr>
<td>min</td>
<td>1.36</td>
<td>1.43</td>
<td>1.40</td>
</tr>
<tr>
<td>max</td>
<td>3.15</td>
<td>2.91</td>
<td>2.89</td>
</tr>
</tbody>
</table>

MGPS, our proposed multiplicative shrinkage prior; mspe, mean squared prediction error; aape, average absolute prediction error; mape, maximum absolute prediction error.