Sparse Bayesian infinite factor models

A. Bhattacharya, D. B. Dunson (2011)

Jae-Hoon Kim

November 22, 2023

Contents

1. Introduction

Factor models

Standard factor model:

$$X - \mu = LF + \epsilon, \quad \textit{where}$$

- $X = (X_i) \in \mathbb{R}^p$: random vector with $\mathrm{E}(X) = \mu$ and $\mathrm{Cov}(X) = \Sigma$
- $L = (l_{ij}) \in \mathbb{R}^{p \times m}$: matrix of factor loadings
- $F = (F_i) \in \mathbb{R}^m$: common factors ($m \leq p$)
- $\epsilon = (\epsilon_i) \in \mathbb{R}^p$: errors (or specific factors)

with assumptions:

- $\mathrm{E}(F) = 0$ and $\mathrm{Cov}(F) = I_m$
- $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Cov}(\epsilon) = \Psi = \mathrm{diag}\,(\psi_i)$
- $\mathrm{Cov}(\epsilon, F) = 0$

Factor models

Standard factor models

- Constraints for identifiability
  - Loading matrix to be lower triangular with positive diagonal entries.
- Undesirable properties due to the constraints
  - Priori order dependence in the off-diagonal entries of the covariance matrix.

A proposed model

- Parameter-expanded loading matrix
  - The assumptions such as dimension of loadings aren't pre-defined.
  - Making induced prior on the covariance matrix invariant to ordering the data.
- Adaptively select a truncation of the infinite loadings.
  - Facilitates the posterior computation.
  - Provides an accurate approximation to the infinite factor model.

2. Model and prior specification

Outline

1. **Model specification**
2. **Prior specification**
3. **Backgrounds & Properties of the model**

Model specification

We assume the model:

$$y_i = \Lambda \eta_i + \epsilon_i, \quad \epsilon_i \sim N_p\left(0, \sigma^2 I_p\right) \quad i = 1, \ldots, p,$$

- y_i is normalized p-dimensional continuous response
- Λ is $p \times k$ factor loading matrix
- $\eta_i \sim N_k(0, I_k)$ are latent factors
- ϵ_i is an idiosyncratic error with covariance $\Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2)$

Model specification

Marginally,

$$y_i \sim N_p(0, \Omega) \quad \textit{with} \quad \Omega = \Lambda\Lambda^T + \Sigma$$

- y_i is assumed to have independent components given the factors
- The dependence among the components is induced by marginalizing over the distribution of the factors

Prior specification

We use a shrinkage-type prior with the degree of shrinkage increasing across the column index:

$$\lambda_{jh} \sim N(0, \phi_{jh}\tau_h),$$

$$\phi_{jh} \sim IG\left(\tfrac{\nu}{2}, \tfrac{\nu}{2}\right), \quad j = 1, \ldots, p,$$

$$\tau_h = \textstyle\prod_{l=1}^{h} \delta_l, \quad h = 1, \ldots, \infty,$$

$$\delta_1 \sim IG(a_1, 1), \quad \delta_l \sim IG(a_2, 1), \quad l \geqslant 2,$$

$$\sigma^2 \sim IG(a_\sigma, b_\sigma).$$

Prior specification

Prior assumption

- δ_l $(l = 1, \ldots, \infty)$ are independent
- τ_h is a global shrinkage parameter for the hth column
- ϕ_{jh}s are local shrinkage parameters for the elements in the hth column

Backgrounds & Properties of the model

Condition 1

All the entries of $\Lambda\Lambda^T$ are finite if and only if

$$\Theta_\Lambda = \left\{ \Lambda = (\lambda_{jh}),\ j = 1, \ldots, p,\ h = 1, \ldots, \infty,\ \max_{1 \leqslant j \leqslant p} \sum_{h=1}^{\infty} \lambda_{jh}^2 \leqslant \infty \right\}$$

where Θ_Λ denote the collection of all matrices Λ with p rows and infinitely many columns such that $\Lambda\Lambda^T$ is a $p \times p$ matrix with all entries finite.

Backgrounds & Properties of the model

The image of $\Theta_\Lambda \times \Theta_\Sigma$ under the function g is the set Θ.

Lemma 1

Let Θ_Σ, Θ denote the set of non-negative diagonal and definite matrix respectively.
For any $(\Lambda, \Sigma) \in \Theta_\Lambda \times \Theta_\Sigma$, we have $g(\Lambda, \Sigma) \in \Theta$,
where the function $g : \Theta_\Lambda \times \Theta_\Sigma \to \Theta$ corresponding to $g(\Lambda, \Sigma) = \Lambda\Lambda^T + \Sigma$.

We can make sure that draws from the above prior are elements of $\Theta_\Lambda \times \Theta_\Sigma$ almost surely.

Proposition 1

$$\textit{If } (\Lambda, \Sigma) \sim \Pi_\Lambda \otimes \Pi_\Sigma, \textit{ then } \Pi_\Lambda \otimes \Pi_\Sigma (\Theta_\Lambda \times \Theta_\Sigma) = 1.$$

Backgrounds & Properties of the model

We can use this regularity property of g to prove sup-norm support of the proposed prior.

Lemma 2

Let (Λ_0, Σ_0) be an arbitrary element of $\Theta_\Lambda \times \Theta_\Sigma$ and ϵ-ball around (Λ_0, Σ_0) be

$$B_\epsilon(\Lambda_0, \Sigma_0) = \{(\Lambda, \Sigma) \in \Theta_\Lambda \times \Theta_\Sigma : d_2(\Lambda, \Lambda_0) < \epsilon, d_\infty(\Sigma, \Sigma_0) < \epsilon\}, \text{ for } \epsilon > 0,$$

where $d_2(\cdot, \cdot)$ denotes the L_2 distance, and $d_\infty = \max_{1 \le r,s \le p} |a_{rs} - b_{rs}|$ where a_{rs}, b_{rs} are sup-norm metric.
Then $g\{B_\epsilon(\Lambda_0, \Sigma_0)\}$ contains values $\Omega \in \Theta$ in $B^\infty_{\epsilon^*}(\Omega_0)$, with ϵ^* decreasing towards zero monotonically as ϵ decreases to zero.

Backgrounds & Properties of the model

We can obtain theoretical bounds on the truncation approximation error as justification.

Theorem 1

If $a_2 \geq 2$, then for any $\epsilon > 0$,

$$pr\left\{d_\infty\left(\Omega, \Omega_H\right) > \epsilon\right\} < \frac{6pb}{\epsilon\left(1-a\right)}a^H, \quad for\ H > log\left\{\frac{6pb}{\epsilon\left(1-a\right)}\right\}/log\left(\frac{1}{a}\right),$$

where $b = E\left(\delta_1\right)$, $a = E\left(\delta_2\right)$, $\Omega_H = \Lambda_H\Lambda_H^T + \Sigma$, and Λ_H denotes the matrix obtained by discarding the columns of Λ from H+1.

Backgrounds & Properties of the model

Proposed prior has large support to place positive probability in arbitrarily small neighborhoods around any covariance matrix.

Proposition 2

Let Π be the induced prior on Θ, $\Pi = (\Pi_\Lambda \otimes \Pi_\Sigma) \circ g^{-1}$.
If Σ_0 is any $p \times p$ covariance matrix and $B_\epsilon^\infty(\Omega_0)$ is an ϵ-neighborhood of Ω_0 under the sup-norm, then

$$\Pi\{B_\epsilon^\infty(\Omega_0)\} > 0 \quad \textit{for any } \epsilon > 0.$$

Finally, the posterior distribution of Ω is weakly consistent.

Theorem 2

For any ϵ, there exists ϵ^* such that $\{\Omega : d_\infty(\Omega_0, \Omega) < \epsilon^*\} \subset \{\Omega : K(\Omega_0, \Omega) < \epsilon\}$, where $K(\Omega_0, \Omega)$ denotes the Kullback-Leibler divergence between $N(0,\ \Omega_0)$ and $N(0,\ \Omega)$.

Backgrounds & Properties of the model

Another attractive properties

Proposed prior is free of order dependence

Property 1

The prior on Ω is invariant to permutations with Ω having the same distribution as Ω_π, where $\Omega_\pi = (w_{\pi_r \pi_s})$ with π any permutation of $\{1, \ldots, p\}$ and $\Omega = (w_{rs})$

We can ensure $E\left(w_{rs}^2\right)$ is finite.

Property 2

If $d = E\left(\delta_1^2\right)$ is finite and $c = E\left(\delta_2^2\right) < 1$,
$E\left(w_{rs}^2\right)$ is finite where $w_{rs}^2 = \sum_{h=1}^{\infty} \lambda_{rh}\lambda_{sh} = \lambda_r^T \lambda_s$

3. Posterior computation

Outline

1. **Choosing the number of factors**
2. **Gibbs sampler with a fixed truncation level**

Choosing the number of factors

Define $k^{*(t)} = \tilde{k} - m^{(t)}$ to the effective number of factors at iteration t,
where $\tilde{k}$ is conservative number of factors and $m^{(t)}$ is number of factors having negligible contribution at iteration t.

Truncation procedure

1. Let $u_t \sim Unif\,(0,1)$ and $p(t) = exp(a_0 + a_1 t)$. Set a_0 and a_1 so that adaptation occurs around every 10 iterations
2. If $u_t \leq p(t)$, monitor the columns in the loadings having all elements within some specified small neighborhood of zero.
3. If there is no such columns, add a column to the loadings and sample parameters from the prior distribution to fill in additional columns.

3-1. Otherwise, discard the redundant columns and retain parameters corresponding to the non-redundant columns.

4. After burn-in, use the median or mode of $\left\{k^{*(t)}\right\}$ as an estimate of $k^{*(t)}$ with credible interval quantifying uncertainty.

Gibbs sampler with a fixed truncation level

After truncating the loading matrix to have $k^* \ll p$ columns, we can conduct a straightforward Gibbs Sampler.

1. **Step 1** Let λ_j is jth row of Λ_{k^*}. Draw λ_j from full conditional posterior:

$$\pi\left(\lambda_j \mid -\right) \sim N_{k^*}\left\{\left(D_j+\sigma_j^{-2}\eta^{\mathrm{T}}\eta\right)^{-1}\eta^{\mathrm{T}}\sigma_j^{-2}y^{(j)},\left(D_j+\sigma_j^{-2}\eta^{\mathrm{T}}\eta\right)^{-1}\right\}$$

where $\eta=\left(\eta_1,\ldots,\eta_n\right)^{\mathrm{T}}$, $D_j=\mathrm{diag}\left(\phi_{j1}\tau_1,\ldots,\phi_{jk}\tau_k\right)$ and $y^{(j)}=\left(y_{1j},\ldots,y_{nj}\right)^{\mathrm{T}}$ for $j=1,\ldots,p$.

Gibbs sampler with a fixed truncation level

2. **Step 2** Sample $\sigma_j^2, j = 1 \ldots, p$, from conditionally independent posteriors:

$$\pi\left(\sigma_j^2 \mid -\right) \sim \mathrm{IG}\left\{a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^{n}\left(y_{ij} - \lambda_j^{\mathrm{T}}\eta_i\right)^2\right\}$$

3. **Step 3** Sample $\eta_i, i = 1 \ldots, n$, from conditionally independent posteriors

$$\pi\left(\eta_i \mid -\right) \sim N_k\left\{\left(I_k + \Lambda_k^{\mathrm{T}}\Sigma^{-1}\Lambda_k\right)^{-1}\Lambda_k^{\mathrm{T}}\Sigma^{-1}y_i, \left(I_k + \Lambda_k^{\mathrm{T}}\Sigma^{-1}\Lambda_k\right)^{-1}\right\}$$

4. **Step 4** Sample ϕ_{jh} from

$$\pi\left(\phi_{jh} \mid -\right) \sim \mathrm{IG}\left(\frac{v+1}{2}, \frac{v + \tau_h\lambda_{jh}^2}{2}\right)$$

Gibbs sampler with a fixed truncation level

5. **Step 5** Sample δ_1 from

$$\pi\left(\delta_1 \mid -\right) \sim \mathrm{IG}\left\{a_1 + \frac{pk}{2}, 1 + \frac{1}{2}\sum_{l=1}^{k} \tau_l^{(1)} \sum_{j=1}^{p} \phi_{jl} \lambda_{jl}^2\right\}$$

and for $h \geqslant 2$, sample δ_h from

$$\pi\left(\delta_h \mid -\right) \sim \mathrm{IG}\left\{a_2 + \frac{p}{2}\left(k - h + 1\right), 1 + \frac{1}{2}\sum_{l=h}^{k} \tau_l^{(h)} \sum_{j=1}^{p} \phi_{jl} \lambda_{jl}^2\right\}$$

where $\tau_l^{(h)} = \prod_{t=1, t\neq h}^{l} \delta_t$ for $h = 1, \ldots, k^*$.

6. **Step 6** Update a_1 and a_2 using a Metropolis-Hastings step within the Gibbs sampler.

Gibbs sampler with a fixed truncation level

After a reasonable burn-in,
$\Omega^{(t)} = \Lambda^{(t)}_{\tilde{k}^{(t)}} \Lambda^{(t)}_{\tilde{k}^{(t)}} + \Sigma^{(t)}$ represent draws from the marginal posterior distribution of Ω given $y_1, \ldots, y_n$, where $\left\{ \Lambda^{(t)}_{\tilde{k}^{(t)}}, \Sigma^{(t)} \right\}$ denotes posterior samples at the t^{th} iteration.

Alternative method: EM approach
For computational efficiency, we can use MAP method by replacing posterior distributions of $\Lambda_{\tilde{k}}, \Sigma,$ *and* ϕ in Steps 1, 2, and 4 by the respective conditional posterior modes.

4. Simulation example

Outline

1. **Factor selection and covariance matrix estimation**
2. **Latent factor regression: Predictive performance**
3. **Latent factor regression: Estimating regression coefficients**
4. **Latent factor regression: Variable selection**

Factor selection and covariance matrix estimation

We simulated the model:

$$y_i \sim N_p(0, \Omega) \quad i = 1, \ldots, 200, \ \textit{with } \Omega = \Lambda\Lambda^T + \Sigma$$

.

- The diagonal elements of Σ are drawn independently from a $IG(1, 0.25)$.
- The number of non-zero elements in each column of are chosen linearly between $2k$ and $k + 1$ in a decreasing fashion.
- the location of the zeros in each column are allocated randomly and simulate the nonzero elements independently from $N(0, 9)$.
- Monitor the columns in the loadings having all elements less than 10^{-4}.
- Hyperparameter settings:
  - $\sigma_j^{-2} \sim IG(1, 0.3)$,
  - $\phi_{jh} \sim \left(\frac{3}{2}, \frac{3}{2}\right)$,
  - $\delta_1, \delta_2 \sim IG(2, 1)$,
  - $p(t) = exp(-1 + 10^{-4}t)$.

Factor selection and covariance matrix estimation

We compared the performances of three methods:

- MGPS: Proposed model using multiplicative gamma process shrinkage prior.
- Banding: A covariance estimation method by Bickel & Levina(2008).
- MAP: Proposed model using EM algorithm.

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*

true (p, k)	(100, 5)			(500, 10)			(1000, 15)		
method	MGPS	Banding	MAP	MGPS	Banding	MAP	MGPS	Banding	MAP
MSE									
mean	0·2	1·3	0·2	0·10	0·4	0·10	0·10	0·3	0·10
min	0·1	0·9	0·1	0·02	0·4	0·05	0·02	0·2	0·05
max	0·3	1·6	0·3	0·20	0·5	0·30	0·4	0·5	0·30
average absolute bias									
mean	1·9	3·1	1·0	0·6	0·6	0·3	0·4	0·5	0·3
min	1·3	2·5	0·6	0·4	0·6	0·2	0·2	0·4	0·2
max	2·5	4·9	1·5	0·9	0·9	0·5	0·6	0·5	0·5
maximum absolute bias									
mean	50·9	111·0	44·8	95·4	117·8	97·7	115·0	115	108·0
min	38·8	99·8	24·7	50·2	105·0	64·4	52·6	111	74·7
max	74·1	131·0	105·0	152·0	131·0	162·0	242·0	240	221·0

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.

Latent factor regression: Predictive performance

Simulation Description:

- Let $y_i = (z_i, x_i^T)^T,\ i = 1, \ldots, n$,
  where x_is are (p-1) dimensional predictors and z_is are the response.
- The objective is to predict the response z_{n+1} based on x_{n+1} and $y_1, \ldots, y_n$.
- Then, the posterior predictive distribution of $z_{n+1}|x_{n+1}, y_1, \ldots, y_n$ is

$$f(z_{n+1}|x_{n+1}, y_1, \ldots, y_n) = \int f(z_{n+1}|x_{n+1}, \Omega)\ \pi(\Omega|y_1, \ldots, y_n) d\Omega.$$

- We randomly selected two locations in the first row of Λ and assigned values 1 and -1 to those locations, with the other elements set to zero.
- The remaining rows of the loadings were simulated as mentioned before.
- We used a randomly chosen training set of size 100 and held out the z_i s for the remaining 100 samples.

Latent factor regression: Predictive performance

We compared the performances of three methods:

- MGPS: Proposed model using multiplicative gamma process shrinkage prior
- Lasso
- Elastic net

Table 2. *Predictive performance in the simulation study. Average, best and worst case performance across* 50 *simulation replicates are reported for the different methods*

true (p, k)	(100, 5)			(500, 10)			(1000, 15)		
method	MGPS	Lasso	Elastic net	MGPS	Lasso	Elastic net	MGPS	Lasso	Elastic net
mspe									
mean	0·63	0·55	0·55	0·41	0·38	0·38	0·95	0·87	0·88
min	0·32	0·33	0·33	0·18	0·22	0·22	0·57	0·55	0·56
max	0·89	0·79	0·78	0·86	0·57	0·56	1·48	1·44	1·44
aape									
mean	0·62	0·59	0·59	0·51	0·49	0·49	0·80	0·77	0·75
min	0·47	0·47	0·47	0·33	0·38	0·37	0·60	0·59	0·59
max	0·85	0·73	0·72	0·80	0·58	0·59	0·99	0·98	0·99
mape									
mean	2·19	2·07	2·07	1·71	1·66	1·68	2·54	2·48	2·48
min	1·36	1·43	1·40	1·21	1·17	1·18	1·83	1·83	1·80
max	3·15	2·91	2·89	2·95	2·70	2·63	3·27	3·07	3·07

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

Latent factor regression: Estimating coefficients

Simulation Description:

- The joint Gaussian model implies that $E(z_i \mid x_i) = x_i^T \beta$, where $\beta = \Omega_{xx}^{-1}\Omega_{zx}$, with the Ω matrix suitably partitioned
- The elements of the $(p-1)$-dimensional vector β can be considered as the true regression coefficients of z on x.
- Letting $\Omega^{(t)}$ denote the posterior samples of Ω, $\beta^{(t)} = \left\{\Omega_{xx}^{(t)}\right\}^{-1} \Omega_{zx}^{(t)}$ give samples from the posterior distribution of β.
- Compute error between β and $\beta^{(t)}$

Latent factor regression: Estimating coefficients

We compared the performances of three methods:

Table 3. *Performance in estimating regression coefficients in the simulation study. We report the mean square error* ($\times 10^3$)*, average absolute bias* ($\times 10^3$) *and maximum absolute bias* ($\times 10^3$) *averaged across* 50 *simulation replicates for the different methods*

true (p, k)	(100, 5)			(500, 10)			(1000, 15)		
method	MGPS	Lasso	Elastic net	MGPS	Lasso	Elastic net	MGPS	Lasso	Elastic net
MSE	1·1	1·2	1·3	0·1	0·3	0·4	0·0	0·1	0·1
aab	10·1	12·4	13·0	1·7	3·9	4·1	0·9	1·8	1·9
mab	176·1	207·3	211·3	172·5	253·3	244·5	102·6	109·0	122·6

MGPS, our proposed multiplicative shrinkage prior; MSE, mean squared error; aab, average absolute bias; mab, maximum absolute bias.

Latent factor regression: Variable selection

Thresholding method for variable selection:

- Let $\hat{\beta}_{(1)} < \cdots < \hat{\beta}_{(p-1)}$ denote the ordered values of the posterior means for the $p-1$ predictors, and let $\pi_j = h$ denote that the jth predictor is the hth smallest in magnitude.
- Sets $\beta_j = 0$ for all j with $\pi_j \leqslant \tilde{h}$, with $\tilde{h}$ chosen to minimize the mean squared prediction error.

Latent factor regression: Variable selection

We compared the performances of three methods:

Table 4. *Variable selection performance in the simulation study. Percentage of false positives and power in detecting the true signal reported across* 50 *simulation replicates* (*average, best and worst case*) *for the different methods*

true (p, k)		(100, 5)			(500, 10)			(1000, 15)	
method	MGPS	Lasso	Elastic net	MGPS	Lasso	Elastic net	MGPS	Lasso	Elastic net
false positives (%)									
mean	0	9	7	0	4·0	3	0	3·0	2·0
min	0	0	0	0	0·2	0	0	0·7	0·7
max	0	26	25	0	14·0	14	0	8·0	10·0
power (%)									
mean	72	76	77	75	76	77	71	72	72
min	68	72	74	73	75	76	70	71	71
max	81	80	83	80	79	79	73	73	72

MGPS, our proposed multiplicative shrinkage prior.

Simulation example

Remarks

- The simulation study clearly highlights improved performance over competitors.
- The prior does not allow any of the loading elements to be exactly zero.
- The results were robust even when the model is not applicable, such as with the errors drawn from an AR(1) process.
- The adaptive method for factor selection was robust with respect to the choice of the threshold, even with a threshold as small as 10^{-9}.

5. Real data application

Diffuse large B-cell lymphoma application

About the data and existing analyses

- Lymphoma is a cancer of the white blood cell which occurs when lymphocytes have abnormal growth.
- Diffuse large B-cell lymphoma is the most common lymphoma.
- Rosenwald et al.(2002) used hierarchical clustering to identify four signature groups whose expressions were correlated with the survival times.
- Gui & Li(2005), Segal(2006), Ma & Huang(2007) analyzed the data using penalized methods.

Diffuse large B-cell lymphoma application

Our approach

- Our interest lies in simultaneously identifying an important subset of the features and obtaining a predictive model for the exact survival times.
- Let

$$y_i = (z_i, x_i^T)^T \text{ where } z_i = log(1 + T_i),$$

where T_i denotes the survival time for the ith patient and x_i denotes the corresponding 7399 dimensional feature vector.

- There were 72 patients in the training set whose survival times were right-censored.
- We thresholded the posterior mean of the regression coefficients as described in §4.4 to perform a variable selection.

Diffuse large B-cell lymphoma application

Results of variable selection

- The approach selected 17 features, with all of the features belonging to signature groups mentioned in Rosenwald et al.(2002).
- The ones with GenBank ID AA729055, AA805575 and X59812 also appear in Gui & Li(2005) and Segal(2006).
- Unlike the penalization method, our approach is designed to allow selection of highly correlated predictors into the same model.

Table 5. *Feature selection in the diffuse large-B-cell lymphoma data*

Unique ID	GenBank ID	Signature	Description
24094	AI476194	lymph	CD63 antigen (melanoma 1 antigen)
17048	AA085368	lymph	CD63 antigen (melanoma 1 antigen)
29636	NM005194	lymph	enhancer binding protein (C/EBP), β
34818	U83461	lymph	solute carrier family 31 (copper transporters), member 2
24394	AA729055	MHC	major histocompatibility complex, class II, DR α

Lymph, lymph-node signature; MHC, major histocompatibility complex; GenBank, National Institute of Health genetic sequence database.

Diffuse large B-cell lymphoma application

Results of predicting survival time

- The 95% predictive intervals for the survival times in the test sample were wide and contained the true survival times for the uncensored observations in all the cases.
- The mean square prediction error and mean absolute prediction error for the uncensored observations were 1.31 and 0.89.
- The same for lasso trained with the uncensored observations in the training sample were 1.28 and 0.90.
- The conclusions were unchanged even with varying v, initial values of a_1, a_2.

Q & A