Sparse Bayesian infinite factor models

Ganchao Wei

October 27, 2021

Overview

Introduction

The generic form of a latent factor model:

$$y_i = \Lambda \eta_i + \epsilon_i$$

, where $y_i \in \mathbb{R}^p$, $\Lambda \in \mathbb{R}^{p \times k}$, $\eta_i \sim N_k(9, I_k)$, $\epsilon_i \sim N_p(0, \Sigma)$ and $\Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2)$. Therefore, marginally, $y_i \sim N_p(0, \Omega)$ with $\Omega = \Lambda\Lambda^T + \Sigma$.

In traditional factor analysis, we need to put constraints on loading. However, from Bayesian perspective, one does not require identifiability of the loading elements for a wide class of applications. By making use of this, they define the prior on a parameter-expanded loadings matrix with redundant parameters, resulting in better computational properties while simplifying the theory.

Prior Specification

Prior for Σ: usual inverse gamma priors on the diagonal elements.

Denote $\Lambda = (\lambda_{jh})$, for $j = 1, \ldots, p$ and $h = 1, \ldots, \infty$. The entries of Λ decrease in magnitude as the column index increases, without any restrictions. More specifically, they use a shrinkage-type prior with the degree of shrinkage increasing across the column index as follows,

$$\lambda_{jh} \mid \phi_{jh}, \tau_h \sim N(0, \phi_{jh}^{-1}\tau_h^{-1}), \quad \phi_{jh} \sim \mathrm{Ga}(\nu/2, \nu/2), \ \tau_h = \prod_{l=1}^{h} \delta_l,$$

$$\delta_1 \sim \mathrm{Ga}(a_1, 1), \quad \delta_l \sim \mathrm{Ga}(a_2, 1), \quad l \geqslant 2, \quad \sigma_j^{-2} \sim \mathrm{Ga}(a_\sigma, b_\sigma) \quad (j = 1, \ldots, p),$$

Prior Specification

$$\lambda_{jh} \mid \phi_{jh}, \tau_h \sim N(0, \phi_{jh}^{-1}\tau_h^{-1}), \quad \phi_{jh} \sim \mathrm{Ga}(\nu/2, \nu/2), \ \tau_h = \prod_{l=1}^{h} \delta_l,$$

$$\delta_1 \sim \mathrm{Ga}(a_1, 1), \quad \delta_l \sim \mathrm{Ga}(a_2, 1), \quad l \geqslant 2, \quad \sigma_j^{-2} \sim \mathrm{Ga}(a_\sigma, b_\sigma) \quad (j = 1, \ldots, p),$$

- $\delta_l (l = 1, \ldots, \infty,)$ are independent
- τ_h is a global shrinkage parameter
- $\phi_{jh}s$ are local shrinkage parameters
- τ_h are stochatially increasing under the restriction $a_2 > 1$

For the shrinkage prior, we can show the weak consistency of posterior and the prior is free of order dependence.

They further place gamma priors on a_1 and a_2 to learn these key hyperparameters from the data.

Gibbs sampler with a fixed truncation level

After truncating the loading matrix to have $k^* \ll p$

Step 1. If we denote the jth row of Λ_{k^*} by λ_j^{T} , then the λ_js have independent conditionally conjugate posteriors,

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

where $\eta = (\eta_1, \ldots, \eta_n)^{\mathrm{T}}$, $D_j^{-1} = \mathrm{diag}(\phi_{j1}\tau_1, \ldots, \phi_{jk^*}\tau_{k^*})$ and $y^{(j)} = (y_{1j}, \ldots, y_{nj})^{\mathrm{T}}$ for $j = 1, \ldots, p$. Given the other parameters, $\pi(\lambda_j \mid -)$ denotes the conditional posterior of λ_j.

Step 2. Sample σ_j^{-2}, $j = 1 \ldots, p$, from conditionally independent posteriors

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

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

$$\pi(\eta_i \mid -) \sim N_{k^*}\left\{\left(I_{k^*} + \Lambda_{k^*}^{\mathrm{T}}\Sigma^{-1}\Lambda_{k^*}\right)^{-1}\Lambda_{k^*}^{\mathrm{T}}\Sigma^{-1}y_i, \left(I_{k^*} + \Lambda_{k^*}^{\mathrm{T}}\Sigma^{-1}\Lambda_{k^*}\right)^{-1}\right\}.$$

Gibbs sampler with a fixed

Step 4. Sample ϕ_{jh} from

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

Step 5. Sample δ_1 from

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

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

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

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

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

Choosing the number of factors adaptively

Do it with adaptive Gibbs sampler: tune the number of factors as the sampler progresses.

- adapt with probability $p(t) = exp(\alpha_0 + \alpha_1 t)$ at the tth iteration
- α_0 and α_1 are chosen so that adaptation occurs around every 10 iterations at the beginning of the chain, but decreases in frequency exponentially fast.
- Generate $u_t \sim U(0, 1)$. If $u_t \leq p(t)$, monitor the columns in the loadings having all elements within some pre-specified small neighborhood of 0.
- if the number of such columns drops to 0 $\Rightarrow$ generate new column from prior
- Otherwise, discard the redundant columns.

Factor selection and covariance matrix estimation

Simulation settings:

- $y_i \in \mathbb{R}^p$ for $i = 1, \ldots, 200$. The diagonal elements of $\Sigma^{-}1$ are drawn independently from Ga(1, 0.25).
- They choose 3 combinations of (p,k): (100,5), (500, 10) and (1000, 15)
- For each pair, there are 50 simulation replicates
- 25000 iterations with 5000 burn-in, and collect every 5th sample to thin the chain
- 3 methods are compared: (1) the proposedmethod MGPS, (2) banding sample covariance and (3) EM for MAP.

Factor selection and covariance matrix estimation

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

Instead of doing regularization (e.g. LASSO & elastic net), we can regard it as the latent factor regression problem.
Consider $E(z_i|x_i) = x_i^T\beta$, with $\beta = \Omega_{xx}^{-1}\Omega_{zx}$ and $\Omega_{xx} = \Lambda_x\Lambda_x^T + \Sigma_{xx}$. Just use the sub-blocks of factor analysis results, then everything is done.

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

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.

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.

Diffuse Large-B-Cell Lymphoma Application

Goal: (1) simultaneously identifying important features and (2) obtain a predictive model for the exact survival times.

- Let T_i denote the survival time for ith patient and let x_i denote the corresponding 7399 dimensional feature vector, for $i = 1, \ldots, 72$.
- Combine all the data into $y_i = (z_i, x_i^T)^T$ and $z_i = \log(1 + T_i)$
- Then do latent factor regression as in simulation example.

After fitting the model, they find 17 features, which matches previous results.