Variational Bayes for nonparametric sparse factor analysis

Bo Y.-C. Ning

University of California, Davis
4242 Mathematical Science Building
Department of Statistics
One Shield Avenue, Davis, CA 95618
e-mail: bycning@ucdavis.edu

Abstract: We study the mean-field variational Bayesian for sparse factor analysis with a sparsity is imposed through the spike-and-slab Indian Buffet Process (SS-IBP) prior. We propose a PXL-CAVI method using parameter expansion to compute the variational posterior. This method can achieve a faster convergence than the regular CAVI method without adopting the parameter expansion trick. We obtain sufficient conditions for the variational posterior to contract at the fast rate. Those conditions provide some guidance for choosing suitable values of hyperparameters in the prior in practice. Simulation studies are conducted to compare various methods under different settings. Our simulation results suggest an excellent performance in the finite sample setting. Finally, the method is applied to study a lung cancer dataset to identify important biological relevant genes.

MSC2020 subject classifications: Primary 62C10, 62H25, 62J07.

Keywords and phrases: Mean-field variational inference, PXL-CAVI method, Spike and slab prior, Sparse factor analysis, Sparse PCA.

1. Introduction

Consider the classical latent factor model

$$X_i = Zw_i + \epsilon_i,\ w_i \overset{i.i.d}{\sim} \mathcal{N}(0, I_K),\ \epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma),\ i = 1, \ldots, n, \tag{1}$$

where $X_i \in \mathbb{R}^p$ is the observed data, $Z \in \mathbb{R}^{p\times K}$ is the factor loading matrix, $K \in \mathbb{N}^+$ is the number of latent factors, $w_i \in \mathbb{R}^K$ and $\epsilon_i \in \mathbb{R}^p$ are mutually independent, and $\Sigma \in \mathbb{R}^p$ is a diagonal matrix $\text{diag}(\Sigma) = (\sigma_1^2, \ldots, \sigma_p^2)$. By marginalizing out w_i, this model can be written as $X_i \overset{i.i.d}{\sim} \mathcal{N}(0, ZZ' + \Sigma)$.

We study Model (1) under the high-dimensional setting $p > n > K$. Let $Z_{\cdot k} = (Z_{1k}, \ldots, Z_{pk})'$ be the k-th column of Z, $k = 1, \ldots, K$ and $Z^\star_{\cdot k}$ be the 'true' vector, we assume each $Z^\star_{\cdot k} \in \ell_0[s_k]$, where

$$\ell_0[s] = \{\mathsf{z} \in \mathbb{R}^p, \#\{j \in \{1, \ldots, p\} : |\mathsf{z}_j| > 0\} \le s \ll p\}. \tag{2}$$

That is, each $Z^\star_{\cdot k}$ has at most s_k nonzero coordinates. Denote S_k as the support of $Z_{\cdot k}$, $S_k = \text{supp}(Z_{\cdot k})$, which is the set containing the indices of non-zero coordinates in $Z_{\cdot k}$. Then, $\text{supp}(Z) = S_1 \cup S_2 \cup \cdots \cup S_K$.

2. The spike-and-slab Indian Buffet Process prior

Let Z_{jk} be the (j,k)-th coordinate in Z. We introduce a set of binary variables (γ_{jk}), $j = 1, \ldots, p$, $k = 1, \ldots, K$, each of which indicating whether Z_{jk} is zero or non-zero. If $\gamma_{jk} = 1$, we model Z_{jk} as a continuous density; if $\gamma_{jk} = 0$, we model Z_{jk} using a Dirac measure concentrating at zero. This construction leads to the spike-and-slab prior given as follows:

$$p(Z_{jk} \,|\, \gamma_{jk}, \lambda_k) = \gamma_{jk}\psi(Z_{jk} \,|\, \lambda_k) + (1 - \gamma_{jk})\delta_0, \tag{3}$$

where $\psi(Z_{jk} \,|\, \lambda_k) = \frac{\lambda_k}{2} \exp\{-\lambda_k |Z_{jk}|\}$ is a Laplace density. We further denote $\boldsymbol{\gamma}$ a $p \times K$ matrix whose (j,k)-th coordinate is γ_{jk}, then $\boldsymbol{\gamma}$ is a sparse binary matrix that has finite rows and possibly infinite columns (as $K \to \infty$). It is often expect that the number of non-zero factors is much smaller than K; hence, an ideal prior for $\boldsymbol{\gamma}$ should impose sparsity not only within each column but also on the number of active features (i.e., nonzero number of factor loadings). A popular prior is the Indian buffet process (IBP) prior introduced by Knowles and Ghahramani (2011) in the sparse factor analysis literature. The IBP prior is a hierarchical prior. It first puts an independent Bernoulli distribution with the success probability θ_k on each γ_{jk}, and then a Beta distribution on each θ_k. The explicit form of the IBP prior is given as follows:

$$\boldsymbol{\gamma} \,|\, \theta_1, \ldots, \theta_K \sim \prod_{j=1}^{p} \prod_{k=1}^{K} \theta_k^{\gamma_{jk}} (1 - \theta_k)^{1 - \gamma_{jk}}, \quad \theta_k \,|\, \alpha \sim \text{Beta}(\alpha/K, 1). \tag{4}$$

Last, we put independent priors on $\sigma_j^2 \sim \text{IG}(a, b)$ for each $j = 1, \ldots, p$.

When $\sigma_1 = \cdots = \sigma_p = \sigma$ and $\langle Z_{\cdot k}, Z_{\cdot k'} \rangle = 0$, then the latent factor model becomes the probabilistic PCA proposed by Tipping and Bishop (1999).

3. The mean-field variational posterior

We consider the mean-field variational class as follows:

$$\begin{aligned} \mathcal{Q}^{\text{MF}} = \Big\{ q(\Theta) : \prod_{j=1}^{p} \prod_{k=1}^{K} \left[\zeta_{jk} \mathcal{N}(\mu_{jk}, v_{jk}^2) + (1 - \zeta_{jk})\delta_0 \right], \mu_j \in \mathbb{R}, \\ v_{jk} \in \mathbb{R}^+, \zeta_{jk} \in [0,1], j = 1, \ldots, p, k = 1, \ldots, K \Big\}, \end{aligned} \tag{5}$$

4. Asymptotic properties of the variational posterior

5. Variational inference

5.1. The evidence lower bound

For $q(\Theta) \in \mathcal{Q}^{MF}$, the variational posterior is obtained by minimizing the Kullback-Leibler (KL) divergence between all $q(\theta) \in \mathcal{Q}^{\text{MF}}$ and the posterior:

$$\hat{q}(\Theta) = \underset{q(\Theta) \in \mathcal{Q}^{\text{MF}}}{\arg\min} \, \text{KL}\left(q(\Theta), \pi(\Theta | X^n)\right), \tag{6}$$

By plugging-in the KL divergence formula, (6) can be written as

$$\begin{aligned}\hat{q}(\Theta) &= \underset{q(\Theta)\in\mathcal{Q}^{\mathrm{MF}}}{\arg\min}\ (\mathbb{E}_q \log q(\Theta) - \mathbb{E}_q \log \pi(\Theta|X^n)) \\ &= \underset{q(\Theta)\in\mathcal{Q}^{\mathrm{MF}}}{\arg\min}\ (\mathbb{E}_q \log q(\Theta) - \mathbb{E}_q \log \pi(\Theta, X^n) + \log \pi(X^n))\,. \end{aligned} \tag{7}$$

Solving $\hat{q}(\Theta)$ in (7) requires knowing the 'evidence' part in the posterior, $\log \pi(X^n)$; however, its explicit expression is typically intractable. In practice, it is custom to drop this term and only focus on the first two terms; i.e., we solve

$$\hat{q}(\Theta) = \underset{q(\Theta)\in\mathcal{Q}^{\mathrm{MF}}}{\arg\max}\ \mathbb{E}_q \log \pi(\Theta, X^n) - \mathbb{E}_q \log q(\Theta). \tag{8}$$

The right hand side expression in (8) is called as the evidence lower bound (ELBO). One can quickly check that it is a lower bound of $\log \pi(X^n)$.

In view of our model and priors, the mean-field variational posterior $\hat{q}(\Theta)$ still cannot be solved directly from (8) as $\pi(\Theta, X^n)$ remains intractable because one needs to solve $\pi(\Theta, X^n) = \int \pi(\Theta, \boldsymbol{w}, X^n) d\,\boldsymbol{w}$ and marginalize out those hyperparameters $\boldsymbol{\gamma}$ and θ in the spike-and-slab IBP prior. Instead of directly solving $\hat{q}(\Theta)$, we will introduce a CAVI (coordinate ascent variational inference) algorithm to solve it in the next section.

5.2. The CAVI algorithm

The CAVI algorithm an iterative method, which sequentially optimizes each of the unknown variable by conditioning on the rest. Our CAVI algorithm consists an expectation step and coordinate updating steps. In the expectation step, we handle latent variables $\boldsymbol{w} = (w_1, \ldots, w_n)$. In those coordinate updating steps, we update $q(Z)$, θ, and σ_j^2. We now introduce each step in detail.

5.2.1. The expectation step

From the latent factor model in (1) and the IBP prior in (4), the posterior distribution $\pi(\Theta, X^n)$ requires solving $\int \pi(\Theta, \boldsymbol{w} \mid X^n) d\,\boldsymbol{w}$, which is intractable. Here, Θ includes all the parameters except for $\boldsymbol{w}$. We introduce an expectation (E-) step similar to that in the EM algorithm. In the E-step, the latent variable $\boldsymbol{w}$ is considered as missing data. We will derive the conditional distribution $\boldsymbol{w}$ given $(\Theta^{(t)}, X^n)$, where $\Theta^{(t)}$ is the value of Θ from the previous (namely, t-th) iteration of the algorithm, and then calculate $\mathbb{E}_{\boldsymbol{w}\mid\Theta^{(t)},X^n} \log \pi(\Theta, \boldsymbol{w} \mid X^n)$, which we denote this quantity $Q(\Theta|\Theta^{(t)})$ for simplicity.

Denote the full model posterior as $\pi(\Theta, \boldsymbol{w}, X^n)$. By calculation, we have

$$\pi(w_i \,|\, \Theta, X^n) = \mathcal{N}(\omega_i, V_w),\ i = 1, \ldots, n \tag{9}$$

where $V_w^{-1} = \mathbb{E}_{Z\sim\hat{q}(Z)}(Z'\Sigma^{-1}Z) + I_K$ and $w_i = V_w(\mathbb{E}_{Z\sim\hat{q}(Z)}Z)'\Sigma^{-1}X_i$, $\hat{q}(Z)$ is the mean-field variational posterior of Z.

In view of the mean-field variational posterior defined in (6), let $Z = (Z_{1\cdot}, \ldots, Z_{p\cdot})'$, each $Z_{j\cdot}$ is a length K vector, then

$$\mathbb{E}_{Z\sim\hat{q}(Z)}(Z'\Sigma^{-1}Z) = \sum_{j=1}^{p}\mathbb{E}_{Z\sim\hat{q}(Z)}(Z_{j\cdot}Z_{j\cdot}')/\sigma_j^2 := \sum_{j=1}^{p}(\zeta_{j\cdot}\zeta_{j\cdot}')\circ\Phi_j/\sigma_j^2, \tag{10}$$

where $\circ$ stands for the coordinate-wise product between two matrices, $\zeta_{j\cdot} = (\zeta_{j1}, \ldots \zeta_{jK})$, $\Phi_j \in \mathbb{R}^{K\times K}$, $\Phi_{j,kk} = u_{jk}^2 + \nu_{jk}^2$ and $\Phi_{j,k'k} = u_{jk}u_{jk'}$ for $k \neq k'$. In ω_i, for each (j,k)-th element of Z, $\mathbb{E}_{Z\sim\hat{q}(Z)}Z_{jk} = \zeta_{jk}u_{jk}$.

5.2.2. Coordinate update equations

To obtain $\hat{q}(Z)$ and $\hat{\theta}_k$, we introduce the stick-breaking representation of the IBP prior. Let $1 \geq \theta_{(1)} > \theta_{(2)} > \cdots > \theta_{(K)} \geq 0$ be a sequence has a decreasing order, then the prior $\theta_k \,|\, \alpha \sim \text{Beta}(\alpha/K, 1)$ admits a stick-breaking representation; i.e.,

$$\theta_{(k)} = \prod_{l=1}^{k} v_l, \quad v_l \overset{i.i.d}{\sim} \text{Beta}(\alpha, 1). \tag{11}$$

The above stick-breaking presentation is used by Ročková and George (2016) for the PXL-EM algorithm for the sparse factor analysis. Recently, Ohn and Kim (2022) studied the sufficient conditions for consistently estimating the dimensional of latent factors of the posterior distribution using this representation under the assumption that Z is jointly row sparse.

From the definition,

$$\hat{q}(\Theta) = \underset{q(\Theta)\in\mathcal{Q}^{MF}}{\arg\max}\ \mathbb{E}_q\left(Q(\Theta \,|\, \Theta^{(t)}) - \log q(\Theta)\right), \tag{12}$$

Simply calculation reveals that

$$\begin{aligned} Q(\Theta \,|\, \Theta^{(t)}) = C - \frac{n}{2}\log\det(\Sigma) - \frac{1}{2}\sum_{i=1}^{n}\left[X_i'\Sigma^{-1}X_i - 2X_i'\Sigma^{-1}Z\omega_i + \text{Tr}(H_iZ'\Sigma^{-1}Z)\right] \\ - \log D_\Pi + \log d\Pi(\Theta), \end{aligned} \tag{13}$$

where $H_i = \omega_i\omega_i' + V_w$ and C is a constant does not depend on the unknown parameters, D_Π is the denominator of the posterior distribution and $\Pi(\Theta)$ stands for the priors for all the parameters in Θ.

The CAVI algorithm solves (12) through iteratively solving $\hat{q}(Z)$, $\hat{\sigma}_j^2$ and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)$.

1) Solving $\hat{q}(Z)$

Assuming $q(Z) \in \mathcal{Q}^{MF}$ in (5) and noticing that the variational posterior distribution of Z_{jk} conditional on $\zeta_{jk} = 1$ is singular to the Dirac measure δ_0 in the Radon-Nikodym derivative, denoted by $dP_{u_{jk},\nu_{jk}\,|\,\zeta_{jk}=1}/d\Pi(Z_{jk})$, $\Pi(Z_{jk})$ the prior distribution, it is sufficient to consider

only the continuous part of the prior measure in the denominator. Therefore, by calculation given in the appendix, we solve each u_{jk} and ν_{jk} by minimizing the equation given by

$$g(u_{jk},\nu_{jk}) = -\sum_{i=1}^{n}\sigma_j^{-2}X_{ij}\zeta_{jk}u_{jk}\omega_{ik} + \frac{1}{2}\sum_{i=1}^{n}\sum_{k'=1}^{K}\sigma_j^{-2}\zeta_{jk}\zeta_{jk'}\Phi_{j,k'k}H_{i,k'k} - \frac{1}{2}\log\nu_{jk}^2 - \lambda_k\mathbb{E}|Z_{jk}|, \tag{14}$$

where $\Phi_{j,k'k}$ is in (10) and the expectation of the last term equals to the mean of the folded norm given by

$$\mathbb{E}|Z_{jk}| = \zeta_{jk}\left(\nu_{jk}\sqrt{2/\pi}\exp(-u_{jk}^2/(2\nu_{jk}^2)) + u_{jk}(1-\Phi_N(-u_{jk}/(\nu_{jk})))\right).$$

The notation $\Phi_N(\cdot)$ in the last display stands for the cumulative distribution function (CDF) of the standard normal density.

Next, from the derivation in the appendix, the mixture weight ζ_{jk} can be obtained from $h_{jk} = \zeta_{jk}/(1-\zeta_{jk})$, where $\hat{h}_{jk}$ is given by

$$\begin{aligned}\hat{h}_{jk} = &\sum_{i=1}^{n}\sigma_j^{-2}X_{ij}u_{jk}\omega_{ik} - \frac{1}{2}\sigma_j^{-2}\sum_{i=1}^{n}\sum_{k'=1}^{K}\zeta_{jk'}\Phi_{j,k'k}H_{i,k'k} + \log\frac{\theta_k}{1-\theta_k} + \frac{1}{2}\\ &- \log\frac{\sqrt{\pi}\lambda_k\nu_{jk}}{\sqrt{2}} - \lambda_k\nu_{jk}^2\sqrt{\frac{2}{\pi}}\exp\left(-\frac{u_{jk}^2}{2\nu_{jk}^2}\right) - \lambda_k\left(1-\Phi_N\left(-\frac{u_{jk}}{\nu_{jk}}\right)\right).\end{aligned} \tag{15}$$

2) Solving θ_k and σ_j^2

Let $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_K)$, by maximizing the log-posterior, we obtain $\hat{\theta}_k = g(\theta_k)$., where

$$g(\boldsymbol{\theta}) = \sum_{j=1}^{p}\sum_{k=1}^{K}[\zeta_{jk}\log\theta_k + (1-\zeta_{jk})\log(1-\theta_k) + (\alpha-1)\log\theta_K] \tag{16}$$

subject to $1 \geq \theta_{(k)} > \theta_{(k-1)} > \cdots > \theta_{(1)} \geq 0$. One can solve $\boldsymbol{\theta}$ by solving a linear programming problem with the constraints: $0 \leq \theta_{(k)} \leq 1$ for all k and $\theta_{(k)} - \theta_{(k-1)} \leq 0$ for $k \geq 2$. As noted by Ročková and George (2016), this step can be further enhanced by performing a permutation of the column based on the corresponding $\theta_{(k)}$ value to solve the constrained problem. This order constraint intrinsically leads to shrinkage of higher-index of factor loadings.

Last, we solve σ_j^2 from

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n}\left(X_{ij}^2 - 2\sum_{k=1}^{K}X_{ij}\zeta_{jk}u_{jk}w_{ik} + \sum_{k=1}^{K}\sum_{k'=1}^{K}\zeta_{jk}\zeta_{jk'}\Phi_{j,kk'}H_{i,kk'}\right) + 2b}{n+2a+2}. \tag{17}$$

The derivation of the last display can be found in the appendix.

5.3. Parameter expansion and the PXL-CAVI algorithm

We apply parameter expansion to increase the convergence speed of the original CAVI algorithm. The parameter expansion strategy was first introduced by Liu et al. (1998) and recently adopted by Ročková and George (2016) in the PXL-EM algorithm for a Bayesian sparse factor analysis. Introducing a positive definite matrix D and define $\tilde{Z} = ZD_L^{-1}$, D_L is the lower triangular part from the Cholesky decomposition, then the likelihood function can be re-written as

$$X_i = \tilde{Z}\tilde{w}_i + \epsilon_i,$$

where $\tilde{w}_i \sim \mathcal{N}(0, D)$.

We put the SS-IBP prior on $\tilde{Z}$ instead of Z. Unless D is a diagonal matrix, the support of Z is not the same as the support of $\tilde{Z}$. However, the number of non-zero factor loadings of Z and $\tilde{Z}$ will be the same. The next steps are similar to the original CAVI algorithm. We replace u_{jk} and ν_{jk} with (14)-(17) respectively.

5.4. Joint row-sparsity on Z

The joint row-sparsity is a special case that all the coordinates within each row are either all zero or all non-zero. Under this assumption, the support of Z is simply $S = S_1 = \cdots = S_k$.

Appendix A: Proofs of (14)-(17)

Proof. We first prove (14). Let $\boldsymbol{u}, \boldsymbol{\nu} \in \mathbb{R}^{p\times K}$, where the (j,k)-th coordinate in $\boldsymbol{u}$ and $\boldsymbol{\nu}$ are u_{jk} and ν_{jk} respectively. Denote $P_{-jk} = P_{-\boldsymbol{u}_{-jk},\boldsymbol{\nu}_{-jk},\boldsymbol{\gamma}_{-jk}}$ be the posterior of $\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{\gamma}$ without the (j,k)-th entries. A similar notation is used for the joint prior Π_{-jk} of $\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{\gamma}$. We have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{u},\boldsymbol{\nu},\boldsymbol{\gamma}\,|\,\zeta_{jk}=1}\left[-Q(\Theta\,|\,\Theta^{(t)}) + \log D_\Pi + \log\frac{dP_{-jk}\otimes N(u_{jk},\nu_{jk}^2)}{d\Pi_{-jk}\otimes\theta_k\mathrm{Lap}(\lambda_k)}\right] + C \\
&= -\sum_{i=1}^{n}\mathbb{E}_{\cdot\,|\,\zeta_{jk}=1}(X_i'\Sigma^{-1}Z\omega_i) + \mathrm{Tr}(H_i\mathbb{E}_{\cdot\,|\,\zeta_{jk}=1}(Z'\Sigma^{-1}Z)) - \frac{1}{2}\log(2\pi\nu_{jk}^2) \qquad (18)\\
&\quad - \mathbb{E}_{\cdot\,|\,\zeta_{jk}=1}\left(\frac{(Z_{jk}-u_{jk})^2}{2\nu_{jk}^2}\right) - \log(\theta_k) - \log(\lambda_k/2) - \lambda_k\mathbb{E}_{\cdot\,|\,\zeta_{jk}=1}|Z_{jk}| + C', \qquad (19)
\end{aligned}$$

where we denote $\mathbb{E}_{\cdot\,|\,\zeta_{jk}=1}$ as $\mathbb{E}_{\boldsymbol{u},\boldsymbol{\nu},\boldsymbol{\gamma}\,|\,\zeta_{jk}=1}$ for convenience. From the variational posterior defined in (5), $Z_{jk}|\zeta_{jk}=1 \sim \mathcal{N}(u_{jk},\nu_{jk}^2)$ and thus,

$$\mathbb{E}_{\cdot\,|\,\zeta_{jk}=1}\left(\frac{(Z_{jk}-u_{jk})^2}{2\nu_{jk}^2}\right) = \frac{1}{2}, \qquad (20)$$

as $((Z_{jk}-u_{jk})/\nu_{jk})^2$ follows the standard χ^2 distribution. For the first two terms in (18), we write $X_i\Sigma^{-1}Z\omega_i = \sum_{j=1}^{p}\sum_{k=1}^{K}\sigma_j^{-2}X_{ij}Z_{jk}\omega_{ik}$. Therefore,

$$\mathbb{E}_{\cdot\,|\,\zeta_{jk}=1}(X_i\Sigma^{-1}Z\omega_i) = \sum_{j=1}^{p}\sum_{k=1}^{K}\sigma_j^{-2}X_{ij}u_{jk}\omega_{ik}. \qquad (21)$$

We can also write

$$\mathrm{Tr}(H_i Z'\Sigma^{-1}Z) = \sum_{j=1}^{p}\sigma_j^{-2}\,\mathrm{Tr}(H_i Z_{j\cdot}Z'_{j\cdot}) = \sum_{j=1}^{p}\sum_{k=1}^{K}\sum_{k'=1}^{K}\sigma_j^{-2}H_{i,kk'}Z_{jk}Z_{jk'}.$$

Taking expectation on both side, we obtain

$$\mathrm{Tr}(H_i\mathbb{E}_{\cdot\,|\,\zeta_{jk}=1}(Z'\Sigma^{-1}Z)) = \sum_{j=1}^{p}\sum_{k=1}^{K}\sum_{k'=1}^{K}\sigma_j^{-2}\zeta_{jk}\zeta_{jk'}H_{i,kk'}\Phi_{j,kk'}. \tag{22}$$

Since the goal is to solve u_{jk} and ν_{jk}, other terms in the equation which do not depend on the two terms can be simply treated as constants. Therefore, by plugging (20)–(22) into (18) and (19), we obtain the expression in (14).

Next, we prove (15). We have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{u},\boldsymbol{\nu},\boldsymbol{\gamma}}\left[-Q(\Theta\,|\,\Theta^{(t)}) + \log D_\Pi + \log\frac{d(\zeta_{jk}N(u_{jk},\nu_{jk}^2)+(1-\zeta_{jk})\delta_0)}{d(\theta_k\mathrm{Lap}(\lambda_k)+(1-\theta_k)\delta_0)}\right] + C\\
&\quad= \mathbb{E}_{\boldsymbol{u},\boldsymbol{\nu},\boldsymbol{\gamma}}\Big[-\sum_{i=1}^{n}X_i'\Sigma^{-1}Zw_i + \frac{1}{2}\mathrm{Tr}(H_iZ'\Sigma^{-1}Z) + \mathbb{1}_{\{\gamma_{jk}=1\}}\left(\log\frac{\zeta_{jk}dN(u_{jk},\nu_{jk}^2)}{\theta_k d\mathrm{Lap}(\lambda_k)}\right)\\
&\qquad+\mathbb{1}_{\{\gamma_{jk}=0\}}\log\frac{1-\zeta_{jk}}{1-\theta_k}\Big]
\end{aligned}$$

Note that $\mathbb{E}_{\boldsymbol{u},\boldsymbol{\nu},\boldsymbol{\gamma}}(\mathbb{1}_{\{\gamma_{jk}=1\}}) = \zeta_{jk}$. Using (21) and (22), the last display can be written as

$$\begin{aligned}
&\zeta_{jk}\Big\{-\sum_{i=1}^{n}\sigma_j^{-2}X_{ij}u_{jk}w_{jk} + \frac{1}{2\sigma_j^2}\sum_{i=1}^{n}\sum_{k'=1}^{K}\zeta_{jk'}H_{i,kk'}\Phi_{j,kk'} + \log\frac{\zeta_{jk}}{\theta_k} - \frac{1}{2}\log(2\pi\nu_{jk}^2) - \frac{1}{2}\\
&\quad-\log\frac{\lambda_k}{2} + \lambda_k\nu_{jk}\sqrt{2/\pi}\exp(-u_{jk}^2/(2\nu_{jk}^2)) + \lambda_k u_{jk}(1-\Phi_N(-u_{jk}/\nu_{jk}))\Big\}\\
&\quad+(1-\zeta_{jk})\log\frac{1-\zeta_{jk}}{1-\theta_k}.
\end{aligned}$$

Taking derivation with respect to ζ_{jk} and set the equation to 0, we thus obtain (15).

Last, we derive (16) and (17). (16) is obtained the same as in Ročková and George (2016). We first express the objection function as function of v_k,

$$g(v) = \sum_{j=1}^{p}\sum_{k=1}^{K}\left\{\zeta_{jk}\sum_{l=1}^{k}\log v_l + (1-\zeta_{jk})\log(1-\prod_{l=1}^{k}v_l)\right\} + (\alpha-1)\sum_{k=1}^{K}\log(v_k).$$

Using the stick-breaking law, we have $v_k = \theta_{(k)}/\theta_{(k-1)}$. By plugging-in this expression, we obtain (16).

To derive (17), we solve σ_j^2 by minimizing the function given by

$$\frac{n}{2}\log\sigma_j^2 - \mathbb{E}_{\boldsymbol{u},\boldsymbol{\nu},\boldsymbol{\gamma}}Q(\Theta\,|\,\Theta^{(t)}) - \log\pi(\sigma^2)$$

$$= C + \frac{n}{2}\log\sigma_j^2 + \frac{1}{2\sigma_j^2}\sum_{i=1}^{n} X_{ij}^2 - \frac{1}{\sigma_j^2}\sum_{i=1}^{n}\sum_{k=1}^{K} X_{ij}\zeta_{jk}u_{jk}w_{ik}$$
$$- \frac{1}{2\sigma_j^2}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{k'=1}^{K}\zeta_{jk}\zeta_{jk'}\Phi_{j,kk'}H_{i,kk'} + (a+1)\log\sigma_j^2 + \frac{b}{\sigma_j^2}.$$

Taking derivative with respect to σ_j^2, we obtain (17).

□

References

Knowles, D. and Z. Ghahramani (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *Annals of Applied Statistics 5*, 1534–1552.

Liu, C., D. B. Rubin, and Y. N. Wu (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika 85*(4), 755–770.

Ohn, I. and Y. Kim (2022). Posterior consistency of factor dimensionality in high-dimensional sparse factor models. *Bayesian Analysis 2*(17), 491–514.

Ročková, V. and E. I. George (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association 111*, 1608–1622.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B 61*(3), 611–622.