Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity

by

Veronika Ročková and Edward I. George [1]

The University of Pennsylvania

August 30, 2014

Abstract

Rotational transformations have traditionally played a key role in enhancing the interpretability of factor analysis via post-hoc modifications of the model orientation. Here, we propose a unified Bayesian approach that incorporates factor rotations within the model fitting process, greatly enhancing the effectiveness of sparsity inducing priors. These automatic transformations are embedded within a new PXL-EM algorithm, a Bayesian variant of parameter-expanded EM for fast posterior mode detection. By iterating between soft-thresholding of small factor loadings and transformations of the factor basis, we obtain dramatic accelerations yielding convergence towards better oriented sparse solutions. For accurate recovery and estimation of factor loadings, we propose a spike-and-slab LASSO prior, a two-component refinement of the Laplace prior. Our approach is automatic, because it does not require any pre-specification of the factor dimension. This assumption is avoided by introducing infinitely many factors with the Indian Buffet Process (IBP) prior. The specification of identifiability constraints is also completely avoided. The PXL-EM, made available by the stick-breaking IBP representation, capitalizes on the very fast LASSO implementations and converges quickly. Dynamic posterior exploration over a sequence of spike-and-slab priors is seen to facilitate the search for a global posterior mode. For mode selection, we propose a criterion based on an integral lower bound to the marginal likelihood. The potential of the proposed procedure is demonstrated on both simulated and real high-dimensional data, which would render posterior simulation impractical.

Keywords: Indian Buffet Process; Infinite-dimensional Factor Analysis; Factor Rotation; PXL-EM; Spike-and-slab LASSO.

[1]Veronika Ročková is a postdoctoral researcher at the Department of Statistics, University of Pennsylvania, Philadelphia, PA, 19104, vrockova@wharton.upenn.edu. Edward I. George is Professor of Statistics, Department of Statistics, University of Pennsylvania, Philadelphia, PA, 19104, edgeorge@wharton.upenn.edu.

1 Bayesian Factor Analysis Revisited

Latent factor models aim to find regularities in the variation among multiple responses, and relate these to a set of hidden causes. This is typically done within a regression framework through a linear superposition of unobserved factors. The traditional setup for factor analysis consists of an $n \times G$ matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]'$ of n independent G-dimensional vector observations. For a fixed factor dimension K, the generic factor model is of the form

$$f(\boldsymbol{y}_i \mid \boldsymbol{\omega}_i, \boldsymbol{B}, \boldsymbol{\Sigma}) \overset{\text{ind}}{\sim} \mathcal{N}_G(\boldsymbol{B}\boldsymbol{\omega}_i, \boldsymbol{\Sigma}), \quad \boldsymbol{\omega}_i \sim \mathcal{N}_K(\boldsymbol{0}, \mathrm{I}_K), \tag{1.1}$$

for $1 \leq i \leq n$, where $\boldsymbol{\Sigma} = \mathsf{diag}\{\sigma_j^2\}_{j=1}^G$ is a diagonal matrix of unknown positive scalars, $\boldsymbol{\omega}_i \in \mathbb{R}^K$ is the i^{th} realization of the unobserved latent factors, and $\boldsymbol{B} \in \mathbb{R}^{G \times K}$ is the matrix of factor loadings that weight the contributions of the individual factors. Marginally, $f(\boldsymbol{y}_i \mid \boldsymbol{B}, \boldsymbol{\Sigma}) = \mathcal{N}_G(\boldsymbol{0}, \boldsymbol{B}\boldsymbol{B}' + \boldsymbol{\Sigma})$, $1 \leq i \leq n$, a decomposition which uses at most $G \times K$ parameters instead of $G(G+1)/2$ parameters in the unconstrained covariance matrix. Note that we have omitted an intercept term, assuming throughout that the responses have been centered.

Fundamentally a multivariate regression with unobserved regressors, factor analysis is made more more challenging by the uncertainty surrounding the number and orientation of the regressors. A persistent difficulty associated with the factor model (1.1) has been that $\boldsymbol{B}$ is unidentified. In particular, any orthogonal transformation of the loading matrix and latent factors $\boldsymbol{B}\boldsymbol{\omega}_i = (\boldsymbol{B}\boldsymbol{P})(\boldsymbol{P}'\boldsymbol{\omega}_i)$ yields exactly the same distribution for $\boldsymbol{Y}$. Although identifiability is not needed for prediction or estimation of the marginal covariance matrix, non-sparse orientations diminish the potential for interpretability, our principal focus here.

The main thrust of this paper is the development of a Bayesian approach for factor analysis that can automatically identify interpretable factor orientations with a fast deterministic implementation. Our approach does not make use of any of the usual identifiability constraints on the allocation of the zero elements of $\boldsymbol{B}$, such as lower-triangular forms (Lopes and West, 2004) and their generalizations (Frühwirth-Schnatter and Lopes, 2009). Nor do we require prespecification of the factor cardinality K. Instead, the loading matrix $\boldsymbol{B}$ is extended to include infinitely many columns, where the number of "effective" factors remains finite with probability one.

Our approach begins with a prior on the individual elements in $\boldsymbol{B} = \{\beta_{jk}\}_{j,k=1}^{G,\infty}$ that induces posterior zeroes with high-probability. Traditionally, this entails some variant of a spike-and-slab prior that naturally segregates important coefficients from coefficients that are ignorable (George and McCulloch, 1993; West, 2003; Carvalho et al., 2008; Rai and Daumé, 2008; Knowles and Ghahramani, 2011). The specification of such priors is facilitated by the introduction of a latent binary allocation matrix $\boldsymbol{\Gamma} = \{\gamma_{jk}\}_{j,k=1}^{G,\infty}$, $\gamma_{jk} \in \{0, 1\}$, where $\gamma_{jk} = 1$ whenever the j^{th} variable is associated with k^{th}

factor. Given each $\gamma_{jk} \in \{0, 1\}$ for whether β_{jk} should be ignored, a particularly appealing spike-and-slab variant has been

$$\pi(\beta_{jk} \mid \gamma_{jk}, \lambda_1) = (1 - \gamma_{jk})\delta_0(\beta_{jk}) + \gamma_{jk}\phi(\beta_{jk} \mid \lambda_1), \tag{1.2}$$

where $\delta_0(\cdot)$ is the "spike distribution" (atom at zero) and $\phi(\cdot \mid \lambda_1)$ is the absolutely continuous "slab distribution" with exponential tails or heavier, indexed by a hyper-parameter λ_1. Coupled with a suitable beta-Bernoulli prior on the binary indicators γ_{jk}, the point-mass generative model (1.2) has been shown to yield optimal rates of posterior concentration, both in linear regression (Castillo and van der Vaart, 2012; Castillo et al., 2014) and covariance matrix estimation (Pati et al., 2014). This "methodological ideal", despite being amenable to posterior simulation, poses serious computational challenges in high-dimensional data. These challenges are even more pronounced with the infinite factor models considered here, where a major difficulty is finding highly probable zero allocation patterns within infinite loading matrices.

We will address this challenge by developing a tractable inferential procedure that does not rely on posterior simulation and thereby is ideally suited for high-dimensional data. At the heart of our approach is a novel spike-and-slab LASSO (SSL) prior, a feasible continuous relaxation of its limiting special case (1.2). Such relaxations transform the obstinate combinatorial problem into one of optimization in continuous systems, permitting the use of EM algorithms (Dempster et al., 1977), a strategy we pursue here. The SSL prior will be coupled with the Indian Buffet Process (IBP) prior, which defines a prior distribution on the allocation patterns of active elements within the infinite loading matrix. Our EM algorithm capitalizes on the stick-breaking representation of the IBP, which induces increasing shrinkage of higher-ordered factor loadings.

The indeterminacy of (1.1) due to rotational invariance is ameliorated with the SSL prior, which anchors on sparse representations. This prior automatically promotes rotations with many zero loadings by creating ridge-lines of posterior probability along coordinate axes, thereby radically reducing the posterior multimodality. Whereas empirical averages resulting from posterior simulation may aggregate probability mass from multiple modes and are non-sparse, our EM algorithm yields sparse modal estimates with exact zeroes in the loading matrix.

The search for promising sparse factor orientations is greatly enhanced with data augmentation by expanding the likelihood with an auxiliary transformation matrix (Liu et al., 1998). Exploiting the rotational invariance of the factor model, we propose a PXL-EM (parameter expanded likelihood EM) algorithm that automatically rotates the loading matrix as a part of the estimation process, gearing the EM trajectory along the orbits of equal likelihood. The SSL prior then guides the selection from the many possible likelihood maximizers. Moreover, the PXL-EM algorithm is far more robust against poor initializations, converging dramatically faster than the parent EM algorithm and inducing

orthogonal latent factor featurizations.

As with many other sparse factor analyzers (Knowles and Ghahramani, 2011; Frühwirth-Schnatter and Lopes, 2009; Carvalho et al., 2008), the number of factors will be inferred through the estimated patterns of sparsity. In over-parametrized models with many redundant factors, this can be hampered by the phenomenon of factor splitting, i.e. the smearing of factor loadings across multiple correlated factors. Such factor splitting is dramatically reduced in our approach, because the IBP construction prioritizes lower indexed loadings and the PXL-EM rotates towards independent factors.

To facilitate the search for a global maximum we implement dynamic posterior exploration, a sequential reinitialization along a ladder of increasing spike penalties. Posterior modes will be evaluated by a criterion motivated as an integral lower bound to a posterior probability of the implied sparsity pattern.

The paper is structured as follows. Section 2 introduces our hierarchical prior formulation for infinite factor loading matrices, establishing some of its appealing theoretical properties. Section 3 develops the construction of our basic EM algorithm, which serves as a basis for the PXL-EM algorithm presented in Section 4. Section 5 describes the dynamic posterior exploration strategy for PXL-EM deployment, demonstrating its effectiveness on simulated data. Section 6 derives and illustrates our criterion for factor model comparison. Sections 7 and 8 present applications of our approach on real data. Section 9 concludes with a discussion.

2 Infinite Factor Model with Spike-and-Slab LASSO

The cornerstone of our Bayesian approach is a hierarchically structured prior on infinite-dimensional loading matrices, based on the spike-and-slab LASSO prior of Rockova and George (2014b). For each loading β_{jk}, we consider a continuous relaxation of (1.2) with two Laplace components: a slab component with a common penalty λ_1, and a spike component with a penalty λ_{0k} that is potentially unique to the k^{th} factor. More formally,

$$\pi(\beta_{jk} \mid \gamma_{jk}, \lambda_{0k}, \lambda_1) = (1-\gamma_{jk})\phi(\beta_{jk} \mid \lambda_{0k}) + \gamma_{jk}\phi(\beta_{jk} \mid \lambda_1), \tag{2.1}$$

where $\phi(\beta \mid \lambda) = \frac{\lambda}{2}\exp\{-\lambda|\beta|\}$ is a Laplace prior with mean 0 and variance $2/\lambda^2$ and $\lambda_{0k} >> \lambda_1 > 0$, $k = 1, \ldots \infty$. The prior (2.1) will be further denoted as $SSL(\lambda_{0k}, \lambda_1)$. Coupled with a prior on γ_{jk}, this mixture prior induces a variant of "selective shrinkage" (Ishwaran and Rao, 2005) that adaptively segregates the active coefficients from the ignorable via differential soft-thresholding. The coefficients β_{jk} with active selection indicators ($\gamma_{jk} = 1$) are left relatively unaffected by keeping the slab penalty λ_1 small. The coefficients β_{jk} that are ignorable ($\gamma_{jk} = 0$) are pulled towards zero by letting the

spike penalty λ_{0k} be substantially larger than λ_1. The point-mass prior (1.2) is obtained as a limiting special case of (2.1) when $\lambda_{0k} \to \infty$.

Despite its continuity at the origin, the $SSL(\lambda_{0k}, \lambda_1)$ mixture prior thresholds smaller β_{jk} to exact zeroes, aligning posterior modes along the coordinate axes. This is in sharp contrast to existing spike-and-slab priors with continuous Gaussian spike distributions (George and McCulloch, 1993; Rockova and George, 2014a), whose non-sparse posterior modes must be thresholded for variable selection. The exact sparsity of the posterior modes here will be crucial for anchoring interpretable factor orientations.

For the diagonal elements of $\mathbf{\Sigma}$, we assume independent inverse gamma priors

$$\sigma_1^2, \ldots, \sigma_G^2 \overset{\text{iid}}{\sim} \text{IG}(\eta/2, \eta\xi/2) \tag{2.2}$$

with the relatively noninfluential choice $\eta = 1$ and $\xi = 1$.

The prior construction is completed with the specification of a prior distribution over the feature allocation matrix $\mathbf{\Gamma} = \{\gamma_{jk}\}_{j,k=1}^{G,\infty}$. As with similar sparse infinite factor analyzers (Knowles and Ghahramani, 2011; Rai and Daumé, 2008), we consider the Indian Buffet Process (IBP) prior of Griffiths and Ghahramani (2005), which defines an exchangeable distribution over equivalence classes $[\mathbf{\Gamma}]$ of infinite-dimensional binary matrices. Formally, the IBP with intensity parameter $\alpha > 0$ arises from the beta-Bernoulli prior

$$\begin{aligned} \pi(\gamma_{jk}|\theta_k) &\overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_k), \\ \pi(\theta_k) &\overset{\text{ind}}{\sim} \mathcal{B}\left(\frac{\alpha}{K}, 1\right), \end{aligned} \tag{2.3}$$

by integrating out the θ_k's and by taking the limit $K \to \infty$ (Griffiths and Ghahramani, 2011). Each equivalence class $[\mathbf{\Gamma}]$ contains all matrices $\mathbf{\Gamma}$ with the same left-ordered form, obtained by ordering the columns from left to right by their binary numbers. Marginally,

$$\pi([\mathbf{\Gamma}]) = \frac{\alpha^{K^+}}{\prod_{h=1}^{2^G-1} K_h!} \exp\left(-\alpha H_G\right) \prod_{k=1}^{K^+} \frac{(G - |\boldsymbol{\gamma}_k|)!\,(|\boldsymbol{\gamma}_k| - 1)!}{G\,!}, \tag{2.4}$$

where $|\boldsymbol{\gamma}_k| = \sum_{j=1}^{G} \gamma_{jk}$, $H_N = \sum_{j=1}^{N} \frac{1}{j}$ is the N-th harmonic number, K^+ is the number of active factors, i.e. $K^+ = \sum_k \mathbb{I}(|\boldsymbol{\gamma}_k| > 0)$, and K_h is the number of columns $\boldsymbol{\gamma}_k$ expressing the same binary number h. Proceeding marginally over θ_k lends itself naturally to a Gibbs sampler (Knowles and Ghahramani, 2011; Rai and Daumé, 2008). However, to obtain an EM algorithm, we will instead proceed conditionally on a particular ordering of the θ_k's. We will capitalize on the following stick-breaking representation of the IBP (Teh et al., 2007).

Theorem 2.1. *(Teh et al., 2007)) Let $\theta_{(1)} > \theta_{(2)} > \ldots > \theta_{(K)}$ be a decreasing ordering of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$, where each $\theta_k \overset{\text{iid}}{\sim} \mathcal{B}\left(\frac{\alpha}{K}, 1\right)$. In the limit as $K \to \infty$, the $\theta_{(k)}$'s obey the following stick-*

breaking law

$$\theta_{(k)} = \prod_{l=1}^{k} \nu_l, \quad \textit{where} \quad \nu_l \overset{\text{iid}}{\sim} \mathcal{B}(\alpha, 1). \tag{2.5}$$

Remark 2.1. *Unlike the stick-breaking construction for the Dirichlet process prior, here we recurse on the length of the remaining piece of the stick rather than the discarded piece.*

Remark 2.2. *The implicit ordering $\theta_{(1)} > \theta_{(2)} > ... > \theta_{(K)}$ induces a soft identifiability constraint against the permutational invariance of the factor model.*

To sum up, our hierarchical prior $\Pi_{\lambda_{0k}}(\boldsymbol{B})$ on infinite factor loading matrices $\boldsymbol{B} \in \mathbb{R}^{G \times \infty}$ is:

$$\pi(\beta_{jk} \,|\, \gamma_{jk}) \sim SSL(\lambda_{0k}, \lambda_1), \quad \gamma_{jk} \,|\, \theta_{(k)} \sim \text{Bernoulli}[\theta_{(k)}], \quad \theta_{(k)} = \prod_{l=1}^{k} \nu_l, \quad \nu_l \overset{\text{iid}}{\sim} \mathcal{B}(\alpha, 1). \tag{2.6}$$

The representation (2.5) naturally provides a truncated stick-breaking approximation to the IBP under which $\theta_{(k)} = 0$ for all $k > K^\star$. By choosing $K^\star$ suitably large, and also assuming $\beta_{jk} = 0$ for all $k > K^\star$, this approximation will play a key role in the implementation of our EM algorithm. The next section summarizes some of the properties of $\Pi_{\lambda_{0k}}(\boldsymbol{B})$, which justify its use as an implicit prior on the marginal covariance matrix and also establish a suitable lower bound on $K^\star$.

2.1 Properties of the Prior

When dealing with infinite-dimensional loading matrices, the implied marginal covariance matrix $\boldsymbol{\Lambda} = \boldsymbol{B}\boldsymbol{B}' + \boldsymbol{\Sigma}$ needs to have all entries finite with probability one (Bhattacharya and Dunson, 2011). This property would be guaranteed under the point-mass prior $\Pi_\infty(\boldsymbol{B})$, where the underlying IBP process places zero probability on allocation matrices with bounded effective dimension K^+ (Griffiths and Ghahramani, 2011). In the next theorem, we show that this property continues to hold for the continuous relaxation $\Pi_{\lambda_{0k}}(\boldsymbol{B})$ with $\lambda_{0k} < \infty$.

Theorem 2.2. *Let $\Pi_{\lambda_{0k}}(\boldsymbol{B})$ denote the prior distribution* (2.6) *on a loading matrix $\boldsymbol{B}$ with infinitely many columns. Assuming $k/\lambda_{0k} = \mathcal{O}(1)$, we have*

$$\left(\max_{1 \le j \le G} \sum_{k=1}^{\infty} \beta_{jk}^2 \right) < \infty \quad \Pi_{\lambda_{0k}}\textit{-almost surely.} \tag{2.7}$$

Proof. Appendix (Section 10.1).

Remark 2.3. *The statement* (2.7) *is equivalent to claiming that entries in $\boldsymbol{B}\boldsymbol{B}'$ are finite, $\Pi_{\lambda_{0k}}$-almost surely.*

In our EM implementation, we will be relying on the truncated approximation to $\boldsymbol{B}$, setting all loadings β_{jk} indexed by $k > K^\star$ to zero. The truncated loading matrix $\boldsymbol{B}^{K^\star}$ yields a marginal covariance matrix $\boldsymbol{\Lambda}_{K^\star} = \boldsymbol{B}^{K^\star}\boldsymbol{B}^{K^{\star\prime}} + \boldsymbol{\Sigma}$, which can be made arbitrarily close to $\boldsymbol{\Lambda}$ by considering $K^\star$ large enough. The closeness of such approximation will be measured in the sup-norm metric $d_\infty(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{K^\star}) = \max_{1\leq j,m\leq G} |\Lambda_{jk} - \Lambda^\star_{mk}|$. In the next theorem we derive a lower bound on the order of the approximation $K^\star$ so that $d_\infty(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{K^\star}) < \varepsilon$ with large probability. There we will assume that the spike penalties increase exponentially, i.e. $\lambda_{0k}^2 = (1/a)^k$, where $0 < a < 1$. This assumption can be relaxed, as noted at the end of this section.

Theorem 2.3. *Assume $\lambda_{0k}^2 = (1/a)^k$ with $0 < a < 1$. Then for any $\varepsilon > 0$, we have*

$$\mathsf{P}(d_\infty(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{K^\star}) \leq \varepsilon) > 1 - \varepsilon,$$

whenever $K^\star > \max\left\{\log\left[\frac{\widetilde{\varepsilon}}{2}(1-a)\right]/\log(a); \log\left[\frac{\lambda_1^2\widetilde{\varepsilon}}{2}(1-\mu)\right]/\log(\mu)\right\}$, where $\mu = \left(\frac{\alpha}{1+\alpha}\right)$ and $\widetilde{\varepsilon} = \varepsilon[1-(1-\varepsilon)^{1/G}]$.

Proof. Appendix (Section 10.2).

Remark 2.4. *Theorem 2.3 shows that higher $K^\star$ values are needed when the reciprocal penalty decay rate a and/or the mean breaking fraction $\mu = \frac{\alpha}{1+\alpha}$ are closer to one.*

Remark 2.5. *As an aside, it also follows from the proof of Theorem 2.3 (equation* (10.1)*) that the probability $\mathsf{P}(d_\infty(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}^{K^\star}) \leq \varepsilon)$ converges to one exponentially fast with $K^\star$.*

Theorem 2.3 can be further extended to show that the prior $\Pi_{\lambda_{0k}}(\boldsymbol{B})$ places positive probability in an arbitrarily small neighborhood around any covariance matrix, and hence the posterior distribution of $\boldsymbol{\Omega}$ is weakly consistent. According to Bhattacharya and Dunson (2011) (proof of Proposition 2 and Theorem 2), it suffices to show that the prior concentrates enough probability in a Frobenius-norm neighborhood around any true loading matrix $\boldsymbol{B}_0$. This result is summarized in the next theorem.

Theorem 2.4. *Let $\boldsymbol{B}_0$ denote any $(G \times \infty)$ loading matrix with $K^+ < \infty$ nonzero columns and let $B_\varepsilon^F(\boldsymbol{B}_0)$ be an ε-neighborhood of $\boldsymbol{B}_0$ under the Frobenious norm. Assume $\lambda_{0k}^2 = (1/a)^k$ with $0 < a < 1$. Then $\Pi_{\lambda_{0k}}\left[B_\varepsilon^F(\boldsymbol{B}_0)\right] > 0$ for any $\varepsilon > 0$.*

Proof. Follows directly from Theorem 2.3 and the proof of Proposition 2 of Bhattacharya and Dunson (2011). □

Pati et al. (2014) studied prior distributions that yield the optimal rate of concentration of the posterior measures around any true covariance matrix in operator norm, when the dimension $G = G_n$ can be much larger than the sample size n. The authors considered a variant of the point-mass mixture prior $\Pi_\infty(\boldsymbol{B})$ and showed that it leads to consistent covariance estimation in high-dimensional settings,

where the posterior concentrates at a rate that is minimax up to a factor of $\sqrt{\log n}$ (Theorem 5.1 of Pati et al. (2014)). This result was shown for a slightly different hierarchical model for the binary allocations. First, the effective factor dimension K^+ is picked from a distribution π_{K^+}, which decays exponentially, i.e. $\pi_{K^+}(K^+ > k) \leq \exp(-C\,k)$ for some $C > 0$. Second, each binary indicator is sampled from a Bernoulli prior with a common inclusion probability θ arising from a beta prior $\mathcal{B}(1, K_{0n}G_n + 1)$ with expectation $1/(2 + K_{0n}G_n)$, where K_{0n} is the true factor dimension.

The IBP prior instead uses factor-specific inclusion probabilities $\theta_{(1)} > \theta_{(2)} > \ldots$ obtained as a limit of ordered sequences of K independent beta random variables with expectation $1/(1 + K/\alpha)$, when $K \to \infty$. The stick-breaking law then yields $\mathsf{E}[\theta_{(k)}] = \left(\frac{1}{1+1/\alpha}\right)^k$. The parallel between the IBP and the prior of Pati et al. (2014) is instructive for choosing a suitable intensity parameter α. Selecting $\alpha \propto 1/G$ leads to an IBP formulation that is more closely related to a prior yielding the desired theoretical properties. Moreover, in the next theorem we show that for such α, the IBP process induces an implicit prior distribution on K^+, which also decays exponentially.

Theorem 2.5. *Let* $[\boldsymbol{\Gamma}]$ *be distributed according to the IBP prior* (2.4) *with an intensity* $0 < \alpha \leq 1$. *Let* K^+ *denote the effective factor dimension, that is the largest index* $K^+ \in \mathbb{N}$, *so that* $\gamma_{jk} = 0$ *for all* $k > K^+, j = 1, \ldots, G$. *Then*

$$\mathsf{P}(K^+ > k) < 2\left[G(\alpha+1) + \frac{4}{3}\right]\exp\left[-(k+1)\log\left(\frac{\alpha+1}{\alpha}\right)\right]$$

Proof. Appendix (Section 10.3).

The properties discussed in Theorems 2.2, 2.3 and 2.4 were studied under the assumption of increasing shrinkage for higher-indexed factors, i.e. $\lambda_{0k}^2 = (1/a)^k$, $0 < a < 1$. The discussed guarantees will continue to hold also assuming $\lambda_{01} = \cdots = \lambda_{0K^\star}$ and $\lambda_{0k} = \infty$ when $k > K^\star$ for some suitable $K^\star$, i.e. $K^\star \geq K^+$. This more practical variant of our prior $\Pi_{\lambda_{0k}}(\boldsymbol{B})$ will be used throughout our simulated examples and for the data analysis.

3 The EM Approach to Bayesian Factor Analysis

We will leverage the resemblance between factor analysis and multivariate regression, and implement a sparse variant of the EM algorithm for probabilistic principal components (Tipping and Bishop, 1999). We will capitalize on EMVS, a fast method for posterior model mode detection in linear regression under spike-and-slab priors (Rockova and George, 2014a). To simplify notation, throughout the section we will denote the truncated approximation $\boldsymbol{B}^{K^\star}$ by $\boldsymbol{B}$, for some pre-specified $K^\star$. Similarly, $\boldsymbol{\theta}$ will be the finite vector of ordered inclusion probabilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{K^\star})'$ and $\lambda_{0k} = \lambda_0$ for $k = 1, \ldots, K^\star$.

Letting $\boldsymbol{\Delta} = (\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\theta})$, the goal of the proposed algorithm will be to find parameter values $\widehat{\boldsymbol{\Delta}}$ which are most likely (a posteriori) to have generated the data, i.e. $\widehat{\boldsymbol{\Delta}} = \arg\max_{\boldsymbol{\Delta}} \log \pi(\boldsymbol{\Delta} \mid \boldsymbol{Y})$.

This task would be trivial if we knew the hidden factors $\boldsymbol{\Omega} = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n]'$ and the latent allocation matrix $\boldsymbol{\Gamma}$. In that case the estimates would be obtained as a unique solution to a series of penalized linear regressions. On the other hand, if $\boldsymbol{\Delta}$ were known, then $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}$ could be easily inferred. This "chicken-and-egg" problem can be resolved iteratively by alternating between two steps. Given $\boldsymbol{\Delta}^{(m)}$ at the m^{th} iteration, the E-step computes expected sufficient statistics of hidden/missing data $(\boldsymbol{\Gamma}, \boldsymbol{\Omega})$. The M-step then follows to find the a-posteriori most likely $\boldsymbol{\Delta}^{(m+1)}$, given the expected sufficient statistics. These two steps form the basis of a vanilla EM algorithm with a guaranteed monotone convergence to at least a local posterior mode.

More formally, the EM algorithm locates modes of $\pi(\boldsymbol{\Delta} \,|\, \boldsymbol{Y})$ iteratively by maximizing the expected logarithm of the augmented posterior. Given an initialization $\boldsymbol{\Delta}^{(0)}$, the $(m+1)^{st}$ step of the algorithm outputs $\boldsymbol{\Delta}^{(m+1)} = \arg\max_{\boldsymbol{\Delta}} Q\left(\boldsymbol{\Delta}\right)$, where

$$Q\left(\boldsymbol{\Delta}\right) = \mathsf{E}_{\boldsymbol{\Gamma},\boldsymbol{\Omega} \,|\, \boldsymbol{Y},\boldsymbol{\Delta}^{(m)}}\left[\log \pi\left(\boldsymbol{\Delta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega} \,|\, \boldsymbol{Y}\right)\right], \tag{3.1}$$

with $\mathsf{E}_{\boldsymbol{\Gamma},\boldsymbol{\Omega} \,|\, \boldsymbol{Y},\boldsymbol{\Delta}^{(m)}}(\cdot)$ denoting the conditional expectation given the observed data and current parameter estimates at the m^{th} iteration. Note that we have parametrized our posterior in terms of the ordered inclusion probabilities $\boldsymbol{\theta}$ rather than the breaking fractions $\boldsymbol{\nu}$. These can be recovered using the stick-breaking relationship $\nu_k = \theta_{(k)}/\theta_{(k-1)}$. This parametrization yields a feasible M-step, as will be seen below.

We now take a closer look at the objective function (3.1). For notational convenience, let $\langle X \rangle$ denote the conditional expectation $\mathsf{E}_{\boldsymbol{\Gamma},\boldsymbol{\Omega} \,|\, \boldsymbol{Y},\boldsymbol{\Delta}^{(m)}}(X)$. As a consequence of the hierarchical separation of model parameters, $(\boldsymbol{B}, \boldsymbol{\Sigma})$ and $\boldsymbol{\theta}$ are conditionally independent given $(\boldsymbol{\Omega}, \boldsymbol{\Gamma})$. Thereby

$$Q(\boldsymbol{\Delta}) = Q_1(\boldsymbol{B}, \boldsymbol{\Sigma}) + Q_2(\boldsymbol{\theta}),$$

where $Q_1(\boldsymbol{B}, \boldsymbol{\Sigma}) = \langle \log \pi(\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}, \boldsymbol{\Gamma} \,|\, \boldsymbol{Y}) \rangle$ and $Q_2(\boldsymbol{\theta}) = \langle \log \pi(\boldsymbol{\theta}, \boldsymbol{\Gamma} \,|\, \boldsymbol{Y}) \rangle$. The function $Q_2(\boldsymbol{\theta})$ can be further simplified by noting that the latent indicators γ_{jk} enter linearly and thereby can be directly replaced by their expectations $\langle \gamma_{jk} \rangle$, yielding $Q_2(\boldsymbol{\theta}) = \log \pi(\langle \boldsymbol{\Gamma} \rangle \,|\, \boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})$. The term $Q_1(\boldsymbol{B}, \boldsymbol{\Sigma})$ is also linear in $\boldsymbol{\Gamma}$, but involves quadratic terms $\boldsymbol{\omega}_i \boldsymbol{\omega}_i'$, namely

$$\begin{aligned} Q_1\left(\boldsymbol{B}, \boldsymbol{\Sigma}\right) =& C - \frac{1}{2} \sum_{i=1}^{n} \left\{ \left(\boldsymbol{y}_i - \boldsymbol{B}\langle \boldsymbol{\omega}_i \rangle\right)' \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{y}_i - \boldsymbol{B}\langle \boldsymbol{\omega}_i \rangle\right) + \mathsf{tr}\left[\boldsymbol{B}' \boldsymbol{\Sigma}^{-1} \boldsymbol{B} (\langle \boldsymbol{\omega}_i \boldsymbol{\omega}_i' \rangle - \langle \boldsymbol{\omega}_i \rangle \langle \boldsymbol{\omega}_i \rangle') \right] \right\} \\ & - \sum_{j=1}^{G} \sum_{k=1}^{K^\star} |\beta_{jk}| \left(\lambda_1 \langle \gamma_{jk} \rangle + \lambda_0 (1 - \langle \gamma_{jk} \rangle)\right) - \frac{n+1}{2} \sum_{j=1}^{G} \log \sigma_j^2 - \sum_{j=1}^{G} \frac{1}{2\sigma_j^2}, \end{aligned} \tag{3.2}$$

where C is a constant not involving $\boldsymbol{\Delta}$. The E-step entails the computation of both first and second conditional moments of the latent factors. The updates are summarized below.

3.1 The E-step

The conditional posterior mean vector $\langle \boldsymbol{\omega}_i \rangle$ is obtained as a solution to a ridge-penalized regression of $\boldsymbol{\Sigma}^{(m)^{-1/2}} \boldsymbol{y}_i$ on $\boldsymbol{\Sigma}^{(m)^{-1/2}} \boldsymbol{B}^{(m)}$. This yields

$$\langle \boldsymbol{\Omega}' \rangle = \left(\boldsymbol{B}^{(m)'} \boldsymbol{\Sigma}^{(m)^{-1}} \boldsymbol{B}^{(m)} + \mathrm{I}_{K^\star} \right)^{-1} \boldsymbol{B}^{(m)'} \boldsymbol{\Sigma}^{(m)^{-1}} \boldsymbol{Y}'. \tag{3.3}$$

The conditional second moments are then obtained from $\langle \boldsymbol{\omega}_i \boldsymbol{\omega}_i' \rangle = \boldsymbol{M} + \langle \boldsymbol{\omega}_i \rangle \langle \boldsymbol{\omega}_i \rangle'$, where $\boldsymbol{M} = \left(\boldsymbol{B}^{(m)'} \boldsymbol{\Sigma}^{(m)^{-1}} \boldsymbol{B}^{(m)} + \mathrm{I}_{K^\star} \right)^{-1}$ is the conditional covariance matrix of the latent factors, which does not depend on i. We note in passing that the covariance matrix $\boldsymbol{M}$ can be regarded as a kernel of a smoothing penalty.

The E-step then proceeds by updating the expectation of the binary allocations $\boldsymbol{\Gamma}$. The entries can be updated individually by noting that conditionally on $\boldsymbol{\Delta}^{(m)}$, the γ_{jk}'s are independent. The model hierarchy separates the indicators from the data through the factor loadings so that $\pi(\boldsymbol{\Gamma} \mid \boldsymbol{\Delta}, \boldsymbol{Y}) = \pi(\boldsymbol{\Gamma} \mid \boldsymbol{\Delta})$ does not depend on $\boldsymbol{Y}$. This leads to rapidly computable updates

$$\langle \gamma_{jk} \rangle \equiv \mathsf{P}\left(\gamma_{jk} = 1 \mid \boldsymbol{\Delta}^{(m)} \right) = \frac{\theta_k^{(m)} \phi(\beta_{jk}^{(m)} | \lambda_1)}{\theta_k^{(m)} \phi(\beta_{jk}^{(m)} | \lambda_1) + (1 - \theta_k^{(m)}) \phi(\beta_{jk}^{(m)} | \lambda_0)}. \tag{3.4}$$

As shown in the next section, the conditional inclusion probabilities $\langle \gamma_{jk} \rangle$ serve as adaptive mixing proportions between spike and slab penalties, determining the amount of shrinkage of the associated β_{jk}'s.

3.2 The M-step

Once the latent sufficient statistics have been updated, the M-step consists of maximizing (3.1) with respect to the unknown parameters $\boldsymbol{\Delta}$. Due to the separability of $(\boldsymbol{B}, \boldsymbol{\Sigma})$ and $\boldsymbol{\theta}$, these groups of parameters can be optimized independently. The next theorem explains how $Q_1(\boldsymbol{B}, \boldsymbol{\Sigma})$ can be interpreted as a log-posterior arising from a series of independent penalized regressions, facilitating the exposition of the M-step. First, let us denote $\boldsymbol{Y} = [\boldsymbol{y}^1, \ldots, \boldsymbol{y}^G]$, $\langle \boldsymbol{\Omega} \rangle = [\langle \boldsymbol{\omega}_1 \rangle, \ldots, \langle \boldsymbol{\omega}_n \rangle]'$ and let $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G$ be the columns of $\boldsymbol{B}'$.

Theorem 3.1. *Denote by* $\widetilde{\boldsymbol{Y}} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0}^{K^\star \times K^\star} \end{pmatrix} \in \mathbb{R}^{(n+K^\star) \times G}$ *a zero-augmented data matrix with column vectors* $\widetilde{\boldsymbol{y}}^1, \ldots, \widetilde{\boldsymbol{y}}^G$. *Let* $\widetilde{\boldsymbol{\Omega}} = \begin{pmatrix} \langle \boldsymbol{\Omega} \rangle \\ \sqrt{n}\, \boldsymbol{M}_L \end{pmatrix} \in \mathbb{R}^{(n+K^\star) \times G}$, *where* $\boldsymbol{M}_L$ *is the lower Cholesky factor of* $\boldsymbol{M}$. *Then*

$$Q_1(\boldsymbol{B}, \boldsymbol{\Sigma}) = \sum_{j=1}^{G} Q_j(\boldsymbol{\beta}_j, \sigma_j), \tag{3.5}$$

where

$$Q_j(\boldsymbol{\beta}_j, \sigma_j) = -\frac{1}{2\sigma_j^2}||\widetilde{\boldsymbol{y}}^j - \widetilde{\boldsymbol{\Omega}}\boldsymbol{\beta}_j||^2 - \sum_{k=1}^{K^\star} |\beta_{jk}|\lambda_{jk} - \frac{n+1}{2}\log\sigma_j^2 - \frac{1}{2\sigma_j^2} \tag{3.6}$$

with $\lambda_{jk} = \langle\gamma_{jk}\rangle\lambda_1 + (1 - \langle\gamma_{jk}\rangle)\lambda_0$.

Proof. The statement follows by rearranging the terms in the first row of (3.2). Namely, in the likelihood term we replace the row summation by a column summation. Then, we rewrite $\frac{n}{2}\,\mathsf{tr}(\boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{M}) = \sum_{j=1}^{G}\frac{n}{2\sigma_j^2}\boldsymbol{\beta}_j'\boldsymbol{M}\boldsymbol{\beta}_j$. This quadratic penalty can be embedded within the likelihood term by augmenting the data rows as stated in the theorem. □

Remark 3.1. *The proof of Theorem 3.1 explains why* $\boldsymbol{M}$ *can be regarded as a kernel of a Markov Random Field smoothing prior, penalizing linear combinations of loadings associated with correlated factors.*

Based on the previous theorem, each $\boldsymbol{\beta}_j^{(m+1)}$ (the j^{th} row of the matrix $\boldsymbol{B}^{(m+1)}$) can be obtained by deploying an "adaptive LASSO" computation (Zou, 2006) with a response $\widetilde{\boldsymbol{y}}^j$ and augmented data matrix $\widetilde{\boldsymbol{\Omega}}$. Each coefficient β_{jk} is associated with a unique penalty parameter $2\sigma_j^{(m)}\lambda_{jk}$, which is proportional to λ_{jk}, an adaptive convex combination of the spike and slab LASSO penalties. Notably, each λ_{jk} yields a "self-adaptive" penalty, informed by the data through the most recent $\boldsymbol{\beta}_{jk}^{(m)}$ at the m^{th} iteration.

The computation is made feasible with the very fast LASSO implementations (Friedman et al., 2010), which scale very well with both $K^\star$ and n. First, the data matrix is reweighted by a vector $(1/\lambda_{j1}, \ldots, 1/\lambda_{jK^\star})'$ and a standard LASSO computation is carried out with a penalty $2\sigma_j^{(m)}$. The resulting estimate is again reweighted by $1/\lambda_{jk}$ (Zou, 2006), yielding $\boldsymbol{\beta}_j^{(m+1)}$. Note that the updates $\boldsymbol{\beta}_j^{(m+1)}$ $(j = 1, \ldots, G)$ are obtained conditionally on $\boldsymbol{\Sigma}^{(m)}$ and are independent of each other, permitting the use of distributed computing. This step is followed by a closed form update $\boldsymbol{\Sigma}^{(m+1)}$ with $\sigma_j^{(m+1)2} = \frac{1}{n+1}(||\boldsymbol{y}^j - \widetilde{\boldsymbol{\Omega}}\boldsymbol{\beta}_j^{(m+1)}||^2 + 1)$, conditionally on the new $\boldsymbol{B}^{(m+1)}$. Despite proceeding conditionally in the M-step, monotone convergence is still guaranteed (Meng and Rubin, 1993).

Now we continue with the update of the ordered inclusion probabilities $\boldsymbol{\theta}$ from the stick-breaking construction. To motivate the benefits of parametrization based on $\boldsymbol{\theta}$, let us for a moment assume that we are actually treating the breaking fractions $\boldsymbol{\nu}$ as the parameters of interest. The corresponding objective function $Q_2^\star(\boldsymbol{\nu}) = \log\pi(\langle\boldsymbol{\Gamma}\rangle \mid \boldsymbol{\nu}) + \log\pi(\boldsymbol{\nu})$ is

$$Q_2^\star(\boldsymbol{\nu}) = \sum_{j=1}^{G}\sum_{k=1}^{K^\star}\left\{\langle\gamma_{jk}\rangle\sum_{l=1}^{k}\log\nu_l + (1 - \langle\gamma_{jk}\rangle)\log\left(1 - \prod_{l=1}^{k}\nu_l\right)\right\} + (\alpha - 1)\sum_{k=1}^{K^\star}\log(\nu_k), \tag{3.7}$$

a nonlinear function that is difficult to optimize. Instead, we use the stick-breaking law and plug $\nu_k = \theta_{(k)}/\theta_{(k-1)}$ into (3.7). The objective function then becomes

$$Q_2(\boldsymbol{\theta}) = \sum_{j=1}^{G}\sum_{k=1}^{K^\star}\left[\langle\gamma_{jk}\rangle \log\theta_k + (1-\langle\gamma_{jk}\rangle)\log(1-\theta_k)\right] + (\alpha-1)\log\theta_{K^\star}, \tag{3.8}$$

whose maximum $\boldsymbol{\theta}^{(m+1)}$ can be found by solving a linear program with a series of constraints

$$\begin{aligned}\theta_k - \theta_{k-1} &\leq 0,\ k = 2,\ldots,K^\star,\\ 0 \leq \theta_k &\leq 1,\ k = 1,\ldots,K^\star.\end{aligned}$$

Had we assumed the finite beta-Bernoulli prior (2.3), the update of the (unordered) occurrence probabilities would simply become $\theta_k^{(m)} = \frac{\sum_{j=1}^{G}\langle\gamma_{jk}\rangle + a - 1}{a+b+G-2}$.

Note that the the ordering constraint here induces increasing shrinkage of higher-indexed factor loadings, thereby controlling the growth of the effective factor cardinality.

4 Rotational Ambiguity and Parameter Expansion

The EM algorithm outlined in the previous section is prone to entrapment in local modes in the vicinity of initialization. This local convergence issue is exacerbated by the rotational ambiguity of the likelihood, which induces highly multimodal posteriors, and by strong couplings between the updates of loadings and factors. These couplings cement the initial factor orientation, which may be suboptimal, and affect the speed of convergence with zigzagging update trajectories. These issues can be alleviated with additional augmentation in the parameter space that can dramatically accelerate the convergence (Liu et al., 1998; van Dyk and Meng, 2010, 2001; Liu and Wu, 1999; Lewandowski et al., 1999). By embedding the complete data model within a larger model with extra parameters, we derive a variant of a parameter expanded EM algorithm (PX-EM by Liu et al. (1998)). This enhancement performs an "automatic rotation to sparsity", gearing the algorithm towards orientations which best match the prior assumptions of independent latent components and sparse loadings. A key to our approach is to employ the parameter expansion only on the likelihood portion of the posterior, while using the SSL prior to guide the algorithm towards sparse factor orientations. We refer to our variant as parameter-expanded-likelihood EM (PXL-EM).

Our PXL-EM algorithm is obtained with the following parameter expanded version of (1.1)

$$\boldsymbol{y}_i \,|\, \boldsymbol{\omega}_i, \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{A} \overset{\text{ind}}{\sim} \mathcal{N}_G(\boldsymbol{B}\boldsymbol{A}_L^{-1}\boldsymbol{\omega}_i, \boldsymbol{\Sigma}), \quad \boldsymbol{\omega}_i \,|\, \boldsymbol{A} \sim \mathcal{N}_K(\boldsymbol{0}, \boldsymbol{A}) \tag{4.1}$$

for $1 \leq i \leq n$, where $\boldsymbol{A}_L$ denotes the lower Cholesky factor of $\boldsymbol{A}$, the newly introduced parameter. The observed-data likelihood here is invariant under the parametrizations indexed by $\boldsymbol{A}$. This is evident

from the marginal distribution $f(\boldsymbol{y}_i \,|\, \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{A}) = \mathcal{N}_G(\boldsymbol{0}, \boldsymbol{B}\boldsymbol{B}' + \boldsymbol{\Sigma})$, $1 \leq i \leq n$, which does not depend on $\boldsymbol{A}$. Although $\boldsymbol{A}$ is indeterminate from the observed data, it can be identified with the complete data. Note that the original factor model is preserved at the null value $\boldsymbol{A}_0 = \mathrm{I}_K$.

To exploit the invariance of the parameter expanded likelihood, we impose the SSL prior (2.6) on $\boldsymbol{B}^\star = \boldsymbol{B}\boldsymbol{A}_L^{-1}$ rather than on $\boldsymbol{B}$. That is,

$$\beta^\star_{jk} \,|\, \gamma_{jk} \overset{\text{ind}}{\sim} SSL(\lambda_{0k}, \lambda_1), \quad \gamma_{jk} \,|\, \theta_{(k)} \overset{\text{ind}}{\sim} \text{Bernoulli}[\theta_{(k)}], \quad \theta_{(k)} = \prod_{l=1}^{k} \nu_l, \quad \nu_l \overset{\text{iid}}{\sim} \mathcal{B}(\alpha, 1), \tag{4.2}$$

where the $\beta^\star_{jk}$'s are the transformed elements of $\boldsymbol{B}^\star$. This yields an implicit prior on $\boldsymbol{B}$ that depends on $\boldsymbol{A}_L$ and therefore is not transformation invariant, a crucial property for anchoring sparse factor orientations. The original factor loadings $\boldsymbol{B}$ can be recovered from $(\boldsymbol{B}^\star, \boldsymbol{A})$ through the reduction function $\boldsymbol{B} = \boldsymbol{B}^\star \boldsymbol{A}_L$. The prior (2.2) on $\boldsymbol{\Sigma}$ remains unchanged.

As the EM algorithm from Section 3, PXL-EM also targets (local) maxima of the posterior $\pi(\boldsymbol{\Delta} \,|\, \boldsymbol{Y})$ implied by (1.1) and (2.1), but does so in a very different way. PXL-EM proceeds indirectly in terms of the parameter expanded posterior $\pi(\boldsymbol{\Delta}^\star \,|\, \boldsymbol{Y})$ indexed by $\boldsymbol{\Delta}^\star = (\boldsymbol{B}^\star, \boldsymbol{\Sigma}, \boldsymbol{\theta}, \boldsymbol{A})$ and implied by (4.1) and (4.2). By iteratively optimizing the conditional expectation of the augmented log posterior $\log \pi(\boldsymbol{\Delta}^\star, \boldsymbol{\Omega}, \boldsymbol{\Gamma} \,|\, \boldsymbol{Y})$, PXL-EM yields a path in the expanded parameter space. This sequence corresponds to a trajectory in the original parameter space through the reduction function $\boldsymbol{B} = \boldsymbol{B}^\star \boldsymbol{A}_L$. At convergence, this trajectory yields a local mode of $\pi(\boldsymbol{\Delta} \,|\, \boldsymbol{Y})$ (as shown in Theorem 4.1). Importantly, the E-step of PXL-EM is taken with respect to the conditional distribution of $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$ under the *original model* governed by $\boldsymbol{B}$ and $\boldsymbol{A}_0$, rather than under the expanded model governed by $\boldsymbol{B}^*$ and $\boldsymbol{A}$. Such anchoring by the E-step is crucial and its implications are discussed later in this section.

The PXL-EM can lead to very different trajectories, with an M-step that updates $\boldsymbol{A}$ together with the remaining parameters. Recall that $\boldsymbol{A}$ indexes continuous transformations yielding the same marginal likelihood. Adding this extra dimension, each mode of the original posterior $\pi(\boldsymbol{\Delta} \mid \boldsymbol{Y})$ corresponds to a curve in the expanded posterior $\pi(\boldsymbol{\Delta}^\star \mid \boldsymbol{Y})$, indexed by $\boldsymbol{A}$. These ridge-lines of accumulated probability, or orbits of equal likelihood, serve as a bridges connecting remote posterior modes. These are likely to be located along coordinate axes due to the SSL prior. By updating $\boldsymbol{A}$ and using the reduction function, the PXL-EM trajectory is geared along the orbits, taking greater steps and accelerating convergence.

More formally, the PXL-EM traverses the expanded parameter space and generates a trajectory $\{\boldsymbol{\Delta}^{\star(1)}, \boldsymbol{\Delta}^{\star(2)}, \ldots\}$, where $\boldsymbol{\Delta}^{\star(m)} = (\boldsymbol{B}^{\star(m)}, \boldsymbol{\Sigma}^{(m)}, \boldsymbol{\theta}^{(m)}, \boldsymbol{A}^{(m)})$. This trajectory corresponds to a sequence $\{\boldsymbol{\Delta}^{(1)}, \boldsymbol{\Delta}^{(2)}, \ldots\}$ in the reduced parameter space, where $\boldsymbol{\Delta}^{(m)} = (\boldsymbol{B}^{(m)}, \boldsymbol{\Sigma}^{(m)}, \boldsymbol{\theta}^{(m)})$ and $\boldsymbol{B}^{(m)} = \boldsymbol{B}^{\star(m)} \boldsymbol{A}^{(m)1/2}$. Beginning with the initialization $\boldsymbol{\Delta}^{(0)}$, every step of the PXL-EM algorithm

outputs an update $\boldsymbol{\Delta}^{\star(m+1)} = \arg\max_{\boldsymbol{\Delta}^\star} Q^{px}(\boldsymbol{\Delta}^\star)$, where

$$Q^{px}(\boldsymbol{\Delta}^\star) = \mathsf{E}_{\boldsymbol{\Omega},\boldsymbol{\Gamma} \mid \boldsymbol{Y},\boldsymbol{\Delta}^{(m)},\boldsymbol{A}_0} \log \pi(\boldsymbol{\Delta}^\star, \boldsymbol{\Omega}, \boldsymbol{\Gamma} \mid \boldsymbol{Y}). \tag{4.3}$$

Each such computation is facilitated by the separability of Q^{px} with respect to $(\boldsymbol{B}^\star, \boldsymbol{\Sigma})$, $\boldsymbol{\theta}$ and $\boldsymbol{A}$, a consequence of the hierarchical structure of the Bayesian model. Thus we can write

$$Q^{px}(\boldsymbol{\Delta}^\star) = C^{px} + Q_1^{px}(\boldsymbol{B}^\star, \boldsymbol{\Sigma}) + Q_2^{px}(\boldsymbol{\theta}) + Q_3^{px}(\boldsymbol{A}). \tag{4.4}$$

The function Q_1^{px} is given by Q_1 in (3.5), Q_2^{px} is given by Q_2 in (3.8) and

$$Q_3^{px}(\boldsymbol{A}) = -\frac{1}{2}\sum_{i=1}^{n} \operatorname{tr}[\boldsymbol{A}^{-1}\mathsf{E}_{\boldsymbol{\Omega} \mid \boldsymbol{\Delta}^{(m)},\boldsymbol{A}_0}(\boldsymbol{\omega}_i\boldsymbol{\omega}_i')] - \frac{n}{2}\log|\boldsymbol{A}|. \tag{4.5}$$

Recall that the expectation in (4.3) is taken with respect to the conditional distribution of $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$ under the original model governed by $\boldsymbol{\Delta}^{(m)}$ and $\boldsymbol{A}_0$. Formally, the calculations remain the same as before (Section 3.1). However, the update $\boldsymbol{B}^{(m)} = \boldsymbol{B}^{\star(m)}\boldsymbol{A}_L^{(m)}$ is now used throughout these expressions. The implications of this substitution are discussed in Examples 4.1 and 4.2 later in this section.

Conditionally on the imputed latent data, the M-step is then performed by maximizing $Q^{px}(\boldsymbol{\Delta}^\star)$ over $\boldsymbol{\Delta}^\star$ in the augmented space. The updates of $(\boldsymbol{B}^{\star(m+1)}, \boldsymbol{\Sigma}^{(m+1)})$ and $\boldsymbol{\theta}^{(m+1)}$ are obtained by maximizing (3.5) and (3.8) as described in Section 3.2. The update of $\boldsymbol{A}^{(m+1)}$, obtained by maximizing (4.5), requires only a fast simple operation,

$$\boldsymbol{A}^{(m+1)} = \max_{\boldsymbol{A}=\boldsymbol{A}',\boldsymbol{A}\geq 0} Q_3^{px}(\boldsymbol{A}) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}_{\boldsymbol{\Omega} \mid \boldsymbol{Y},\boldsymbol{\Delta}^{(m)},\boldsymbol{A}_0}(\boldsymbol{\omega}_i\boldsymbol{\omega}_i') = \frac{1}{n}\langle\boldsymbol{\Omega}'\boldsymbol{\Omega}\rangle = \frac{1}{n}\langle\boldsymbol{\Omega}\rangle'\langle\boldsymbol{\Omega}\rangle + \boldsymbol{M}. \tag{4.6}$$

The new coefficient updates in the reduced parameter space are then obtained by the following step $\boldsymbol{B}^{(m+1)} = \boldsymbol{B}^{\star(m+1)}\boldsymbol{A}_L^{(m+1)}$, a "rotation" along an orbit of equal likelihood. The additional computational cost is rather small because the transformations are performed in the lower-dimensional latent subspace.

The following theorem shows that PXL-EM and EM from Section 3 have the same fixed points.

Theorem 4.1. *A value* $(\boldsymbol{\Delta}_{MAP}, \boldsymbol{A}_0)$ *is a fixed point of the PXL-EM algorithm.*

Proof. The proof is analogous to the one of Liu et al. (1998) for PX-EM.

Remark 4.1. *According to the theorem,* $\boldsymbol{A}^{(m+1)} \approx \boldsymbol{A}_0$ *near convergence, where PXL-EM approximately corresponds to the traditional EM algorithm. Equation* (4.6) *then yields* $\frac{1}{n}\langle\boldsymbol{\Omega}'\boldsymbol{\Omega}\rangle \approx \mathrm{I}_K$*, implying that PXL-EM is ultimately enforcing an implicit identifiability constraint based on orthogonal features* $\widetilde{\boldsymbol{\Omega}}$*.*

Remark 4.2. *Although $\boldsymbol{A}_L$ is not strictly a rotation matrix in the sense of being orthonormal, we will refer to its action of changing the factor model orientation as the "rotation by $\boldsymbol{A}_L$". From the polar decomposition of $\boldsymbol{A}_L = \boldsymbol{U}\boldsymbol{P}$, transformation by $\boldsymbol{A}_L$ is the composition of a rotation represented by the orthogonal matrix $\boldsymbol{U} = \boldsymbol{A}_L(\boldsymbol{A}_L'\boldsymbol{A}_L)^{-1/2}$, and a dilation represented by the symmetric matrix $\boldsymbol{P}$. Thus, when we refer to the "rotation" by $\boldsymbol{A}_L$, what is meant is the rotational aspect of $\boldsymbol{A}_L$ contained within $\boldsymbol{U}$. The result of applying the sequence of affine transformations $\boldsymbol{A}_L$ throughout the computation will be referred to as a "rotation to sparsity". At convergence, these transformations are likely to yield better oriented sparse representations.*

Remark 4.3. *Instead of $\boldsymbol{A}_L$, we could as well deploy the square root $\boldsymbol{A}^{1/2}$. We have chosen $\boldsymbol{A}_L$ due to its lower-triangular structure, whose benefits will be highlighted in Example 4.2.*

To sum up, the default EM algorithm proceeds by finding $\boldsymbol{B}^{(m)}$ at the M-step, and then using that $\boldsymbol{B}^{(m)}$ for the next E-step. In contrast, the PXL-EM algorithm finds $\boldsymbol{B}^{\star(m)}$ at the M-step, but then uses the value of $\boldsymbol{B}^{(m)} = \boldsymbol{B}^{\star(m)}\boldsymbol{A}_L^{(m)}$ for the next E-step. Each transformation $\boldsymbol{B}^{(m)} = \boldsymbol{B}^{\star(m)}\boldsymbol{A}_L^{(m)}$ decouples the most recent updates of the latent factors and factor loadings, enabling the EM trajectory to escape the attraction of suboptimal orientations. In this, the "rotation" induced by $\boldsymbol{A}_L^{(m)}$ plays a crucial role for the detection of sparse representations which are tied to the orientation of the factors.

In the following two intentionally simple examples, we convey the intuition of entries in $\boldsymbol{A}$ as penalties that encourage featurizations with fewer and more informative factors. The first example describes the scaling aspect of the transformation by $\boldsymbol{A}_L$, assuming $\boldsymbol{A}_L$ is diagonal.

Example 4.1. *(Diagonal $\boldsymbol{A}$) We show that for $\boldsymbol{A} = \operatorname{diag}\{\alpha_1, \ldots, \alpha_K\}$, each α_k plays a role of a penalty parameter, determining the size of new features as well as the amount of shrinkage. This is seen from the E-step, which (a) creates new features $\langle\boldsymbol{\Omega}\rangle$, (b) determines penalties for variable selection $\langle\boldsymbol{\Gamma}\rangle$, (c) creates a smoothing penalty matrix $\mathsf{Cov}\,(\boldsymbol{\omega}_i \mid \boldsymbol{B}, \boldsymbol{\Sigma})$. Here is how inserting $\boldsymbol{B}$ instead of $\boldsymbol{B}^\star$ affects these three steps. For simplicity, assume $\boldsymbol{\Sigma} = \mathrm{I}_K$, $\boldsymbol{B}^{\star\prime}\boldsymbol{B}^\star = \mathrm{I}_K$ and $\boldsymbol{\theta} = (0.5, \ldots, 0.5)'$. From* (3.3)*, the update of latent features is*

$$\mathsf{E}_{\boldsymbol{\Omega}\,|\,\boldsymbol{Y},\boldsymbol{B}}(\boldsymbol{\Omega}') = \boldsymbol{A}_L^{-1}(\mathrm{I}_K + \boldsymbol{A}^{-1})^{-1}\boldsymbol{B}^{\star\prime}\boldsymbol{Y}' = \operatorname{diag}\left\{\frac{\sqrt{\alpha}_k}{1+\alpha_k}\right\}\boldsymbol{B}^{\star\prime}\boldsymbol{Y}'. \tag{4.7}$$

Note that (4.7) *with $\alpha_k = 1\,(k = 1, \ldots, K)$ corresponds to no parameter expansion. The function $f(\alpha) = \frac{\sqrt{\alpha}}{1+\alpha}$ steeply increases up to its maximum at $\alpha = 1$ and then slowly decreases. Before the convergence (which corresponds to $\alpha_k \approx 1$), PXL-EM performs shrinkage of features, which is more dramatic if the k^{th} variance α_k is close to zero. Regarding the Markov-field kernel smoothing penalty, the coordinates with higher variances α_k are penalized less. This is seen from $\mathsf{Cov}\,(\boldsymbol{\omega}_i \mid \boldsymbol{B}, \boldsymbol{\Sigma}) = \boldsymbol{A}_L^{-1}(\mathrm{I}_K + \boldsymbol{A}^{-1})^{-1}\boldsymbol{A}_L^{-1} = \operatorname{diag}\{1/(1+\alpha_k)\}$. The E-step is then completed with the update of variable*

selection penalty mixing weights $\langle \boldsymbol{\Gamma} \rangle$. *Here*

$$\mathsf{E}_{\boldsymbol{\Gamma} \mid \boldsymbol{B},\boldsymbol{\theta}}(\gamma_{jk}) = \left[1 + \frac{\lambda_0}{\lambda_1} \exp\left(-|\beta^\star_{jk}|\alpha_k(\lambda_0 - \lambda_1)\right)\right]^{-1}.$$

This probability is exponentially increasing in α_k. *Higher variances* $\alpha_k > 1$ *increase the inclusion probability as compared to no parameter expansion* $\alpha_k = 1$. *The coefficients of the newly created features with larger* α_k *are more likely to be selected.*

The next example illustrates the rotational aspect of $\boldsymbol{A}_L$, where the off-diagonal elements perform linear aggregation.

Example 4.2. *(Unit lower-triangular* $\boldsymbol{A}_L$*) Suppose that* $\boldsymbol{A} = (\alpha_{jk})_{j,k=1}^K$ *with* $\alpha_{jk} = \min\{j,k\}$. *This matrix has a unit-lower-triangular Cholesky factor with entries* $A^L_{jk} = \mathbb{I}(j \geq k)$. *For* $\boldsymbol{\Sigma} = \mathrm{I}_K$ *and* $\boldsymbol{B}^{\star\prime}\boldsymbol{B}^\star = \mathrm{I}_K$, *we have again* $\mathsf{E}_{\boldsymbol{\Omega} \mid \boldsymbol{Y},\boldsymbol{B}}(\boldsymbol{\Omega}') = \boldsymbol{A}_L^{-1}(\mathrm{I}_K + \boldsymbol{A}^{-1})^{-1}\boldsymbol{B}^{\star\prime}\boldsymbol{Y}'$, *where now the matrix* $\boldsymbol{A}_L^{-1}(\mathrm{I}_K + \boldsymbol{A}^{-1})^{-1}$ *has positive elements in the upper triangle and negative elements in the lower triangle. The first feature thus aggregates information from all the columns* $\boldsymbol{Y}\boldsymbol{B}^\star$, *whereas the last feature correlates negatively with all but the last column. The variable selection indicators are computed for linear aggregates of the coefficients* $\boldsymbol{B}^\star$, *as determined by the entires* $\boldsymbol{A}_L$, *i.e.*

$$\mathsf{E}_{\boldsymbol{\Gamma} \mid \boldsymbol{B},\boldsymbol{\theta}}(\gamma_{jk}) = \left[1 + \frac{\lambda_0}{\lambda_1} \exp\left(-\sum_{l \geq k}^{K^\star} |\beta^\star_{jl}|(\lambda_0 - \lambda_1)\right)\right]^{-1},$$

where $\boldsymbol{\theta} = (0.5, \ldots, 0.5)'$. *The lower ordered inclusion indicators are likely to be higher since they aggregate more coefficients, a consequence of the lower-triangular form* $\boldsymbol{A}_L$. *As a result, lower ordered new features are more likely to be selected. The smoothness penalty matrix* $\mathsf{Cov}\left(\boldsymbol{\omega}_i | \boldsymbol{B}, \boldsymbol{\Sigma}\right) = (\boldsymbol{A}_L'\boldsymbol{A}_L + \mathrm{I}_K)^{-1}$ *has increasing values on the diagonal and all the off-diagonal elements are negative. The quadratic penalty* $\boldsymbol{\beta}_j'\mathsf{Cov}\left(\boldsymbol{\omega}_i \mid \boldsymbol{B}, \boldsymbol{\Sigma}\right)\boldsymbol{\beta}_j$ *thus forces the loadings to be similar (due to the positive covariances between the factors as given in* $\boldsymbol{A}$*), where the penalty is stronger between coefficients of higher-ordered factors.*

Further insight into the role of the matrix $\boldsymbol{A}_L$ can be gained by recasting the LASSO penalized likelihood in the M-step of PXL-EM in terms of the original model parameters $\boldsymbol{B} = \boldsymbol{B}^\star\boldsymbol{A}_L$. From (3.6), the PXL M-step yields

$$\boldsymbol{\beta}_j^{\star(m+1)} = \arg\max_{\boldsymbol{\beta}_j^\star} \left\{ -||\widetilde{\boldsymbol{y}}^j - \widetilde{\boldsymbol{\Omega}}\boldsymbol{\beta}_j^\star||^2 - 2\sigma_j^{(m)2} \sum_{k=1}^{K^\star} |\beta^\star_{jk}|\lambda_{jk} \right\},$$

for each $j = 1, \ldots, G$. However, in terms of the columns of $\boldsymbol{B}^{(m)\prime}$, where $\boldsymbol{\beta}_j^{(m+1)} = \boldsymbol{A}_L'\boldsymbol{\beta}_j^{\star(m+1)}$, these

solutions become

$$\boldsymbol{\beta}_j^{(m+1)} = \arg\max_{\boldsymbol{\beta}_j} \left\{ -||\widetilde{\boldsymbol{y}}^j - (\widetilde{\boldsymbol{\Omega}}\boldsymbol{A}_L^{'-1})\boldsymbol{\beta}_j||^2 - 2\sigma_j^{(m)2} \sum_{k=1}^{K^\star} \left| \sum_{l \geq k}^{K^\star} (\boldsymbol{A}_L^{-1})_{lk}\beta_{jl} \right| \lambda_{jk} \right\}. \tag{4.8}$$

Thus, the $\boldsymbol{\beta}_j^{(m+1)}$ are solutions to modified penalized regressions of $\widetilde{\boldsymbol{y}}^j$ on $\widetilde{\boldsymbol{\Omega}}\boldsymbol{A}_L^{'-1}$ under a series of linear constraints. Here $\boldsymbol{A}_L^{-1}$ serves to "rotate" the factor basis. The linear constraints are a bit more general than those in the fused LASSO (Tibshirani et al., 2005), since they involve linear combinations determined by $\boldsymbol{A}_L^{-1}$, which is not required to be a banded diagonal matrix. This illustrates how PXL-EM cycles through "rotations" of the factor basis and sparsification by soft-thresholding.

The PXL-EM algorithm outlined in this section can be regarded as a one-step-late PX-EM (van Dyk and Tang, 2003) or more generally as a one-step-late EM (Green, 1990). The PXL-EM differs from the traditional PX-EM of Liu et al. (1998) by not requiring the SSL prior be invariant under transformations $\boldsymbol{A}_L$. PXL-EM purposefully leaves only the likelihood invariant, offering (a) tremendous accelerations without sacrificing the computational simplicity, (b) automatic rotation to sparsity and (c) robustness against poor initializations. The price we pay is the guarantee of monotone convergence. Let $(\boldsymbol{\Delta}^{(m)}, \boldsymbol{A}_0)$ be an update of $\boldsymbol{\Delta}^\star$ at the m^{th} iteration. It follows from the information inequality, that for any $\boldsymbol{\Delta} = (\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\theta})$, where $\boldsymbol{B} = \boldsymbol{B}^\star\boldsymbol{A}_L$,

$$\begin{aligned} \log \pi(\boldsymbol{\Delta} \mid \boldsymbol{Y}) - \log \pi(\boldsymbol{\Delta}^{(m)} \mid \boldsymbol{Y}) \geq Q^{px}(\boldsymbol{\Delta}^\star \mid \boldsymbol{\Delta}^{(m)}, \boldsymbol{A}_0) - Q^{px}(\boldsymbol{\Delta}^{(m)} \mid \boldsymbol{\Delta}^{(m)}, \boldsymbol{A}_0) \\ + \mathsf{E}_{\boldsymbol{\Gamma} \mid \boldsymbol{\Delta}^{(m)}, \boldsymbol{A}_0} \log \left(\frac{\pi(\boldsymbol{B}^\star\boldsymbol{A}_L)}{\pi(\boldsymbol{B}^\star)} \right). \end{aligned} \tag{4.9}$$

Whereas $\boldsymbol{\Delta}^{\star(m+1)} = \arg\max Q^{px}(\boldsymbol{\Delta}^\star)$ increases the Q^{px} function, the log prior ratio evaluated at $(\boldsymbol{B}^{\star(m+1)}, \boldsymbol{A}^{(m+1)})$ is generally not positive. van Dyk and Tang (2003) proposed a simple adjustment to monotonize their one-step-late PX-EM, where the new proposal $\boldsymbol{B}^{(m+1)} = \boldsymbol{B}^{\star(m+1)}\boldsymbol{A}_L^{(m+1)}$ is only accepted when the value of the right hand side of (4.9) is positive. Otherwise, the classical EM step is performed with $\boldsymbol{B}^{(m+1)} = \boldsymbol{B}^{\star(m+1)}\boldsymbol{A}_0$. Although this adjustment guarantees the convergence towards the nearest stationary point (Wu, 1983), poor initializations may gear the monotone trajectories towards peripheral modes. It may therefore be beneficial to perform the first couple of iterations according to PXL-EM to escape such initializations, not necessarily improving on the value of the objective, and then to switch to EM or to the monotone adjustment. Monitoring the criterion (6.6) throughout the iterations, we can track the steps in the trajectory that are guaranteed to be monotone. If convergent, PXL-EM converges no slower than EM algorithm (Green, 1990) and the accelerations are dramatic, as will be illustrated in the next section.

The EM acceleration with parameter expansion is related to the general framework of parameter expanded variational Bayes (VB) methods (Qi and Jaakkola, 2006), whose variants were implemented

for factor analysis by Luttinen and Ilin (2010). The main difference here is that we use a parameterization that completely separates the update of auxiliary and model parameters, while breaking up the dependence between factors and loadings. PXL-EM yields transformations that accelerate the convergence towards sparse modes of the actual posterior, not only its lower bound. Parameter expansion has already proven useful in accelerating convergence of sampling procedures, generally (Liu and Wu, 1999) and in factor analysis (Ghosh and Dunson, 2009). What we have considered here is an expansion by a full prior factor covariance matrix, not only its diagonal, to obtain even faster accelerations (Liu et al., 1998).

4.1 Anchoring Factor Rotation: A Synthetic Example

To illustrate the effectiveness of the symbiosis between factor model "rotations" and the spike-and-slab LASSO soft-thresholding, we generated a dataset from model (1.1) with $n = 100$ observations, $G = 1\,956$ responses and $K_{true} = 5$ factors. The true loading matrix $\boldsymbol{B}_{true}$ (Figure 1 left) has a block-diagonal pattern of nonzero elements $\boldsymbol{\Gamma}_{true}$ with overlapping response-factor allocations, where $\sum_j \gamma_{jk}^{true} = 500$ and $\sum_j \gamma_{jk}^{true}\gamma_{jk+1}^{true} = 136$ is the size of the overlap. We set $b_{jk}^{true} = \gamma_{jk}^{true}$ and $\boldsymbol{\Sigma}^{true} = \mathrm{I}_G$. The implied covariance matrix is again block-diagonal (Figure 1 middle). For the EM and PXL-EM factor model explorations, we use $\lambda_{0k} = \lambda_0$, as advocated at the end of the Section 2.1. We set $\lambda_1 = 0.001, \lambda_0 = 20, \alpha = 1/G$ and $K^\star = 20$. The slab penalty λ_1 was set small to ameliorate bias in estimation. The spike penalty λ_0 was set much larger to improve the accuracy of recovery. Its tuning will be addressed in the next section. All the entries in the initialization $\boldsymbol{B}^{(0)}$ were sampled independently from the standard normal distribution, $\boldsymbol{\Sigma}^{(0)} = \mathrm{I}_G$ and $\boldsymbol{\theta}_{(k)}^{(0)} = 0.5$, $k = 1, \ldots, K^\star$.

We compared the EM and PXL-EM implementations with regard to the number of iterations to convergence and the accurateness of the recovery of the loading matrix. Convergence was claimed whenever $d_\infty(\boldsymbol{B}^{\star(m+1)}, \boldsymbol{B}^{\star(m)}) < 0.05$ in the PXL-EM and $d_\infty(\boldsymbol{B}^{(m+1)}, \boldsymbol{B}^{(m)}) < 0.05$ in the EM algorithm.

The results without parameter expansion were rather disappointing. Figure 2 depicts four snapshots of the EM trajectory, from the initialization to the 100^{th} iteration. The plot depicts heat-maps of $|\boldsymbol{B}^{(m)}|$ (a matrix of absolute values of $\boldsymbol{B}^{(m)}$) for $m \in \{0, 1, 10, 100\}$, where the blank entries correspond to zeroes. The EM algorithm did not converge even after 100 iterations, where the recovered factor allocation pattern is nowhere close to the generating truth. On the other hand, parameter expansion fared superbly. Figure 3 shows snapshots of $|\boldsymbol{B}^{\star(m)}|$ for the PXL-EM trajectory at $m \in \{0, 1, 10, 23\}$, where convergence was achieved after merely 23 iterations. Even at the first iteration, PXL-EM began to gravitate towards a sparser and more structured solution. At convergence, PXL-EM recovers the true pattern of nonzero elements in the loading matrix (up to a permutation) with merely 2 false

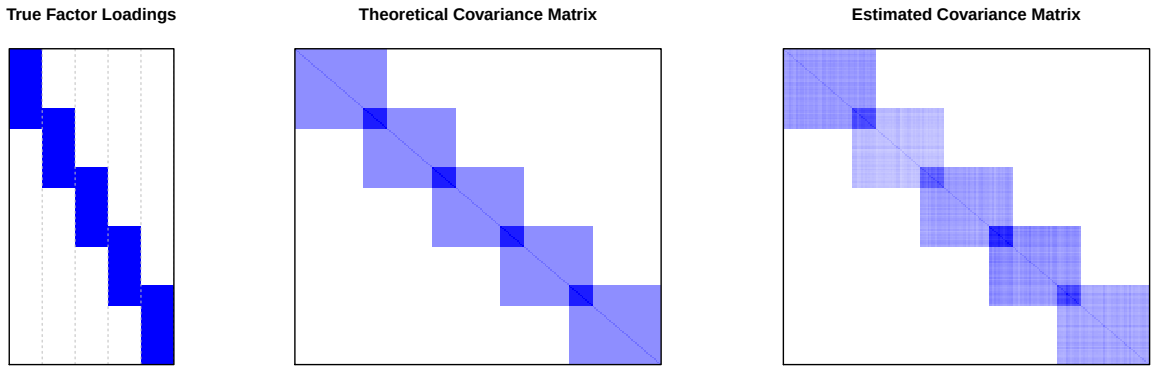
positives and 2 false negative. In addition, we obtain a rather accurate estimate of the marginal covariance matrix (Figure 1(c)). This estimate will be compared with the solution obtained using a one-component Laplace prior in the next section.

The PXL-EM is seen to be robust against poor initializations. After repeating the experiment with different random starting locations $\boldsymbol{B}^{(0)}$ sampled element-wise from Gaussian distributions with larger variances, PXL-EM yielded almost identical loading matrices, again with only a few false positives and negatives. The computing time required for one iteration of PXL-EM was between 3-4 seconds in R without parallelization on a 1.7 GHz machine.

Given the vastness of the posterior with its intricate multimodality, and the arbitrariness of the initialization, the results of this experiment are very encouraging. There are some lingering issues, such as the selection of the penalty parameter λ_0. We shall show in the next section, that the selection is avoided with a sequential deployment of PXL-EM over a sequence of increasing λ_0 values. The larger λ_0, the closer we are to the methodological ideal, and so large values that are practically indistinguishable from the limiting case are preferred.

5 Dynamic Posterior Exploration

Under our hierarchical SSL prior with $\lambda_{0k} = \lambda_0$, the character of the posterior landscape is regulated by the two penalty parameters $\lambda_0 >> \lambda_1$, which determine the degree of multi-modality and spikiness. In the context of simple linear regression, $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \mathrm{I}_n)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, the posterior distribution



(a) $\boldsymbol{B}_{true}$ (b) $\boldsymbol{B}_{true}\boldsymbol{B}'_{true} + \mathrm{I}_5$ (c) $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}' + \mathrm{diag}\{\widehat{\Sigma}\}$

Figure 1: The true pattern of nonzero values in the loading matrix (left), a heat-map of the theoretical covariance matrix $\boldsymbol{B}_{true}\boldsymbol{B}'_{true} + \mathrm{I}_5$ (middle), estimated covariance matrix (right).

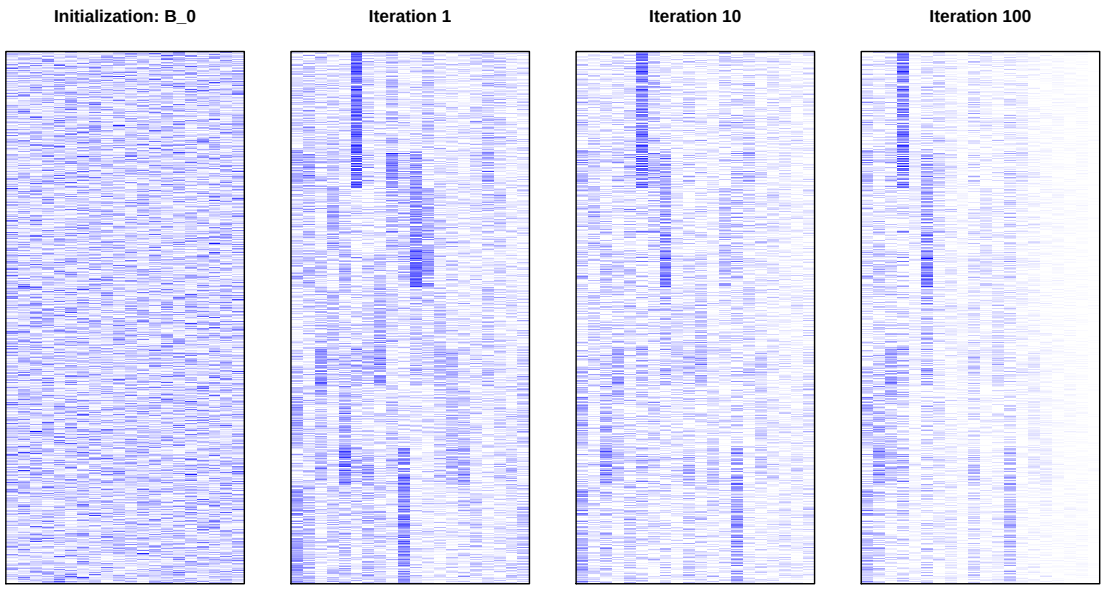


Figure 2: A trajectory of the EM algorithm, convergence not achieved even after 100 iterations

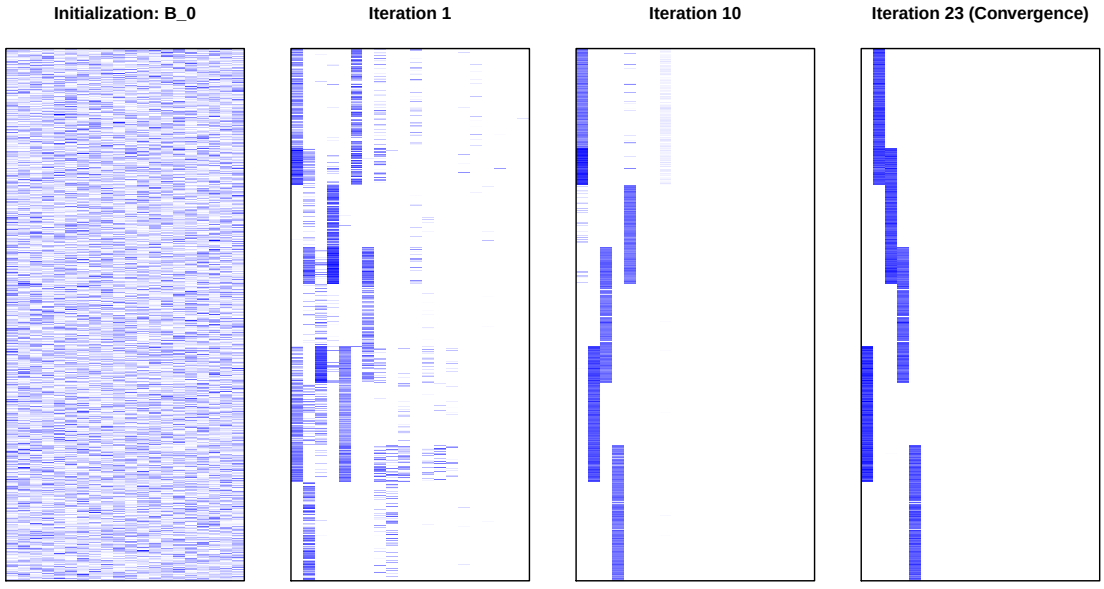
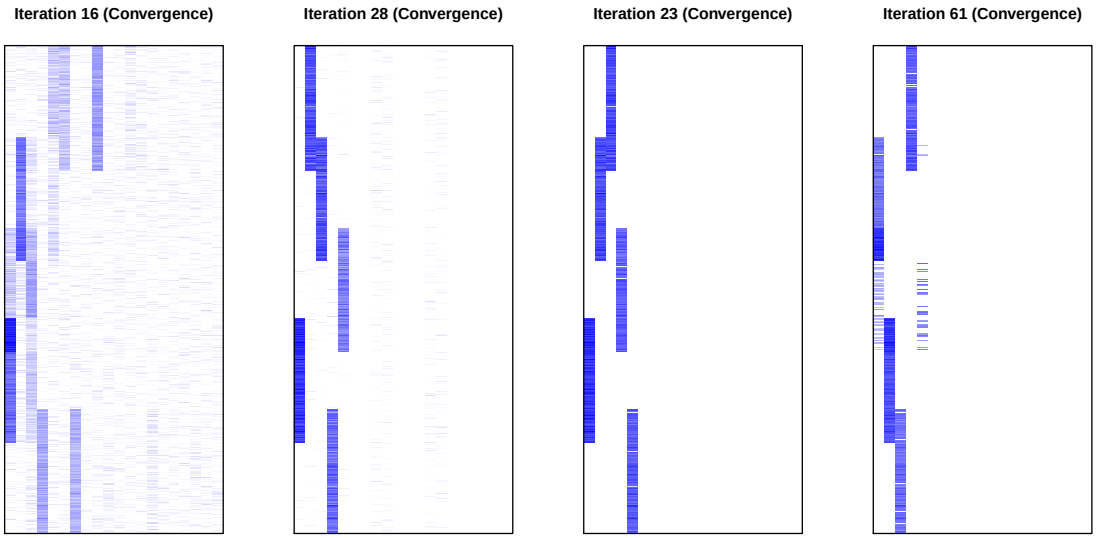


Figure 3: A trajectory of the PXL-EM algorithm, convergence achieved after 23 iterations



(a) $\lambda_0 = 5$ (b) $\lambda_0 = 10$ (c) $\lambda_0 = 20$ (d) $\lambda_0 = 30$

Figure 4: Recovered loading matrices of PXL-EM for different values of λ_0. The computations are independent, each initialized at the same $\boldsymbol{B}^{(0)}$ used in the previous section.

induced by the SSL prior on $\boldsymbol{\beta}$ will be unimodal as long λ_0 and λ_1 are not too different.

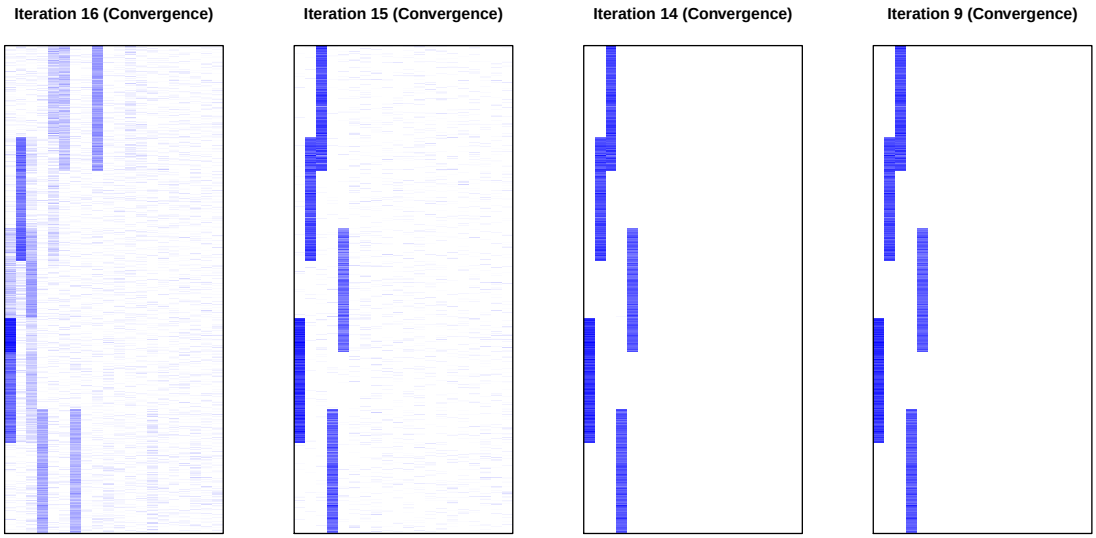
Theorem 5.1. *Assume $p < n$ and denote by λ_{min} the minimal eigen-value of $n^{-1}\boldsymbol{X}'\boldsymbol{X}$. Then for a given $\theta \in (0,1)$ the posterior distribution arising from a θ-weighted spike-and-slab LASSO mixture prior with penalties λ_0 and λ_1 has a unique mode, whenever*

$$(\lambda_0 - \lambda_1)^2 \leq 4\lambda_{min}. \tag{5.1}$$

Proof. Shown by Rockova and George (2014b).

Remark 5.1. *Under the PXL-EM regime, the estimated features $\widetilde{\boldsymbol{\Omega}}$ are nearly orthogonal as the algorithm converges. This is the most ideal scenario in the light of this theorem, where $\lambda_{min}(\widetilde{\boldsymbol{\Omega}}) \approx 1$.*

Priors with large differences $(\lambda_0 - \lambda_1)$, however, induce posteriors with many isolated sharp spikes, a difficult environment for the EM trajectories to move around. The effect of this phenomenon is illustrated in Figure 4, where the PXL-EM algorithm was run for a series of spike penalties $\lambda_0 \in I = \{5, 10, 20, 30\}$ using the same initialization and tuning as in the previous section. Clearly, larger penalties λ_0 are needed to shrink the redundant coefficients to zero. However, as λ_0 approaches the methodological ideal ($\lambda_0 \to \infty$), it becomes increasingly more difficult to find the optimal solution.



(a) $\lambda_0 = 5$ (b) $\lambda_0 = 10$ (c) $\lambda_0 = 20$ (d) $\lambda_0 = 30$

Figure 5: Recovered loading matrices of PXL-EM for different values of λ_0. The first computation ($\lambda_0 = 5$) initialized at $\boldsymbol{\beta}^{(0)}$ from the previous section, then reinitialized sequentially.

In order to facilitate the search for the global maximum in the multimodal landscape, we build on ideas from the deterministic annealing EM algorithm proposed by Ueda and Nakano (1998). The idea there is to sequentially initialize the EM calculation at optima of a series of modified posteriors, each raised to a power $1/t$, an inverse temperature parameter. By starting with a high temperature $t_1 >> 1$, where the tempered posterior is nearly unimodal, and continuing along a temperature ladder $t_1 > t_2 > \cdots > t_T = 1$ by sequentially reinitializing computations at t_i with the solutions obtained at t_{i-1}, this algorithm is apt to be more successful in finding higher posterior modes. This approach was successfully implemented by Rockova and George (2014a) in the EMVS method for variable selection in linear regression and by Yoshida and West (2010) in the variational Bayesian approach to graphical factor model. Here, we pursue a different strategy, sharing conceptual similarities with deterministic annealing, to mitigate the multimodality associated with the variable selection priors.

We will regard λ_0 as an analogue of the inverse temperature parameter, where small values (in the range suggested by Theorem 5.1) yield unimodal posteriors (in linear regression). The solutions obtained for these less interesting parameter pairs (λ_0, λ_1) can be used as warm starts for more interesting choices (λ_0, λ_1), where $\lambda_0 >> \lambda_1$. By keeping the slab variance steady and gradually increasing the spike variance λ_0 over a ladder of values $\lambda_0 \in I = \{\lambda_0^1 < \lambda_0^2 < \cdots < \lambda_0^L\}$, we perform

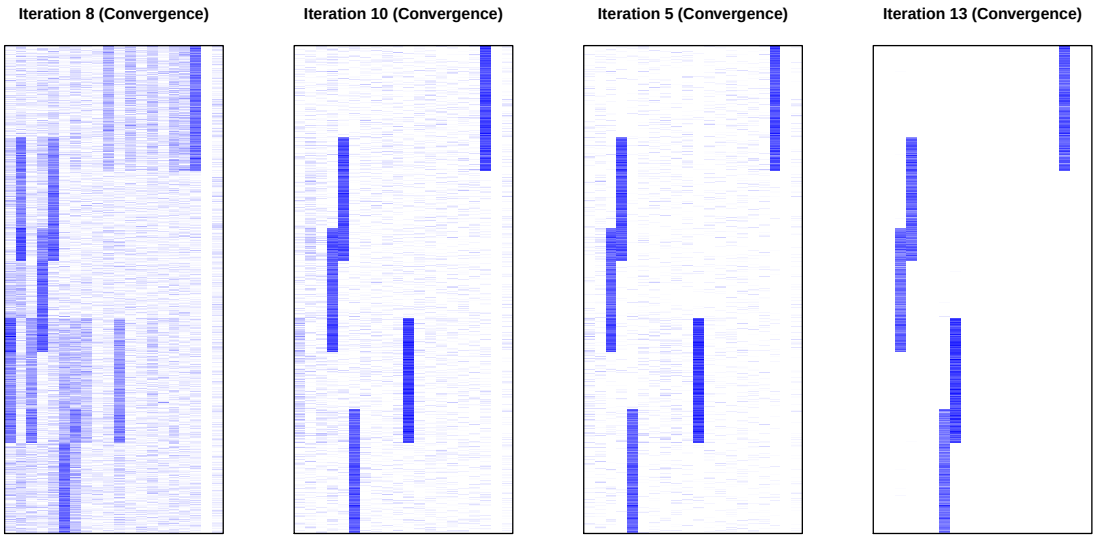
a "dynamic posterior exploration", sequentially reinitializing the calculations along the solution path. Accelerated dynamic posterior exploration is obtained by reinitializing only the loading matrix $\boldsymbol{B}$, using the same $\boldsymbol{\Sigma}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ as initial values throughout the solution path. This strategy was applied on our example with $\lambda_0 \in I = \{5, 10, 20, 30\}$ (Figure 5). Compared to Figure 4, fewer iterations are needed while factor splitting is reduced at $\lambda_0 = 30$, evidently a result of the reinitialization at the mode with $\lambda_0 = 20$. The solution stabilizes after a certain value λ_0, where further increase of λ_0 did not impact the solution. Thus, the obtained solution for sufficiently large λ_0, if a global maximum, can be regarded as an approximation to the MAP estimator under the point-mass prior.

Finally, we explored what would happen if we instead used the one-component Laplace prior obtained by $\lambda_0 = \lambda_1$. With the absence of a slab distribution, this would correspond to a LASSO-like solution path, but with automatic transformations due to the PXL-EM algorithm. We performed the PXL-EM computation with sequential initialization for the one-component Laplace prior ($\lambda_0 = \lambda_1$) assuming $\lambda_0 \in I = \{0.1, 5, 10, 20\}$ (Figure 6). In contrast to the SSL implementation above with $\lambda_0 >> \lambda_1$, it was necessary to begin with $\lambda_0 = \lambda_1 = 0.1$. In terms of identifying the nonzero loadings, PXL-EM with the one-component Laplace prior did reasonably well, generating 45 false positives in the best case where $\lambda_0 = \lambda_1 = 20$. On this example, the PXL-EM implementation of a l_1-penalized likelihood method dramatically enhances the sparsity recovery over existing sparse PCA techniques, which do not alter the factor orientation throughout the computation (Witten and Hastie, 2009). However, the estimates of the non-zero loadings were quite poor as is evidenced by Figure 7, which compares the estimated entries in the marginal covariance matrix obtained with the one-component Laplace and the SSL priors. Whereas the SSL prior achieves great accuracy in both recovery and estimation, the one-component Laplace prior must sacrifice unbiasedness to improve recovery.

6 Factor Mode Evaluation

The PXL-EM algorithm in concert with dynamic posterior exploration rapidly elicits a sequence of loading matrices $\{\widehat{\boldsymbol{B}}_{\lambda_0} : \lambda_0 \in I\}$ of varying factor cardinality and sparsity. Each such $\widehat{\boldsymbol{B}}_{\lambda_0}$ yields an estimate $\widehat{\boldsymbol{\Gamma}}_{\lambda_0}$ of the feature allocation matrix $\boldsymbol{\Gamma}$, where $\widehat{\gamma}_{ij}^{\lambda_0} = \mathbb{I}(\widehat{\beta}_{ij}^{\lambda_0} \neq 0)$. The matrix $\boldsymbol{\Gamma}$ can be regarded as a set of constraints imposed on the factor model, restricting the placement of nonzero values, both in $\boldsymbol{B}$ and $\boldsymbol{\Lambda} = \boldsymbol{B}\boldsymbol{B}' + \boldsymbol{\Sigma}$. Each $\widehat{\boldsymbol{\Gamma}}_{\lambda_0}$ provides an estimate of the effective factor dimension, the number of free parameters and the allocation of response-factor couplings. Assuming $\boldsymbol{\Gamma}$ is left-ordered (i.e. the columns sorted by their binary numbers) to guarantee uniqueness, $\boldsymbol{\Gamma}$ can be thought of as a "model" index, although not a model per se.

For purpose of comparison and selection from $\{\widehat{\boldsymbol{\Gamma}}_{\lambda_0} : \lambda_0 \in I\}$, a natural and appealing criterion is



(a) $\lambda_0 = \lambda_1 = 0.1$ (b) $\lambda_0 = \lambda_1 = 5$ (c) $\lambda_0 = \lambda_1 = 10$ (d) $\lambda_0 = \lambda_1 = 20$

Figure 6: PXL-EM algorithm with sequential reinitialization along the solution path with one-component Laplace prior.

the posterior model probability

$$\pi(\boldsymbol{\Gamma} \mid \boldsymbol{Y}) \propto \pi(\boldsymbol{Y} \mid \boldsymbol{\Gamma})\pi(\boldsymbol{\Gamma}). \tag{6.1}$$

Whereas the continuous relaxation $\Pi_{\lambda_0}(\boldsymbol{B})$ was useful for model exploration, the point-mass mixture prior $\Pi_{\infty}(\boldsymbol{B})$ will be of interest for model evaluation. Unfortunately, computing the marginal likelihood $\pi(\boldsymbol{Y} \mid \boldsymbol{\Gamma})$ under these priors is hampered because tractable closed forms are unavailable and Monte Carlo integration would be impractical. Instead, we replace $\pi(\boldsymbol{Y} \mid \boldsymbol{\Gamma})$ in (6.1) by a surrogate function, which can be interpreted as an integral lower bound to the marginal likelihood (Minka, 2001).

Schematically, the lower bound integration is as follows (Minka, 2001). We begin with the particular integral form for the marginal likelihood

$$\pi(\boldsymbol{Y} \mid \boldsymbol{\Gamma}) = \int_{\boldsymbol{\Omega}} \pi(\boldsymbol{Y}, \boldsymbol{\Omega} \mid \boldsymbol{\Gamma}) \mathrm{d}\, \pi(\boldsymbol{\Omega}), \tag{6.2}$$

for which analytical evaluation is intractable. We proceed to find an approximation to (6.2) by lower-bounding the integrand

$$\pi(\boldsymbol{Y}, \boldsymbol{\Omega} \mid \boldsymbol{\Gamma}) \geq g_{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}, \boldsymbol{\phi}), \forall(\boldsymbol{\Omega}, \boldsymbol{\phi}), \tag{6.3}$$

so that $G_{\boldsymbol{\Gamma}}(\boldsymbol{\phi}) = \int_{\boldsymbol{\Omega}} g_{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}, \boldsymbol{\phi}) \mathrm{d}\, \boldsymbol{\Omega}$ is easily integrable. The function $G_{\boldsymbol{\Gamma}}(\boldsymbol{\phi}) \leq \pi(\boldsymbol{Y} \mid \boldsymbol{\Gamma})$ then constitutes a lower bound to the marginal likelihood for any $\boldsymbol{\phi}$. The problem of integration is thus transformed

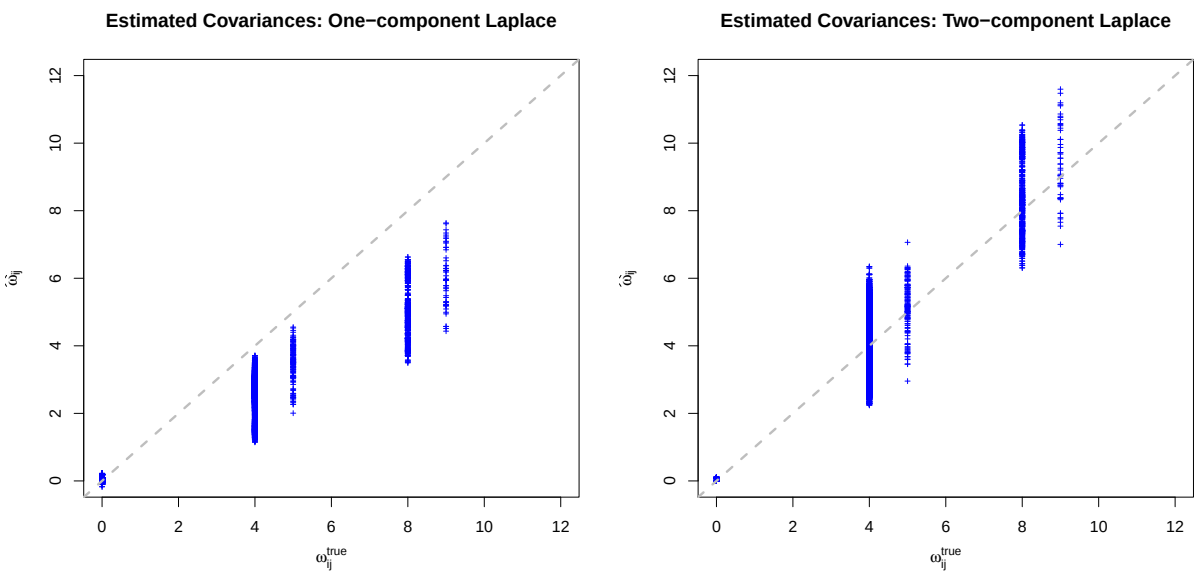


Figure 7: Estimated covariances: LASSO prior vs spike-and-slab LASSO prior

into a problem of optimization, where we search for $\widehat{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi}} G_{\boldsymbol{\Gamma}}(\boldsymbol{\phi})$ to obtain the tightest bound.

In our context, a lower bound that satisfies (6.3) is

$$g_{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}, \boldsymbol{\phi}) = C\, \pi(\boldsymbol{Y}, \boldsymbol{\Omega}, \boldsymbol{\phi} \mid \boldsymbol{\Gamma}), \quad \text{where} \quad \boldsymbol{\phi} = (\boldsymbol{B}, \boldsymbol{\Sigma}), \tag{6.4}$$

and $C = 1/\max_{\boldsymbol{\phi},\boldsymbol{\Omega}}[\pi(\boldsymbol{B}, \boldsymbol{\Sigma} \mid \boldsymbol{Y}, \boldsymbol{\Omega}, \boldsymbol{\Gamma})]$. This yields the closed form integral bound

$$G_{\boldsymbol{\Gamma}}(\boldsymbol{\phi}) = C\, \pi(\boldsymbol{B} \mid \boldsymbol{\Gamma})\pi(\boldsymbol{\Sigma})(2\pi)^{-nG/2}|\boldsymbol{\Psi}|^{n/2} \exp(-0.5 \sum_{i=1}^{n} \text{tr}(\boldsymbol{\Psi} \boldsymbol{y}_i \boldsymbol{y}_i'), \tag{6.5}$$

where $\boldsymbol{\Psi} = (\boldsymbol{B}\boldsymbol{B}' + \boldsymbol{\Sigma})^{-1}$.

By treating $G_{\boldsymbol{\Gamma}}(\boldsymbol{\phi})$ as the "complete-data" likelihood, finding $\widehat{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi}} \int_{\boldsymbol{\Omega}} \pi(\boldsymbol{Y}, \boldsymbol{\Omega}, \boldsymbol{\phi} \mid \boldsymbol{\Gamma}) \mathrm{d}\, \boldsymbol{\Omega}$ can be carried out with an EM algorithm. In particular, we can directly use the EM steps derived in Section 3, but now with $\boldsymbol{\Gamma}$ no longer treated as missing. As would be done in a confirmatory factor analysis, the calculations are now conditional on the particular $\boldsymbol{\Gamma}$ of interest. The EM calculations here are also in principle performed assuming $\lambda_0 = \infty$. As a practical matter, this will be equivalent to setting λ_0 equal to a very large number ($\lambda_0 = 1\,000$ in our examples). Thus, our EM procedure has two regimens: (a) exploration regime assuming $\lambda_0 < \infty$ and treating $\boldsymbol{\Gamma}$ as missing to find $\widehat{\boldsymbol{\Gamma}}$, (b) evaluation regime assuming $\lambda_0 \to \infty$ and fixing $\boldsymbol{\Gamma} = \widehat{\boldsymbol{\Gamma}}$. The evaluation regime can be initialized at the output values $(\widehat{\boldsymbol{B}}_{\lambda_0}, \widehat{\boldsymbol{\Sigma}}_{\lambda_0}, \widehat{\boldsymbol{\theta}}_{\lambda_0})$ from the exploratory run.

The surrogate criterion (6.5) is fundamentally the height of the posterior mode $\pi(\widehat{\boldsymbol{\phi}} \mid \boldsymbol{Y}, \boldsymbol{\Gamma})$ under the point-mass prior $\Pi_\infty(\boldsymbol{B})$, assuming $\boldsymbol{\Gamma} = \widehat{\boldsymbol{\Gamma}}$. Despite being a rather crude approximation to the

Exploratory PXL-EM Regime						Evaluation PXL-EM Regime		
Figure 4: SSL prior								
λ_0	FDR	FNR	$\sum_{jk}\widehat{\gamma}_{jk}$	$\widehat{K^+}$	Recovery Error	λ_0	Recovery Error	$\widetilde{G}(\widehat{\mathbf{\Gamma}})$
5	0.693	0	24150	20	459.209	1 000	410.333	-464171.3
10	0.616	0	10933	20	354.428	1 000	330.263	-367567.8
20	0.001	0.001	2500	5	290.074	1 000	255.104	-300765.3
30	0.169	0.102	2483	11	533.475	1 000	533.234	-310708.6
Figure 5: SSL prior								
λ_0	FDR	FNR	$\sum_{jk}\widehat{\gamma}_{jk}$	$\widehat{K^+}$	Recovery Error	λ_0	Recovery Error	$\widetilde{G}(\widehat{\mathbf{\Gamma}})$
5	0.693	0	24150	20	459.209	1 000	410.333	-464171.3
10	0.629	0	11563	20	326.355	1 000	332.73	-372386.7
20	0.003	0.001	2502	5	256.417	1 000	255.054	-300774.0
30	0	0.002	2498	5	256.606	1 000	256.061	-300771.4
Figure 6: LASSO prior								
λ_0	FDR	FNR	$\sum_{jk}\widehat{\gamma}_{jk}$	$\widehat{K^+}$	Recovery Error	λ_0	Recovery Error	$\widetilde{G}(\widehat{\mathbf{\Gamma}})$
0.1	0.693	0	36879	19	409.983	1 000	420.32	-536836.2
5	0.692	0	21873	19	365.805	1 000	398.78	-447489.0
10	0.64	0	11657	19	570.104	1 000	315.316	-373339.2
20	0.024	0	2533	5	892.244	1 000	233.419	-300933.3

Table 1: Table summarizes the quality of the reconstruction of the marginal covariance matrix $\mathbf{\Lambda}$, namely (a) FDR, (b) FNR, (c) the estimated number of nonzero loadings, (d) the estimated effective factor cardinality, (d) the Frobenius norm $d_F(\widehat{\mathbf{\Lambda}}, \mathbf{\Lambda}_0)$ (Recovery Error). The recovery error is computed twice, once after the exploratory run of the PXL-EM algorithm ($\lambda_0 < \infty$) and after the evaluation run ($\lambda_0 \approx \infty$). The evaluation is run **always** with the SSL prior. For the exploration, we used both the SSL prior as well as the LASSO (one-component Laplace) prior.

posterior model probability, the function

$$\widetilde{G}(\mathbf{\Gamma}) = G_{\mathbf{\Gamma}}(\widehat{\phi})\pi(\mathbf{\Gamma}), \tag{6.6}$$

is a practical criterion that can discriminate well between candidate models.

Using the sequential initialization from previous section, we hope to improve upon this criterion at every step, until some sufficiently large value λ_0 which is practically indistinguishable from the limiting case. The stabilization of this criterion will be then indication, that no further increase in λ_0 is needed. We now illustrate the use of this criterion on our models from from Section 5.

We computed the $\widetilde{G}(\mathbf{\Gamma})$ criterion for all the models discovered with the PXL-EM algorithm, both using independent initialization (Figure 4) and sequential initialization (Figure 5 and Figure 6). We

also evaluated the quality of the reconstructed marginal covariance matrix, namely the (a) the proportion of falsely identified nonzero values (FDR), (b) the proportion of falsely non-identified nonzeroes (FNR), (c) the estimated number of nonzero loadings, (d) the estimated effective factor cardinality, (d) the Frobenius norm $d_F(\widehat{\boldsymbol{\Lambda}}, \boldsymbol{\Lambda}_0)$ (Recovery Error). The recovery error is computed twice, once after the exploratory run of the PXL-EM algorithm ($\lambda_0 < \infty$) and then after the evaluation run of the EM algorithm ($\lambda_0 \approx \infty$). The exploration was performed both using the SSL prior (Figures 4 and 5) as well as the one-component Laplace (LASSO) prior (Figure 6). The evaluation is run *always* with the SSL prior.

The results indicate that the criterion $\widetilde{G}(\boldsymbol{\Gamma})$ is higher for models with fewer false negative/positive discoveries and effectively discriminates the models with the best reconstruction properties. It is worth noting that the output from the exploratory run is greatly refined with the point-mass SSL prior ($\lambda_0 \approx \infty$), reducing the reconstruction error. This is particularly evident for the one-component LASSO prior, which achieves good reconstruction properties (estimating the pattern of sparsity) for larger penalty values, at the expense of sacrificing the recovery of the coefficients (Figure 7).

Importantly, the sequential initialization of SSL (Figure 5) results in the stabilization of the estimated loading pattern and thereby the criterion (6.6). This is an indication that further increase in λ_0 may not dramatically change the result and that the output obtained with such large λ_0 is ready for interpretation.

Lastly, to see how our approach would fare in the presence of no signal, a similar simulated experiment was conducted with $\boldsymbol{B}_{true} = \mathbf{0}_{G \times K^\star}$. The randomly initiated dynamic posterior exploration soon yielded the null model $\widehat{\boldsymbol{B}} = \boldsymbol{B}_{true}$, where the criterion $\widetilde{G}(\boldsymbol{\Gamma})$ was also the highest. Our approach did not find a signal where there was none.

7 Kendall's Applicant Data

We first illustrate our method on a familiar dataset analysed previously by multiple authors including (Kendall, 1975; Rowe, 2003; Frühwirth-Schnatter and Lopes, 2009). The data consists of scores on a ten-point scale involving 15 characteristics of $n = 48$ applicants for a certain job. Kendall extracts four factors on the basis of a principal component analysis with 4 eigenvalues greater than one, accounting for 81.5% of explained variance.

After centering the scores, we run our PXL-EM algorithm assuming $K^\star = 10$, $\lambda_1 = 0.001, \alpha = 1/15$. For factor model exploration with sequential reinitialization, we consider a random starting point $\boldsymbol{B}^{(0)}$ (standard Gaussian entries) and a tempering schedule $\lambda_0 \in I = \{1, 2, \ldots, 50\}$. We used a convergence margin $\varepsilon = 0.01$. We report a model obtained with $\lambda_0 = 50$, confirmed by the highest value of (6.6), which yielded $\widehat{K^+} = 6$ factors. The associated loading matrix together with estimated residual

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	$\widehat{\sigma}_j^2$
Application	0.88	**-1.29**	0.35	-0.71	**-1.94**	0	0.17
Appearance	0	0	0	0	0	0	1.93
Academic ability	0	0	0	0	0	0	1.95
Likability	**1.40**	0	**2.35**	0	0	0	0.20
Self-confidence	**2.03**	0	0	0	0	0	1.20
Lucidity	**2.82**	0	0	0	0	0	1.25
Honesty	0.94	0.63	**1.40**	**1.70**	0	0	2.49
Salesmanship	**3.13**	0	0	0	0	0	1.28
Experience	0.87	**-2.17**	0	-0.34	-0.50	**-2.17**	0.16
Drive	**2.51**	0	0	0	0	0	1.44
Ambition	**2.61**	0	0	0	0	0	1.18
Grasp	**2.72**	0	0	0	0	0	1.20
Potential	**2.79**	0	0	0	0	0	1.41
Keenness	**1.67**	0	0	0	0	0	0.19
Suitability	**1.81**	**-2.68**	0	0	0	0	0.17

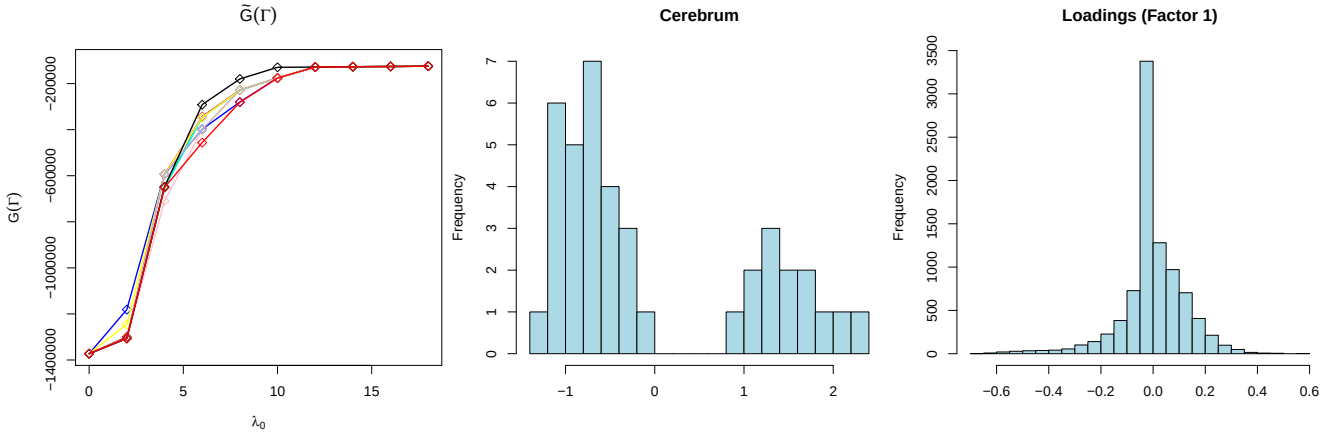
Table 2: Kendall's data: estimated loading matrix and residual variance parameters, in bold are loading estimates that are greater than one in absolute value

variances is displayed in Table 2.

The loading matrix can be naturally interpreted as follows. Factor 1 is a "success precursor", involving abilities such as self-confidence, drive, ambition, and lucidity. A similar factor was found also by Frühwirth-Schnatter and Lopes (2009) (Factor 3) and by Kendall (1975) (Factor 1). The Factor 2 can be interpreted as experience (Factor 2 of Kendall (1975) and Factor 1 of Frühwirth-Schnatter and Lopes (2009)). The Factor 3 can be interpreted as a general likability of a person (Factor 3 in Kendall (1975)).

8 The AGEMAP Data

We further illustrate our approach on a high-dimensional dataset extracted from AGEMAP (Atlas of Gene Expression in Mouse Aging Project) database of Zahn and et al. (2007), which catalogs age-related changes in gene expression in mice. Included in the experiment were mice of ages 1, 6, 16, and 24 months, with ten mice per age cohort (five mice of each sex). For each of these 40 mice, 16 tissues were dissected and tissue-specific microarrays were prepared. From each microarray, values from 8 932 probes were obtained. The collection of standardized measurements is available online `http://cmgm.stanford.edu/~kimlab/aging_mouse/`. Factor analysis in genomic studies provides



(a) Evolution of $\widetilde{G}(\widehat{\mathbf{\Gamma}})$ (b) The new latent feature (c) The factor loadings

Figure 8: (Left) Dynamic posterior exploration, evolution of the $\widetilde{G}(\cdot)$ function, one line for each of the 10 initializations; (Middle) Histogram of the newly created feature; (Right) Histogram of the factor loadings of the new factor

an opportunity to look for groups of functionally related genes, whose expression may be affected by shared hidden causes. In this analysis we will also focus on the ability to featurize the underlying hidden variables. The success of the featurization is also tied to the orientation of the factor model.

The AGEMAP dataset was analyzed previously by Perry and Owen (2010), who verified the existence of some apparent latent structures using rotation tests. Here we will focus only on one tissue, cerebrum, which exhibited strong evidence for the presence of a binary latent variable, confirmed by a rotation test Perry and Owen (2010). We will first deploy a series of linear regressions, regressing out the effect of an intercept, sex and age on each of the 8 932 responses. Taking the residuals from these regressions as new outcomes, we proceed to apply our infinite factor model, hoping to recover the hidden binary variable.

We assume that there are at most $K^\star = 20$ latent factors and run our PXL-EM algorithm with the SSL prior and $\lambda_1 = 0.001, \alpha = 1/G$. For factor model exploration, we deploy dynamic posterior exploration, i.e. sequential reinitialization of the loading matrix along the solution path. The solution path will be evaluated along the following tempering schedule $\lambda_0 \in \{\lambda_1 + k \times 2 : 0 \leq k \leq 9\}$, initiated at the trivial case $\lambda_0 = \lambda_1$ to mitigate the multimodality associated with the variable selection priors (as explained in Section 5). To investigate the sensitivity to initialization of the dynamic posterior exploration, we consider 10 random starting matrices used to initialize the first computation with $\lambda_1 = \lambda_0$. These matrices were each generated element-wise from a standard Gaussian distribution. We will use $\mathbf{\Sigma}^{(0)} = \mathrm{I}_G$, $\boldsymbol{\theta}^{(0)} = (0.5, \ldots, 0.5)'$ as starting values for every λ_0 to accelerate the exploration.

The margin $\varepsilon = 0.01$ is used to claim convergence.

The results of dynamic posterior exploration are summarized in Table 3. The table reports the estimated factor dimension $\widehat{K^+}$ (i.e. the number of factors with at least one nonzero estimated loading), estimated number of nonzero factor loadings $\sum_{jk} \widehat{\gamma}_{jk}$ and the value of the surrogate criterion $\widetilde{G}(\widehat{\mathbf{\Gamma}})$. The evolution of $\widetilde{G}(\widehat{\mathbf{\Gamma}})$ along the solution path is also depicted on Figure and shows a remarkably similar pattern, despite the very arbitrary initializations. From both Table 3 and Figure 8(a) we observe that the estimation has stabilized after $\lambda_0 = 12.001$, yielding factor models of effective dimension $\widehat{K^+} = 1$ with a similar number of nonzero factor loadings (all nonzero factor loadings are associated with just one factor). Such stabilization is an indication that further increase in λ_0 will not affect very much the solution. Based on this analysis, we would select just one factor. Using the output obtained at $\lambda_0 = 18.001$ from 6^{th} initiation (the highest value $\widetilde{G}(\widehat{\mathbf{\Gamma}})$), we investigate the pattern of the estimated latent feature (histogram on Figure 8(b)). We discover a strikingly dichotomous pattern across the 40 mice, suggesting the presence of an underlying binary latent variable. A similar histogram was reported also by Perry and Owen (2010). Their finding was also strongly supported by a statistical test.

The data representation found by PXL-EM is sparse in terms of the number of factors, suggesting the presence of a single latent variable. The representation is however not sparse in terms of factor loadings, where the factor is loaded on the majority of considered genes (78%). The mechanism underlying the binary latent variable thus cannot be attributed to only a few responsible genes. The histogram of estimates loadings associated with this factor (Figure 8(c)) suggests the presence of a few very active genes that could potentially be interpreted as leading genes for the factor.

The concise representation using just one latent factor could not obtained using, for instance, sparse principal components that do not perform the rotation and thereby smear the signal across multiple factors when the factor dimension is overfitted.

9 Discussion

We have presented a new paradigm for the discovery of interpretable latent factor models through automatic rotations to sparsity. Rotational transformations have played a key role in enhancing the interpretability of principal components via post-hoc modifications of the model orientation. Sparse principal components have partially avoided the need for such transformations by penalizing non-sparse orientations. Here we have combined the benefits of the two perspectives within an integrated procedure. The new and crucial aspect here is that the modifications of the factor basis are performed throughout the computation rather than after the model estimates have been obtained. These automatic transformations provide an opportunity to find the right coordinate system for the latent

λ_0	$\widetilde{G}(\widehat{\mathbf{\Gamma}})$	$\widehat{K^+}$	$\sum_{jk}\widehat{\gamma}_{jk}$	$\widetilde{G}(\widehat{\mathbf{\Gamma}})$	$\widehat{K^+}$	$\sum_{jk}\widehat{\gamma}_{jk}$	$\widetilde{G}(\widehat{\mathbf{\Gamma}})$	$\widehat{K^+}$	$\sum_{jk}\widehat{\gamma}_{jk}$	$\widetilde{G}(\widehat{\mathbf{\Gamma}})$	$\widehat{K^+}$	$\sum_{jk}\widehat{\gamma}_{jk}$	$\widetilde{G}(\widehat{\mathbf{\Gamma}})$	$\widehat{K^+}$	$\sum_{jk}\widehat{\gamma}_{jk}$
	Init 1			Init 2			Init 3			Init 4			Init 5		
0.001	-1372666.6	20	178626	-1372640.9	20	178623	-1372662.2	20	178624	-415766.0	20	178624	-1372657.5	20	178630
2.001	-1306835.9	20	163282	-1299649.4	20	161958	-1303685.7	20	162686	-175881.2.0	18	147234	-1304573.4	20	162908
4.001	-709828.9	11	85056	-649129.3	10	77327	-649686.7	10	77438	-126214.0	9	70548	-653144.7	10	77992
6.001	-401190.8	6	44856	-343720.6	5	37216	-345125.0	5	37437	-123184.9	6	44209	-346137.2	5	37596
8.001	-228202.7	3	21534	-228247.6	3	21541	-228293.2	3	21548	-120121.2	4	28510	-228720.5	3	21615
10.001	-175878.9	2	14310	-175862.5	2	14307	-175868.7	2	14308	-116653.7	2	14307	-175925.4	2	14319
12.001	-128111.9	1	7595	-128094.7	1	7592	-128111.9	1	7595	-112851.9	1	7602	-128100.5	1	7593
14.001	-126894.2	1	7380	-126567.9	1	7323	-126894.2	1	7380	-109596.9	1	7367	-126894.2	1	7380
16.001	-125623.1	1	7159	-125746.2	1	7180	-125746.3	1	7180	-106815.9	1	7145	-125722.7	1	7176
18.001	-124481.5	1	6961	-124398.2	1	6947	-124368.2	1	6942	-104005.2	1	6953	-124428.1	1	6952
	Init 6			Init 7			Init 8			Init 9			Init 10		
0.001	-1372649.5	20	178627	-1372674.4	20	178634	-1372668.6	20	178629	-1372722.8	20	178637	-1372751.8	20	178633
2.001	-1305285.1	20	163025	-1245378.2	19	155526	-1305644.2	20	163048	-1306693.6	20	163278	-1306569.3	20	163219
4.001	-592586	9	70291	-593542.5	9	70434	-592662.1	9	70294	-648525.8	10	77225	-651142.3	10	77673
6.001	-343951	5	37254	-346380.2	5	37630	-398551.3	6	44375	-291575.3	4	30435	-456168.5	7	51886
8.001	-228280.2	3	21546	-228485.1	3	21578	-228250.4	3	21541	-179853.2	2	14979	-281223.4	4	28721
10.001	-175891.6	2	14312	-175931.8	2	14319	-175885.2	2	14311	-129360.1	1	7819	-175872.2	2	14309
12.001	-128111.9	1	7595	-128094.7	1	7592	-128111.9	1	7595	-128094.7	1	7592	-128111.9	1	7595
14.001	-126882.6	1	7378	-126894.2	1	7380	-126870.8	1	7376	-126876.7	1	7377	-126876.7	1	7377
16.001	-125623.1	1	7159	-125576.8	1	7151	-125743.6	1	7181	-125758.2	1	7182	-125743.6	1	7181
18.001	-124326.5	1	6935	-124326.5	1	6935	-124404.3	1	6948	-124451.9	1	6956	-124422.2	1	6951

Table 3: Evolutions of the $\widetilde{G}(\widehat{\mathbf{\Gamma}})$ function together with estimated factor dimension $\widehat{K^+}$ and estimated number of model parameters $\sum_{jk}\widehat{\gamma}_{jk}$ in dynamic posterior exploration using 10 random initializations.

features that admits the sparse representation. By incorporating the rotational aspect of the factor model we greatly enhance the search for sparse representations. Here, we introduce the rotation parameter indirectly, through parameter expansion in our PXL-EM algorithm, which iterates between soft-thresholding and transformations of the factor basis.

Although we have proposed the PXL-EM algorithm in conjunction with the new spike-and-slab LASSO prior and Indian Buffet Process, it can also be used to improve other methods, such as penalized-likelihood variants of probabilistic principal components. By modifying the factor orientation throughout the computation, PXL-EM increases the sparsity recovery potential of such methods, as seen in our simulated example.

Our approach does not require any pre-specification of the factor dimension and any identifiability constraints. Deployed with the fast PXL-EM algorithm, the proposed methodology is suitable for high dimensional data, eliminating the need for posterior simulation. Finally, our methodology can be further extended to canonical correlation analysis or to latent factor augmentations of multivariate regression (Rockova and Lesaffre, 2013; Bargi et al., 2014).

10 Appendix

10.1 Proof of Theorem 2.2

Proof. We will prove the theorem by showing that random variables $M_j = \sum_{k=1}^{\infty} \beta_{jk}^2$ have finite expectations. We can write $\mathsf{E}(M_j) = \sum_{k=1}^{\infty} \mathsf{E}[\mathsf{E}(\beta_{jk}^2 \mid \theta_{(k)})]$, where

$$\mathsf{E}(\beta_{jk}^2 \mid \theta_{(k)}) = \frac{2}{\lambda_1^2}\theta_{(k)} + \frac{2}{\lambda_{0k}^2}(1-\theta_{(k)}) = 2\left(\frac{1}{\lambda_1^2} - \frac{1}{\lambda_{0k}^2}\right)\theta_{(k)} + \frac{2}{\lambda_{0k}^2}.$$

From the stick-breaking construction we obtain

$$\mathsf{E}[\theta_{(k)}] = \prod_{l=1}^{k} \mathsf{E}\,\nu_l = \left(\frac{\alpha}{\alpha+1}\right)^k.$$

The sum $\sum_{k=1}^{\infty}\left(\frac{2}{\lambda_1^2} + \frac{2}{\lambda_{0k}^2}\right)\left(\frac{\alpha}{\alpha+1}\right)^k$ is dominated by a convergent geometric series and hence converges. The finiteness of $\sum_{k=1}^{\infty}\frac{2}{\lambda_{0k}^2}$ follows from the assumption that $1/\lambda_{0k}$ goes to zero at least as fast as $1/k$. □

10.2 Proof of Theorem 2.3

Proof. By definition, $d_\infty(\mathbf{\Lambda}, \mathbf{\Lambda}^{K^\star}) = \max_{1\le j,m\le G} |a_{jm}^{K^\star}|$, where $a_{jm}^{K^\star} = \sum_{k=K^\star+1}^{\infty} \beta_{jk}\beta_{mk}$. By the Cauchy-Schwartz inequality, we have $d_\infty(\mathbf{\Lambda}, \mathbf{\Lambda}^{K^\star}) = \max_{1\le j\le G} a_{jj}^{K^\star}$. By the Jensen and Chebyshev inequalities, we obtain

$$\mathsf{P}(d_\infty(\mathbf{\Lambda}, \mathbf{\Lambda}^{K^\star}) \le \varepsilon) \ge \left(1 - \frac{\mathsf{E}[a_{11}]}{\varepsilon}\right)^G, \tag{10.1}$$

where

$$\mathsf{E}[a_{11}] = \mathsf{E}\left(\sum_{k=K^\star+1}^{\infty} \mathsf{E}[\beta_{jk}^2 \mid \theta_{(k)}]\right) = 2\left[\frac{1}{\lambda_1^2}\frac{\mu^{K^\star+1}}{(1-\mu)} - \frac{(a\mu)^{K^\star+1}}{(1-a\mu)} + \frac{a^{K^\star+1}}{(1-a)}\right]. \tag{10.2}$$

We will now find a lower bound on $K^\star$, so that $\left(1 - \frac{\mathsf{E}[a_{11}]}{\varepsilon}\right)^G > 1-\varepsilon$. Such $K^\star$ will have to satisfy $\mathsf{E}[a_{11}] < \widetilde{\varepsilon}$, where $\widetilde{\varepsilon} = \varepsilon[1-(1-\varepsilon)^{1/G}]$. This will be fulfilled by setting

$$K^\star > \max\left\{\log\left[\frac{\widetilde{\varepsilon}}{t}(1-a)\right] / \log(a); \log\left[\lambda_1^2\widetilde{\varepsilon}\left(1-\frac{1}{t}\right)(1-\mu)\right] / \log(\mu)\right\}$$

for any $0 < t < 1$. The theorem follows by setting $t = 1/2$. □

10.3 Proof of Theorem 2.5

Proof. We can write

$$\begin{aligned}\mathsf{P}(K^+ \leq k) &= \mathsf{P}(\gamma_{jl} = 0; l > k, j = 1, \ldots, G) = \mathsf{E}[\mathsf{P}(\gamma_{jl} = 0; l > k, j = 1, \ldots, G \mid \boldsymbol{\theta})] \\ &= \mathsf{E}\left(\prod_{l>k}(1-\theta_{(l)})^G\right)\end{aligned}$$

Now we use the inequality $(1 - x/2) > \exp(-x)$ if $0 < x \leq 1.5$ to get $(1-\theta_{(l)})^G > \exp(-2\,G\theta_{(l)})$ for $\theta_{(l)} < 0.75$. Denote the event $\mathcal{E} = \{\theta_{(l)} \leq 0.75 : l > k\}$. We can write

$$\begin{aligned}\mathsf{P}(K^+ \leq k) &> \mathsf{E}\left[\left(\prod_{l>k}(1-\theta_{(l)})^G\right)\mathrm{I}_{\mathcal{E}}(\boldsymbol{\theta})\right] = \mathsf{P}(\mathcal{E})\mathsf{E}\left[\left(\prod_{l>k}(1-\theta_{(l)})^G\right) \mid \boldsymbol{\theta} \in \mathcal{E}\right] \\ &> \mathsf{P}(\mathcal{E})\mathsf{E}\left[\exp\left(-2\,G\sum_{l>k}\theta_{(l)}\right) \mid \boldsymbol{\theta} \in \mathcal{E}\right] > \mathsf{P}(\mathcal{E})\exp\left(-2\,G\sum_{l>k}\mathsf{E}[\theta_{(l)} \mid \theta_{(l)} \leq 0.75]\right) \\ &\geq \mathsf{P}(\mathcal{E})\exp\left(-2\,G\sum_{l>k}\mathsf{E}[\theta_{(l)}]\right) = \mathsf{P}(\mathcal{E})\exp\left[-2\,G\sum_{l>k}\left(\frac{\alpha}{\alpha+1}\right)^l\right] \\ &= \mathsf{P}(\mathcal{E})\exp\left[-2\,G(\alpha+1)\left(\frac{\alpha}{\alpha+1}\right)^{k+1}\right]\end{aligned}$$

Because $1 > \theta_{(1)} > \theta_{(2)} > \ldots$, we have $\mathsf{P}(\mathcal{E}) = \mathsf{P}(\theta_{(k+1)} \leq 0.75)$. By Markov's inequality, we obtain $\mathsf{P}(\mathcal{E}^C) < \frac{4}{3}\left(\frac{\alpha}{\alpha+1}\right)^{k+1}$ and therefore

$$\mathsf{P}(\mathcal{E}) \geq 1 - \frac{4}{3}\left(\frac{\alpha}{\alpha+1}\right)^{k+1} > \exp\left[-\frac{8}{3}\left(\frac{\alpha}{\alpha+1}\right)^{k+1}\right].$$

The last inequality holds because $\frac{8}{3}\left(\frac{\alpha}{\alpha+1}\right)^{k+1} < 1.5$ for all $0 < \alpha \leq 1$ and $k \in \mathbb{N}$. Then we have

$$\mathsf{P}(K^+ > k) \leq 1 - \exp\left\{-2\left[G(\alpha+1)+\frac{4}{3}\right]\left(\frac{\alpha}{\alpha+1}\right)^{k+1}\right\} < 2\left[G(\alpha+1)+\frac{4}{3}\right]\left(\frac{\alpha}{\alpha+1}\right)^{k+1}. \quad \square$$

References

Bargi, A., Piccardi, M., and Ghahramani, Z. (2014), "A non-parametric conditional factor regression model for multi-dimensional input and response," in "17^{th} International Conference on Artificial Intelligence and Statistics," .

Bhattacharya, A. and Dunson, D. (2011), "Sparse Bayesian infinite factor models," *Biometrika*, 98, 291–306.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008), "High-dimensional sparse factor modelling: Applications in gene expression genomics," *Journal of the American Statistical Association*, 103, 1438–1456.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2014), "Bayesian linear regression with sparse priors," *Submitted manuscript*.

Castillo, I. and van der Vaart, A. (2012), "Needles and straw in a haystack: Posterior concentration for possibly sparse sequences," *The Annals of Statistics*, 40, 2069–2101.

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, 39, 1–38.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, 22, 1–22.

Frühwirth-Schnatter, S. and Lopes, H. (2009), *Parsimonious Bayesian factor analysis when the number of factors is unknown*, Technical report, University of Chicago Booth School of Business.

George, E. I. and McCulloch, R. E. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

Ghosh, J. and Dunson, D. (2009), "Default prior distributions and efficient posterior computation in Bayesian factor analysis," *Journal of Computational and Graphical Statistics*, 18, 306320.

Green, P. J. (1990), "On use of the EM for penalized likelihood estimation," *Journal of the Royal Statistical Society. Series B*, 52.

Griffiths, T. and Ghahramani, Z. (2005), *Infinite latent feature models and the Indian buffet process*, Technical report, Gatsby Computational Neuroscience Unit.

Griffiths, T. and Ghahramani, Z. (2011), "The Indian buffet process: An introduction and review," *Journal of Machine Learning Research*, 12, 1185–1224.

Ishwaran, H. and Rao, J. S. (2005), "Spike and slab variable selection: frequentist and Bayesian strategies," *Annals of Statistics*, 33, 730–773.

Kendall, M. (1975), *Multivariate Analysis*, London: Griffin.

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

Lewandowski, A., Liu, C., and van der Wiel, S. (1999), "Parameter expansion and efficient inference," *Statistical Science*, 25, 533–544.

Liu, C., Rubin, D., and Wu, Y. N. (1998), "Parameter expansion to accelerate EM: The PX-EM algorithm," *Biometrika*, 85, 755–0770.

Liu, J. S. and Wu, Y. N. (1999), "Parameter expansion for data augmentation," *Journal of the American Statistical Association*, 94, 1264–1274.

Lopes, H. and West, M. (2004), "Bayesian model assessment in factor analysis," *Statistica Sinica*, 14, 41–67.

Luttinen, J. and Ilin, A. (2010), "Transformations in variational Bayesian factor analysis to speed up learning," *Neurocomputing*, 73, 1093–1102.

Meng, X. L. and Rubin, D. B. (1993), "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, 80, 267–278.

Minka, T. (2001), *Using lower bounds to approximate integrals*, Technical report, Microsoft Research.

Pati, D., Bhattacharya, A., Pillai, N., and Dunson, D. (2014), "Posterior contraction in sparse Bayesian factor models for massive covariance matrices," *The Annals of Statistics*, 42, 1102–1130.

Perry, P. O. and Owen, A. B. (2010), "A rotation test to verify latent structure," *Journal of Machine Learning Research*, 11, 603–624.

Qi, Y. and Jaakkola, T. (2006), "Parameter expanded variational Bayesian methods," in "Neural Information Processing Systems," .

Rai, P. and Daumé, H. (2008), "The infinite hierarchical factor regression model," in "Neural Information Processing Systems," .

Rockova, V. and George, E. (2014a), "EMVS: The EM approach to Bayesian variable selection," *Journal of the American Statistical Association*, 109, 828–846.

Rockova, V. and George, E. (2014b), *The Spike-and-Slab LASSO*, Manuscript in preparation.

Rockova, V. and Lesaffre, E. (2013), "Bayesian sparse factor regression approach to genomic data integration," in "28^{th} International Workshop on Statistical Modelling," .

Rowe, D. (2003), *Multivariate Bayesian Statistics*, London: Chapman & Hall.

Teh, Y., Gorur, D., and Ghahramani, Z. (2007), "Stick-breaking construction for the Indian buffet process," in "11th Conference on Artificial Intelligence and Statistics," .

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and smoothness via the fused LASSO," *Journal of the Royal Statistical Society. Series B*, 67, 91–108.

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

Ueda, N. and Nakano, R. (1998), "Deterministic annealing EM algorithm," *Neural Networks*, 11, 271–282.

van Dyk, D. and Meng, X. L. (2001), "The art of data augmentation," *Journal of Computational and Graphical Statistics*, 10, 1–111.

van Dyk, D. and Meng, X. L. (2010), "Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: A graphical guide book," *Statistical Science*, 25, 429–449.

van Dyk, D. and Tang, R. (2003), "The one-step-late PXEM algorithm," *Statistics and Computing*, 13, 137–152.

West, M. (2003), "Bayesian factor regression models in the "large p, small n" paradigm," in "Bayesian Statistics," pages 723–732, Oxford University Press.

Witten, D. and Hastie, T. (2009), "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, 10, 515–534.

Wu, C. F. J. (1983), "On the convergence properties of the EM algorithm," *The Annals of Statistics*, 11, 95–103.

Yoshida, R. and West, M. (2010), "Bayesian learning in sparse graphical factor models via variational mean-field annealing," *Journal of Machine Learning Research*, 11, 1771–1798.

Zahn, J. M. and et al. (2007), "AGEMAP: A gene expression database for aging in mice," *PLOS Genetics*, 3, 2326–2337.

Zou, H. (2006), "The adaptive LASSO and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.