Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity

by

Veronika Ročková and Edward I. George [1]

Revised September 12^{th}, 2015

Abstract

Rotational post-hoc transformations have traditionally played a key role in enhancing the interpretability of factor analysis. Regularization methods also serve to achieve this goal by prioritizing sparse loading matrices. In this work, we bridge these two paradigms with a unifying Bayesian framework. Our approach deploys *intermediate* factor rotations throughout the learning process, greatly enhancing the effectiveness of sparsity inducing priors. These automatic rotations to sparsity are embedded within a PXL-EM algorithm, a Bayesian variant of parameter-expanded EM for posterior mode detection. By iterating between soft-thresholding of small factor loadings and transformations of the factor basis, we obtain (a) dramatic accelerations, (b) robustness against poor initializations and (c) better oriented sparse solutions. To avoid the pre-specification of the factor cardinality, we extend the loading matrix to have infinitely many columns with the Indian Buffet Process (IBP) prior. The factor dimensionality is learned from the posterior, which is shown to concentrate on sparse matrices. Our deployment of PXL-EM performs a dynamic posterior exploration, outputting a solution path indexed by a sequence of spike-and-slab priors. For accurate recovery of the factor loadings, we deploy the Spike-and-Slab LASSO prior, a two-component refinement of the Laplace prior (Ročková, 2015). A companion criterion, motivated as an integral lower bound, is provided to effectively select the best recovery. The potential of the proposed procedure is demonstrated on both simulated and real high-dimensional data, which would render posterior simulation impractical.

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 $n \times G$

[1] Veronika Ročková is Postdoctoral Research Associate, 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.

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

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

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, $f(\boldsymbol{y}_i \mid \boldsymbol{B}, \boldsymbol{\Sigma}) = \mathcal{N}_G(\mathbf{0}, \boldsymbol{B}\boldsymbol{B}' + \boldsymbol{\Sigma})$, $1 \leq i \leq n$, a decomposition which uses at most $G \times (K+1)$ 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 *factor dimensionality* K and the *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 necessary for prediction or estimation of the marginal covariance matrix (Bhattacharya and Dunson, 2011), non-sparse orientations diminish the potential for interpretability, our principal focus here.

Traditional approaches to obtaining interpretable loading patterns have entailed post-hoc rotations of the original solution. For instance, the varimax post-processing step (Kaiser, 1958) finds a rotation $\boldsymbol{P}$ that minimizes a complexity criterion and yields new loadings that are either very small (near zero) or large. As an alternative, regularization methods prioritize matrices with exact zeroes via a penalty/sparsity prior (Witten et al., 2009; Carvalho et al., 2008). The main thrust of this contribution is to cross-fertilize these two paradigms within one unified framework. In our approach, model rotations are embedded within the learning process, greatly enhancing the effectiveness of regularization and providing an opportunity to search for a factor basis which best supports sparsity, when it in fact exists. Going further, our approach does not require pre-specification of the factor cardinality K. As with similar factor analyzers (Knowles and Ghahramani, 2011; Bhattacharya and Dunson, 2011), here

the loading matrix $\boldsymbol{B}$ is extended to include infinitely many columns.

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 the important coefficients from the ignorable (West, 2003; Carvalho et al., 2008; Rai and Daumé, 2008; Knowles and Ghahramani, 2011; Frühwirth-Schnatter and Lopes, 2009). A particularly appealing spike-and-slab variant has been the mixture of a point mass spike and an absolutely continuous slab distribution which, unfortunately, poses serious computational challenges in high-dimensional data.

We address this challenge by developing a tractable inferential procedure that performs deterministic rather than stochastic posterior exploration. At the heart of our approach is a feasible continuous relaxation of the point-mass spike-and-slab mixture, the Spike-and-Slab LASSO (SSL) prior of Ročková (2015). This prior transforms the obstinate combinatorial search problem into one of optimization in continuous systems, permitting the use of EM algorithms (Dempster et al., 1977), a strategy we pursue here.

The search for promising sparse factor orientations is greatly enhanced with data augmentation by expanding the likelihood with an auxiliary rotation matrix. Exploiting the invariance of the factor model, we propose a PXL-EM (parameter expanded likelihood EM) algorithm, a variant of the PX-EM algorithm of (Liu et al., 1998) and the one-step late PX-EM of van Dyk and Tang (2003) for Bayesian factor analysis. PXL-EM automatically rotates the loading matrix as a part of the estimation process, gearing the EM trajectory along the orbits of equal likelihood. The PXL-EM algorithm is far more robust against poor initializations, converging dramatically faster than the parent EM algorithm.

The SSL prior is coupled with the Indian Buffet Process (IBP) prior, which provides an opportunity to learn about the ambient factor dimensionality. Confirming its potential for this purpose, we provide a tail bound on the expected posterior factor dimensionality, showing that it reflects the true levels of underlying sparsity. In over-parametrized models with many redundant factors, inference about the factor cardinality 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 with our approach, because our IBP construction prioritizes lower indexed loadings and the PXL-EM rotates towards independent factors.

A variety of further steps are proposed to enhance the effectiveness of our approach. To facilitate the search for higher posterior modes we implement a dynamic posterior exploration with a sequential reinitialization of PXL-EM along a ladder of increasing spike penalties. For selection among identified posterior modes, we recommend an evaluation criterion motivated as an integral lower bound to a posterior probability of the implied sparsity pattern. Finally, we introduce an optional varimax rotation step within PXL-EM, to provide further robustification against local convergence issues.

The paper is structured as follows. Section 2 introduces our hierarchical prior formulation and shows some properties of the posterior. Section 3 develops the construction of our PXL-EM algorithm. Section 4 describes the dynamic posterior exploration strategy for PXL-EM deployment. Section 5 derives and illustrates our criterion for factor model comparison. Section 6 illustrates the potential of varimax robustification. Sections 7 presents an applications of our approach on real high-dimensional data. Section 8 concludes with a discussion. Further developments and proofs are provided in the Supplemental material.

2 Infinite Factor Model with the Indian Buffet Process

The cornerstone of our Bayesian approach is a hierarchically structured prior on infinite-dimensional loading matrices, based on the Spike-and-Slab LASSO (SSL) prior of Ročková (2015). Independently for each loading β_{jk}, we consider a two-point mixture of 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,

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

where $\psi(\beta \,|\, \lambda) = \frac{\lambda}{2}\exp\{-\lambda|\beta|\}$ 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 $SSL(\lambda_{0k}, \lambda_1)$. The SSL priors form a

continuum between a single Laplace (LASSO) prior, obtained with $\lambda_{0k} = \lambda_1$, and the point-mass mixture prior, obtained as a limiting case when $\lambda_{0k} \to \infty$.

Coupled with a prior on γ_{jk}, $SL(\lambda_{0k}, \lambda_1)$ generates a spike-and-slab posterior that performs "selective shrinkage" (Ishwaran and Rao, 2005; Ročková, 2015). Posterior modes under the SSL prior are adaptively thresholded, smaller values shrunk to exact zeroes. This is in sharp contrast to spike-and-slab priors with a Gaussian spike (George and McCulloch, 1993; Ročková and George, 2014), whose non-sparse posterior modes must be thresholded for variable selection. The exact sparsity here is crucial for anchoring on interpretable factor orientations and alleviating identifiability issues.

For a prior over the feature allocation matrix $\mathbf{\Gamma} = \{\gamma_{jk}\}_{j,k=1}^{G,\infty}$, we adopt the Indian Buffet Process (IBP) prior (Griffiths and Ghahramani, 2005), which defines an exchangeable distribution over equivalence classes[2] $[\mathbf{\Gamma}]$ of infinite-dimensional binary matrices. Formally, the IBP with an intensity parameter $\alpha > 0$ arises from the beta-Bernoulli prior

$$\begin{aligned} \pi(\gamma_{jk}|\theta_k) &\overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_k), \\ \pi(\theta_k) &\overset{\text{ind}}{\sim} \mathcal{B}\left(\frac{\alpha}{K}, 1\right) \end{aligned} \tag{2.2}$$

by integrating out the θ_k's and by taking the limit $K \to \infty$ (Griffiths and Ghahramani, 2005). Due to its flexibility, the IBP prior has been used in various factor analytic contexts (Knowles and Ghahramani, 2011; Rai and Daumé, 2008; Paisley and Carin, 2009). Our IBP deployment differs from these existing procedures in two important aspects. First, we couple IBP with a continuous spike-and-slab prior rather than the point-mass mixture (compared with Knowles and Ghahramani (2011); Rai and Daumé (2008)). Second, the IBP process is imposed here on the matrix of factor loadings rather than on the matrix of latent factors (compared with Paisley et al. (2012)).

Whereas posterior simulation with the IBP is facilitated by margining over θ_k in (2.2) (Knowles and Ghahramani, 2011; Rai and Daumé, 2008), we instead proceed conditionally on a particular ordering of the θ_k's. As with similar deterministic algorithms for sparse latent allocation models

[2]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.

(Paisley and Carin, 2009), our EM algorithms capitalize on the stick-breaking representation of the IBP derived by Teh et al. (2007).

Theorem 2.1. *(Teh et al., 2007)) Let $\theta_{(1)} > \theta_{(2)} > ... > \theta_{(K)}$ 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: $\theta_{(k)} = \prod_{l=1}^{k} \nu_l$, where $\nu_l \overset{\text{iid}}{\sim} \mathcal{B}(\alpha, 1)$.*

Remark 2.1. *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 on infinite factor loading matrices $\boldsymbol{B} \in \mathbb{R}^{G \times \infty}$, further referred to as SSL-IBP$(\{\lambda_{0k}\}; \lambda_1; \alpha)$, is as follows:

$$\pi(\beta_{jk} \mid \gamma_{jk}) \sim SSL(\lambda_{0k}, \lambda_1), \quad \gamma_{jk} \mid \theta_{(k)} \sim \text{Bernoulli}[\theta_{(k)}], \quad \theta_{(k)} = \prod_{l=1}^{k} \nu_l, \quad \nu_l \overset{\text{iid}}{\sim} \mathcal{B}(\alpha, 1). \quad (2.3)$$

Note that the inclusion of loadings associated with the k^{th} factor is governed by the k^{th} largest inclusion probability $\theta_{(k)}$. According to Theorem 2.1, these $\theta_{(k)}$ decay exponentially with k, hampering the activation of binary indicators γ_{jk} when k is large, and thereby controlling the growth of the ambient factor dimensionality. Thus, the most prominent features are associated with small column indices k.

The prior specification is completed with a prior on diagonal elements of $\boldsymbol{\Sigma}$. We assume independent inverse gamma priors

$$\sigma_1^2, \ldots, \sigma_G^2 \overset{\text{iid}}{\sim} \text{IG}(\eta/2, \eta\xi/2) \quad (2.4)$$

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

The stick-breaking representation in Theorem 2.1 suggests a natural truncated 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 algorithms. The issue remains how to select the order of truncation $K^\star$.

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. Measuring the closeness

of such approximation in the sup-norm metric $d_\infty(\mathbf{\Lambda}, \mathbf{\Lambda}^{K^\star}) = \max_{1\leq j,m\leq G} |\Lambda_{jk} - \Lambda^\star_{mk}|$, we show (Lemma G1 in Supplemental material) that $\mathsf{P}[d_\infty(\mathbf{\Lambda}, \mathbf{\Lambda}^{K^\star}) \leq \varepsilon] > 1 - \varepsilon$ for $K^\star$ sufficiently large. Whereas these considerations are based solely on the prior distribution, in the next section we provide practical guidance for choosing $K^\star$ based on the posterior distribution.

2.1 Learning about Factor Dimensionality

In this section, we investigate the ability of the posterior distribution to identify the factor dimensionality, taking into account the rotational invariance of the likelihood. To begin, we show that the IBP prior concentrates on sparse matrices.

Note that the SSL-IBP prior puts zero mass on factor loading matrices with exact zeros. In such contexts, Pati et al. (2014) and Ročková (2015) deploy a generalized notion of sparsity, regarding coefficients below a certain threshold as negligible. For our continuous SSL-IBP prior, we proceed similarly with the threshold $\delta(\lambda_1, \lambda_0, \theta) = \frac{1}{\lambda_0 - \lambda_1} \log\left[\frac{\lambda_0(1-\theta)}{\lambda_1 \theta}\right]$, the intersection point between the θ-weighted spike-and-slab LASSO densities with penalties (λ_0, λ_1).

Definition 2.1. *Given a loading matrix*[3] $\boldsymbol{B}$*, define the "effective factor dimensionality"* $K(\boldsymbol{B})$ *as the largest column index* K *such that* $|\beta_{jk}| < \delta(\lambda_1, \lambda_{0k}, \theta_{(k)})$ *for* $k > K$ *and* $j = 1, \ldots, G$*.*

The SSL-IBP prior induces an implicit prior distribution on the effective factor dimensionality. In the next theorem, we show that this prior rewards sparse matrices in the sense that it is exponentially decaying. The proof is presented in Section G.1 of the Supplemental material.

Theorem 2.2. *Let* $\boldsymbol{B}_{G\times\infty}$ *be distributed according to the SSL-IBP*$(\{\lambda_{0k}\}, \lambda_1, \alpha)$ *with an intensity parameter* $\alpha = c/G$ *and* $\lambda_1 < \mathrm{e}^{-2}$*, where* $0 < c < G$*. Then for a suitable constant* $D > 0$*, we have*

$$\mathsf{P}[K(\boldsymbol{B}) > K] < D\,\mathrm{e}^{-K\log(1+G/c)}. \tag{2.5}$$

Exponentially decaying priors on the dimensionality are essential for obtaining optimally behaving posteriors in sparse situations. In finite sparse factor analysis, such priors were deployed by Pati

[3] assumed to be left-ordered

et al. (2014) to obtain rate-optimal posterior concentration around any true covariance matrix in spectral norm, when the dimension $G = G_n$ can be much larger than the sample size n.

We now proceed to show the property (2.5) penetrates through the data into the posterior. Let us first introduce some additional notation and assumptions. We assume that (G, K_0) increase with n and thereby supply the index n. Let $\boldsymbol{B}_{0n} = [\boldsymbol{\beta}_0^1, \boldsymbol{\beta}_0^2, \dots]$ be the $(G_n \times \infty)$ true loading matrix with $K_{0n} < G_n$ nonzero columns (ordered to be the leftmost ones) and $\sigma_{0j}, 1 \leq j \leq G$, the true residual variances. Furthermore, assume that each of the K_{0n} active columns satisfies $||\boldsymbol{\beta}_0^k||_0 \leq S_{nk} < G_n/2$, $1 \leq k \leq K_{0n}$. Let s_n^{min} denote the smallest eigenvalue of $\boldsymbol{\Lambda}_{0n}$, the true covariance matrix. Let $||\boldsymbol{B}_{0n}||_2 \equiv |||[\boldsymbol{\beta}_0^1, \dots, \boldsymbol{\beta}_0^{K_{0n}}]|||_2$ be the spectral norm of the sub-matrix of nonzero the columns of $\boldsymbol{B}_{0n}$.

We assume the following: (A) $n < G_n$ and $n \log \sqrt{n} < \log(G_n+1) \sum_k S_{nk}$; (B) $\sqrt{\sum_k S_{nk}}/s_n^{min} \to \infty$ and $\sqrt{\sum_k S_{nk}/n}/s_n^{min} \to 0$; (C) $||\boldsymbol{B}_{0n}||_2 < \min_k \sqrt{S_{nk}}$; and (D) $\sigma_{01}^2 = \dots = \sigma_{0G}^2 = 1$. The assumption (A) limits our considerations to high-dimensional scenarios. In addition, the growth of n should be reasonably slow relative to G_n. The assumption (B)[4] requires that the number of estimable parameters grows slower than n. The assumption (C) is an analog of the assumption (A3) of Pati et al. (2014) and holds in any case with probability at least $1 - \mathrm{e}^{-C\,K_0}$. The assumption (D) avoids the "singleton" identifiability issue and can be relaxed (Pati et al. (2014)).

The following theorem states that, just like the prior, the posterior also concentrates on sparse matrices. Namely, the average posterior probability that $K(\boldsymbol{B})$ "overshoots" the true dimensionality K_{0n} by a multiple of a true sparsity level goes to zero. The proof is presented in Section G.2 of the Supplemental material.

Theorem 2.3. *Assume model* (1.1)*, where the true loading matrix* $\boldsymbol{B}_{0n}$ *has* K_{0n} *nonzero leftmost columns. For* $1 \leq k \leq K_{0n}$ *assume* $\sum_{j=1}^{G_n} \mathbb{I}(\beta_{jk} \neq 0) = S_{nk}$ *and* $S_{nk}/G_n < 1/2$*. Assume* $\boldsymbol{B} \sim$ *SSL-IBP* $(\{\lambda_{0k}\}; \lambda_1; \alpha)$ *with* $\alpha = c/G_n, \lambda_1 < \mathrm{e}^{-2}$ *and* $\lambda_{0k} \geq d\, G_n^2 k^3 n/S_{nk}$*, where* $0 < d$ *and* $0 < c < G$*. Assume* $(A) - (D)$*. Denote by* $\bar{S}_n = \frac{1}{K_{0n}} \sum_{k=1}^{K_{0n}} S_{nk}$*, then*

$$\mathsf{E}_{\boldsymbol{B}_0} \mathsf{P}\left[K(\boldsymbol{B}) > C\, K_{0n} \bar{S}_n \,|\, \boldsymbol{Y}^{(n)}\right] \xrightarrow[n\to\infty]{} 0. \tag{2.6}$$

[4]Under (D), s_n^{min} converges to 1 a.s. as $n \to \infty$, when $\boldsymbol{B}_{0n}$ are `iid` with mean 0 and variance 1 and $G_n > K_{0n}$.

for some $C > 0$.

Note that $K_{0n}\bar{S}_n$ is the actual number of nonzero components in $\boldsymbol{B}_0$. Thus, the tail bound (2.6) reflects the ability of the posterior to identify the ambient sparsity level. However, it also acknowledges the uncertainty about how the nonzero values are distributed among the columns of $\boldsymbol{B}$. This uncertainty is due to the rotational invariance of the likelihood. In the absence of knowledge about $K_{0n}\bar{S}_n$, one would deploy as many columns as are needed to obtain at least one vacant factor (zero loading column). Such an empirical strategy may require gradually increasing the number of columns in repeated runs, or adapting the computational procedure to add extra columns throughout the calculation. This strategy is justified by Theorem 2.3, which shows that only a finite number of columns is needed in the presence of sparsity. If scientific context is available about the likely number of factors, one might like to tune the complexity parameter α of the IBP prior so that it reflects this knowledge.

2.2 Identifiability Considerations

There are two sources of indeterminacy in traditional factor analysis: rotational ambiguity of the likelihood and lack of information when estimating over-parametrized models.

The rotational invariance of the likelihood manifests itself through a multimodal posterior. This source of indeterminacy is ameliorated with the SSL prior, which anchors on sparse representations. This prior promotes factor orientations with many zero loadings by creating ridge-lines of posterior probability along coordinate axes, radically reducing posterior multimodality and exposing sparse models. Posterior simulation typically requires identifiability constraints on the allocation of the zero elements of $\boldsymbol{B}$, such as lower-triangular forms (Geweke and Zhou, 1996) and their generalizations (Frühwirth-Schnatter and Lopes, 2009), to prevent the aggregation of probability mass from multiple modes. Such restrictive constraints are not needed in our mode detection approach. Rather than impose constraints up front, our approach uses the data to find zero allocation patterns. Given such an allocation, it can be verified whether the non-zero loadings are conditionally identifiable (up to column permutations and column sign changes). If not identifiable, the interpretation of such an over-parameterized model may be problematic. In any case, interpretation will only be meaningful in

relation to the scientific context of the problem at hand.

To avoid the lack of identifiability that would occur between a singleton factor loading (a single nonzero loading parameter in a factor) and its idiosyncratic variance, we restrict estimated $\widehat{\boldsymbol{B}}$ to matrices with at least two nonzero factor loadings per factor (Frühwirth-Schnatter and Lopes, 2009). The contribution of the singleton can be absorbed by the residual variance parameter. A practical implementation of this constraint is described in Section 5. Finally, we check to make sure that number of free parameters does not exceed the number of parameters in $\boldsymbol{B}\boldsymbol{B}' + \boldsymbol{\Sigma}$, which is $G(G+1)/2$. For suitably sparse $\widehat{\boldsymbol{B}}$, this condition will be met.

3 The Parameter Expanded Likehood EM (PXL-EM)

3.1 The Vanilla EM Algorithm

As a stepping stone towards the PXL-EM development, we first lay out a vanilla EM algorithm, leveraging the resemblance between factor analysis and multivariate regression. A sparse variant of the EM algorithm for probabilistic principal components (Tipping and Bishop, 1999), we capitalize on the ideas behind EMVS (Ročková and George, 2014), a fast method for posterior model mode detection under spike-and-slab priors in linear regression. To simplify notation, the truncated approximation $\boldsymbol{B}^{K^\star}$ will now be denoted by $\boldsymbol{B}$, for some pre-specified $K^\star$. Similarly, $\boldsymbol{\Gamma} = \{\gamma_{jk}\}_{j,k=1}^{G,K^\star}$ and $\boldsymbol{\theta}$ 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 EM algorithm is 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 is achieved indirectly by iteratively maximizing the expected logarithm of the augmented posterior, treating both the hidden factors $\boldsymbol{\Omega} = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n]'$ and $\boldsymbol{\Gamma}$ as missing data. Given an initialization $\boldsymbol{\Delta}^{(0)}$, the $(m+1)^{st}$ step of the algorithm outputs $\boldsymbol{\Delta}^{(m+1)} = \arg\max_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta})$, where $Q(\boldsymbol{\Delta}) = \mathsf{E}_{\boldsymbol{\Gamma},\boldsymbol{\Omega} \mid \boldsymbol{Y},\boldsymbol{\Delta}^{(m)}} [\log \pi(\boldsymbol{\Delta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega} \mid \boldsymbol{Y})]$,with $\mathsf{E}_{\boldsymbol{\Gamma},\boldsymbol{\Omega} \mid \boldsymbol{Y},\boldsymbol{\Delta}^{(m)}}(\cdot)$ denoting the conditional expectation given the observed data and current parameter estimates at the m^{th} iteration. The E-step is

obtained with simple and fast closed form updates. The M-step boils down to solving a series of independent LASSO regressions, for which fast implementations exist. The IBP stick breaking fractions are updated with a non-linear program. For a detailed development of the EM algorithm with the derivation of the E-step and the M-step, we refer the reader to the Supplemental material A. A brief description of the calculations is presented in Table 1, where the EM algorithm is obtained as a special case by setting $\boldsymbol{A} = \mathrm{I}_{K^\star}$.

3.2 Rotational Ambiguity and Parameter Expansion

The rotational invariance of the likelihood inevitably renders the posterior multimodal, hampering posterior simulation and (global) mode detection. In particular, the EM algorithm outlined in the previous section is vulnerable to entrapment at local modes in the vicinity of initialization. The local convergence issue is exacerbated by strong ties between the loadings and latent factors. Such couplings cement the initial factor orientation, which may be suboptimal, and affects the speed of convergence by 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. Thus, 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)

$$\boldsymbol{y}_i \,|\, \boldsymbol{\omega}_i, \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{A} \overset{\text{ind}}{\sim} \mathcal{N}_G(\boldsymbol{B}\boldsymbol{A}_L^{-1}\boldsymbol{\omega}_i, \boldsymbol{\Sigma}), \quad \boldsymbol{\omega}_i \,|\, \boldsymbol{A} \sim \mathcal{N}_K(\boldsymbol{0}, \boldsymbol{A}), \quad \boldsymbol{A} \sim \pi(\boldsymbol{A}) \tag{3.1}$$

for $1 \leq i \leq n$, where $\boldsymbol{A}_L$ denotes the lower Cholesky factor of $\boldsymbol{A}$, the newly introduced parameter.

Algorithm: *PXL-EM Algorithm for Automatic Rotations to Sparsity*	
Initialize $\boldsymbol{B} = \boldsymbol{B}^{(0)}, \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(0)}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$	
Repeat the following steps until convergence	
The E-Step	
E1: Latent features	$\langle \boldsymbol{\Omega} \rangle' = \boldsymbol{M}\boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'$ where $\boldsymbol{M} = \left(\boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B} + \mathrm{I}_{K^\star}\right)^{-1}$
E2: Latent indicators	$\langle \gamma_{jk} \rangle = \left[1 + \frac{\lambda_0}{\lambda_1}\frac{1-\theta_{(k)}}{\theta_{(k)}}\mathrm{e}^{-\lvert\beta_{jk}\rvert(\lambda_0-\lambda_1)}\right]^{-1}$ $1 \le j \le G, 1 \le k \le K^\star$
The M-Step	
	Set $\widetilde{\boldsymbol{Y}} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0}_{K^\star \times K^\star} \end{pmatrix}$ and $\widetilde{\boldsymbol{\Omega}} = \begin{pmatrix} \langle \boldsymbol{\Omega} \rangle \\ \sqrt{n}\,\boldsymbol{M}_L \end{pmatrix}$
	For $j = 1, \ldots, G$
M1: Loadings	$\boldsymbol{\beta}_j^\star = \arg\max_{\boldsymbol{\beta}} \left\{ -\lvert\lvert\widetilde{\boldsymbol{y}}^j - \widetilde{\boldsymbol{\Omega}}\boldsymbol{\beta}\rvert\rvert^2 - 2\sigma_j^2 \sum_{k=1}^{K^\star} \lvert\beta_k\rvert\lambda_{jk} \right\},$
M2: Variances	$\sigma_j^2 = \frac{1}{n+1}(\lvert\lvert\widetilde{\boldsymbol{y}}^j - \widetilde{\boldsymbol{\Omega}}\boldsymbol{\beta}_j^\star\rvert\rvert^2 + 1)$
M3: Rotation Matrix	$\boldsymbol{A} = \frac{1}{n}\langle \boldsymbol{\Omega} \rangle'\langle \boldsymbol{\Omega} \rangle + \boldsymbol{M}$
M4: Weights	$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} Q_2(\boldsymbol{\theta})$ as described in Section A2 (Supplemental material)
The Rotation Step	
R: Rotation	$\boldsymbol{B} = \boldsymbol{B}^\star \boldsymbol{A}_L$

Legend: $\boldsymbol{M}_L$ is a lower Cholesky factor of $\boldsymbol{M}$, $\langle X \rangle = \mathrm{E}[X \mid \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\theta}, \boldsymbol{Y}]$, $\widetilde{\boldsymbol{Y}} = [\widetilde{\boldsymbol{y}}^1, \ldots, \widetilde{\boldsymbol{y}}^G]$, $\boldsymbol{B}^\star = [\boldsymbol{\beta}_1^\star, \ldots, \boldsymbol{\beta}_G^\star]'$

Table 1: PXL-EM algorithm for sparse Bayesian factor analysis, EM algorithm obtained with $\boldsymbol{A} = \boldsymbol{I}_{K^\star}$

This expansion was used for traditional factor analysis by Liu et al. (1998). The observed-data likelihood here is invariant under the parametrizations indexed by $\boldsymbol{A}$. This is evident from the marginal distribution $f(\boldsymbol{y}_i \mid \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{A}) = \mathcal{N}_G(\boldsymbol{0}, \boldsymbol{B}\boldsymbol{B}' + \boldsymbol{\Sigma})$, $1 \le i \le n$, 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 (2.3) on $\boldsymbol{B}^\star = \boldsymbol{B}\boldsymbol{A}_L^{-1}$ rather than on $\boldsymbol{B}$. That is,

$$\beta_{jk}^\star \mid \gamma_{jk} \overset{\text{ind}}{\sim} SSL(\lambda_{0k}, \lambda_1), \quad \gamma_{jk} \mid \theta_{(k)} \overset{\text{ind}}{\sim} \text{Bernoulli}[\theta_{(k)}], \quad \theta_{(k)} = \prod_{l=1}^{k} \nu_l, \quad \nu_l \overset{\text{iid}}{\sim} \mathcal{B}(\alpha, 1), \tag{3.2}$$

where the $\beta_{jk}^\star$'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 (2.4) on $\boldsymbol{\Sigma}$ remains unchanged.

Just as the vanilla EM algorithm, PXL-EM also targets (local) maxima of the posterior $\pi(\boldsymbol{\Delta} \mid \boldsymbol{Y})$ (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 $\pi(\boldsymbol{\Delta}^\star \mid \boldsymbol{Y})$ indexed by $\boldsymbol{\Delta}^\star = (\boldsymbol{B}^\star, \boldsymbol{\Sigma}, \boldsymbol{\theta}, \boldsymbol{A})$ and implied by (3.1) and (3.2). By iteratively optimizing the conditional expectation of the augmented log posterior $\log \pi^X(\boldsymbol{\Delta}^\star, \boldsymbol{\Omega}, \boldsymbol{\Gamma} \mid \boldsymbol{Y})$, PXL-EM yields a path of $\boldsymbol{\Delta}^\star$ updates through the expanded parameter space. This sequence corresponds to a trajectory in the original parameter space through the reduction function $\boldsymbol{B} = \boldsymbol{B}^\star \boldsymbol{A}_L$. 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 $\boldsymbol{A}$. Thus, the updated $\boldsymbol{A}$ is not carried forward throughout the iterations. Instead, each E-step is anchored on $\boldsymbol{A} = \boldsymbol{A}_0$. As is elaborated on in Section 3.6, $\boldsymbol{A} = \boldsymbol{A}_0$ upon convergence and thus the PXL-EM trajectory converges to local modes of the *original posterior* $\pi(\boldsymbol{\Delta} \mid \boldsymbol{Y})$.

The prior $\pi(\boldsymbol{A})$ influences the orientation of the augmented feature matrix upon convergence. Whereas proper prior distributions $\pi(\boldsymbol{A})$ can be implemented within our framework, and may be a fruitful avenue for future research, here we use $\pi(\boldsymbol{A}) \propto 1$. This improper prior has an "orthogonalization property" which can be exploited for more efficient calculations (Section 3.4). In contrast to marginal augmentation (Liu and Wu, 1999; Meng and van Dyk, 1999), where an improper working prior may cause instability in posterior simulation, here it is more innocuous. This is because PXL-EM does not use the update $\boldsymbol{A}$ for the next E-step.

The PXL M-step uses $\boldsymbol{A}$ to guide the trajectory, which can be very different from the vanilla EM. 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. Due to the thresholding ability of the SSL prior, promising modes are located at the intersection of the orbits with coordinate axes. Obtaining $\boldsymbol{A}$ and subsequently performing the reduction step $\boldsymbol{B} = \boldsymbol{B}^\star \boldsymbol{A}_L$, the PXL-EM trajectory is geared along the orbits, taking larger steps over posterior valleys to conquer multimodality.

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 $\boldsymbol{B}^{(m)} = \boldsymbol{B}^{\star(m)} \boldsymbol{A}_L^{(m)}$. Beginning with the initialization $\boldsymbol{\Delta}^{(0)}$, every step of the PXL-EM algorithm outputs an update $\boldsymbol{\Delta}^{\star(m+1)} = \arg\max_{\boldsymbol{\Delta}^\star} Q^X(\boldsymbol{\Delta}^\star)$, where $Q^X(\boldsymbol{\Delta}^\star) = \mathsf{E}_{\boldsymbol{\Omega},\boldsymbol{\Gamma} \mid \boldsymbol{Y}, \boldsymbol{\Delta}^{(m)}, \boldsymbol{A}_0} \log \pi(\boldsymbol{\Delta}^\star, \boldsymbol{\Omega}, \boldsymbol{\Gamma} | \boldsymbol{Y})$. Each such computation is facilitated by the separability of Q^X 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

$$Q^X(\boldsymbol{\Delta}^\star) = C^X + Q_1(\boldsymbol{B}^\star, \boldsymbol{\Sigma}) + Q_2(\boldsymbol{\theta}) + Q_3^X(\boldsymbol{A}). \tag{3.3}$$

The functions $Q_1(\cdot)$ and $Q_2(\cdot)$ (defined in (A.2) and (A.8) in the Supplemental material) appear in the objective function of the vanilla EM algorithm, suggesting that the M-step will be analogous. In addition, PXL-EM includes an extra term

$$Q_3^X(\boldsymbol{A}) = -\frac{1}{2} \sum_{i=1}^{n} \operatorname{tr}[\boldsymbol{A}^{-1} \mathsf{E}_{\boldsymbol{\Omega} \mid \boldsymbol{\Delta}^{(m)}, \boldsymbol{A}_0} (\boldsymbol{\omega}_i \boldsymbol{\omega}_i')] - \frac{n}{2} \log |\boldsymbol{A}| \tag{3.4}$$

for obtaining a suitable transformation matrix $\boldsymbol{A}$.

3.3 The PXL E-step

The exact calculation of the E-step is presented in Table 1, involving the update of first and second moments of the latent factors $\boldsymbol{\omega}_i$ and the expectation of binary indicators γ_{ij}. These expectations are 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 are the same as for the plain EM algorithm, derived in Section A.1 in the Supplemental material. However, the update $\boldsymbol{B}^{(m)} = \boldsymbol{B}^{\star(m)} \boldsymbol{A}_L^{(m)}$ is now used instead of $\boldsymbol{B}^{\star(m)}$ throughout. The implications of this substitution are discussed in the following intentionally simple example, which conveys the intuition of entries in $\boldsymbol{A}$ as penalties that encourage featurizations with fewer more informative factors. This example highlights the scaling aspect of the transformation induced by $\boldsymbol{A}_L$, assuming $\boldsymbol{A}_L$ is diagonal.

Example 3.1. *(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, (b) determines penalties for variable selection, (c) creates a smoothing penalty matrix $\mathsf{Cov}\,(\boldsymbol{\omega}_i|\boldsymbol{B},\boldsymbol{\Sigma})$. *Here is how inserting* $\boldsymbol{B}=\boldsymbol{B}^\star\boldsymbol{A}_L$ *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 (E1) in Table 1, the new latent features are* $\mathsf{E}_{\boldsymbol{\Omega}\,|\,\boldsymbol{Y},\boldsymbol{B}}(\boldsymbol{\Omega}')=\boldsymbol{A}_L^{-1}(\mathrm{I}_K+\boldsymbol{A}^{-1})^{-1}\boldsymbol{B}^{\star\prime}\boldsymbol{Y}'=\mathrm{diag}\left\{\frac{\sqrt{\alpha}_k}{1+\alpha_k}\right\}\boldsymbol{B}^{\star\prime}\boldsymbol{Y}'$. *Recall that* $\alpha_k=1$ *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* α_k *is close to zero. Regarding the second moments of the latent factors, the coordinates with higher variances* α_k *are penalized less. This is seen from* $\mathsf{Cov}\,(\boldsymbol{\omega}_i\mid\boldsymbol{B},\boldsymbol{\Sigma})=(\boldsymbol{A}_L'\boldsymbol{A}_L+\mathrm{I}_K)^{-1}=\mathrm{diag}\{1/(1+\alpha_k)\}$. *The conditional mixing weights* $\mathsf{E}_{\boldsymbol{\Gamma}\,|\,\boldsymbol{B},\boldsymbol{\theta}}(\gamma_{jk})=\left[1+\frac{\lambda_0}{\lambda_1}\exp\left(-|\beta^\star_{jk}|\alpha_k(\lambda_0-\lambda_1)\right)\right]^{-1}$ *increase exponentially with* α_k. *Higher variances* $\alpha_k>1$ *increase the inclusion probability as compared to no parameter expansion* $\alpha_k=1$. *Thus, the loadings of the newly created features associated with larger* α_k *are more likely to be selected.*

Another example presented in the Supplemental material B illustrates the rotational aspect of $\boldsymbol{A}_L$, when it is non-diagonal. The off-diagonal elements are seen to perform linear aggregation. This example also highlights the benefits of the lower-triangular structure of $\boldsymbol{A}_L$.

3.4 The PXL M-step

Conditionally on the imputed latent data, the M-step is then performed by maximizing $Q^X(\boldsymbol{\Delta}^\star)$ over $\boldsymbol{\Delta}^\star$ in the augmented space. These steps are described in Table 1. The updates of $(\boldsymbol{B}^{\star(m+1)},\boldsymbol{\Sigma}^{(m+1)})$ and $\boldsymbol{\theta}^{(m+1)}$ can be obtained as in the vanilla EM algorithm (Section A.2 in the Supplemental material). PXL-EM requires one additional update $\boldsymbol{A}^{(m+1)}$, obtained by maximizing (3.4). This is a very fast simple operation,

$$\boldsymbol{A}^{(m+1)}=\max_{\boldsymbol{A}=\boldsymbol{A}',\boldsymbol{A}\geq 0}Q_3^X(\boldsymbol{A})=\frac{1}{n}\sum_{i=1}^{n}\mathsf{E}_{\boldsymbol{\Omega}\,|\,\boldsymbol{Y},\boldsymbol{\Delta}^{(m)},\boldsymbol{A}_0}(\boldsymbol{\omega}_i\boldsymbol{\omega}_i')=\frac{1}{n}\langle\boldsymbol{\Omega}'\boldsymbol{\Omega}\rangle=\frac{1}{n}\langle\boldsymbol{\Omega}\rangle'\langle\boldsymbol{\Omega}\rangle+\boldsymbol{M}. \quad (3.5)$$

The new coefficient updates in the reduced parameter space can be 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. This step is missing

from the vanilla EM algorithm, which assumes throughout that $\boldsymbol{A} = \boldsymbol{A}_0 = \mathrm{I}_{K^\star}$. The consequences of this rotation are explored below.

Remark 3.1. *Although $\boldsymbol{A}_L$ is not strictly a rotation matrix in the sense of being orthonormal, we 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 $\boldsymbol{A}_L$, namely the action of $\boldsymbol{U}$.*

More insight into the role of $\boldsymbol{A}_L$ can be gained by recasting the PXL M-step in terms of the original model parameters $\boldsymbol{B} = \boldsymbol{B}^\star\boldsymbol{A}_L$. From (M1) in Table 1, the PXL M-step yields

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

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

$$\boldsymbol{\beta}_j^{(m+1)} = \arg\max_{\boldsymbol{\beta}_j} \left\{ -||\widetilde{\boldsymbol{y}}^j - (\widetilde{\boldsymbol{\Omega}}\boldsymbol{A}_L^{'-1})\boldsymbol{\beta}_j||^2 - 2\sigma_j^{(m)2} \sum_{k=1}^{K^\star} \left| \sum_{l\geq k}^{K^\star} (\boldsymbol{A}_L^{-1})_{lk}\beta_{jl} \right| \lambda_{jk} \right\}. \tag{3.6}$$

Thus, the rotated parameters $\boldsymbol{\beta}_j^{(m+1)}$ are solutions to modified penalized regressions of $\widetilde{\boldsymbol{y}}^j$ on $\widetilde{\boldsymbol{\Omega}}\boldsymbol{A}_L^{'-1}$ under a series of triangular linear constraints. As seen from (3.5) and (3.6), $\boldsymbol{A}_L^{-1}$ serves to "orthogonalize" the factor basis.

Because $\widetilde{\boldsymbol{\Omega}}\boldsymbol{A}_L^{'-1}$ in (3.6) is orthogonal, $\boldsymbol{B}^{(m+1)}$ can be approximated using a closed form update, removing the need for first computing $\boldsymbol{B}^{\star(m+1)}$ and then rotating it by $\boldsymbol{A}_L$. By noting (a) LASSO has a closed form solution in orthogonal designs, (b) the system of constraints in (4.8) is triangular with "dominant" entries on the diagonal, we can deploy back-substitution to quickly obtain an approximate update $\boldsymbol{B}^{(m+1)}$ in just one sweep. Denote by $\boldsymbol{z}^j = \boldsymbol{A}_L^{-1}\widetilde{\boldsymbol{\Omega}}'\boldsymbol{y}^j/n$ and $z_{k+}^j = \sum_{l>k}(\boldsymbol{A}_L^{-1})_{lk}\beta_{jl}/(\boldsymbol{A}_L^{-1})_{kk}$. Then,

$$\beta_{jk}^{(m+1)} \approx \left(|z| - \sigma_j^{(m)2}\lambda_{jk}(\boldsymbol{A}_L^{-1})_{kk}/n \right)_+ \mathrm{sign}(z) - z_{k+}^j, \tag{3.7}$$

where $z = z_k^j + z_{k+}^j$. This approximate step dramatically reduces the computational cost, as discussed in Section C.2 in the Supplemental material, and is therefore worthwhile deploying in large problems. Performing (3.7) instead of the proper M-step (M1) in Table 1 yields a slightly different trajectory. However, both PXL-EM and this trajectory have the same fixed points $(\widehat{\boldsymbol{B}}, \boldsymbol{A}_0)$. Towards convergence when $\boldsymbol{A} \approx \boldsymbol{A}_0$, the approximation (3.7) is close to exact.

To sum up, the default EM algorithm proceeds by finding $\boldsymbol{B}^{(m)}$ at the M-step, and then using this $\boldsymbol{B}^{(m)}$ 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.

Remark 3.2. *PXL-EM performs orthonormalization of the features $\widetilde{\boldsymbol{\Omega}}$ upon convergence. According to* (3.5)*, when PXL-EM converges to its fixed point $(\widehat{\boldsymbol{\Delta}}, \boldsymbol{A}_0)$, we obtain $\frac{1}{n}\langle \boldsymbol{\Omega}'\boldsymbol{\Omega} \rangle = \mathrm{I}_K$. Thus, PXL-EM forces the feature matrix to be orthonormal.*

3.5 Modulating the Trajectory

Our PXL-EM algorithm 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 $\boldsymbol{B} = \boldsymbol{B}^\star \boldsymbol{A}_L$,

$$\log \pi(\boldsymbol{\Delta} \mid \boldsymbol{Y}) - \log \pi(\boldsymbol{\Delta}^{(m)} \mid \boldsymbol{Y}) \geq Q^X(\boldsymbol{\Delta}^\star) - Q^X(\boldsymbol{\Delta}^{(m)}) + \mathsf{E}_{\boldsymbol{\Gamma} \mid \boldsymbol{\Delta}^{(m)}, \boldsymbol{A}_0} \log\left(\frac{\pi(\boldsymbol{B}^\star \boldsymbol{A}_L, \boldsymbol{\Gamma})}{\pi(\boldsymbol{B}^\star, \boldsymbol{\Gamma})}\right). \quad (3.8)$$

Whereas $\boldsymbol{\Delta}^{\star(m+1)} = \arg\max Q^X(\boldsymbol{\Delta}^\star)$ increases the Q^X 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 on the right hand side of (3.8) is positive. Otherwise, the vanilla 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, 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 (3.8) throughout the iterations, we can track the steps in the trajectory that are guaranteed to be monotone.

Apart from monotonization, one could also divert the PXL-EM trajectory with occasional jumps between orbits. Note that PXL-EM moves along orbits indexed by oblique rotations. One might also consider moves along orbits indexed by orthogonal rotations such as varimax (Kaiser (1958)). One might argue that performing a varimax rotation instead of the oblique rotation throughout the EM computation would be equally, if not more, successful. However, the plain EM may fail to provide a sufficiently structured intermediate input for varimax. On the other hand, PXL-EM identifies enough structure early on in the trajectory and may benefit from further varimax rotations. The potential for further improvement with this optional step is demonstrated in Section 7. In the next section we show that PXL-EM is an efficient scheme, i.e. it converges fast.

3.6 Convergence Speed: EM versus PXL-EM

The speed of convergence of the EM algorithm (for MAP estimation) is defined as the smallest eigenvalue of the matrix fraction of the observed information $\boldsymbol{S} = I_{aug}^{-1} I_{obs}$, where

$$I_{obs} = -\frac{\partial^2 \log \pi(\boldsymbol{\Delta} \mid \boldsymbol{Y})}{\partial \boldsymbol{\Delta} \partial \boldsymbol{\Delta}'}\Big|_{\boldsymbol{\Delta}=\widehat{\boldsymbol{\Delta}}}, \quad I_{aug} = -\frac{\partial^2 \log Q(\boldsymbol{\Delta} \mid \boldsymbol{\Delta})}{\partial \boldsymbol{\Delta} \partial \boldsymbol{\Delta}'}\Big|_{\boldsymbol{\Delta}=\widehat{\boldsymbol{\Delta}}} \tag{3.9}$$

and where $\widehat{\boldsymbol{\Delta}} = (\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\theta}})$ is a target posterior mode (Dempster, Laird and Rubin (1968)). The speed matrix satisfies $\boldsymbol{S} = \mathrm{I} - \boldsymbol{DM}$, where $\boldsymbol{DM}$ is the Jacobian of the EM mapping $\boldsymbol{\Delta}^{(t+1)} = \boldsymbol{M}(\boldsymbol{\Delta}^{(t)})$ evaluated at $\widehat{\boldsymbol{\Delta}}$, governing the behavior of the EM algorithm near its fixed point $\widehat{\boldsymbol{\Delta}}$. Due to the fact that $\boldsymbol{M}(\cdot)$ is a soft-thresholding operator on the loadings (and is hence non-differentiable at zero),

we confine attention only to nonzero directions of $\widehat{\boldsymbol{B}}$. This notion of convergence speed supports the intuition: the sparser the mode, the faster the convergence[5]. We obtain an analog of a result of Liu et al. (1998), showing that PXL-EM converges probably faster than EM. The proof is deferred to the Supplemental material (Section C.1).

Theorem 3.1. *Given that PXL-EM converges to* $(\widehat{\boldsymbol{\Delta}}, \boldsymbol{A}_0)$*, it dominates* EM *in terms of the speed of convergence.*

In addition to converging rapidly, PXL-EM also computes quickly. The complexity analysis is presented in Section C.2 in the Supplemental material.

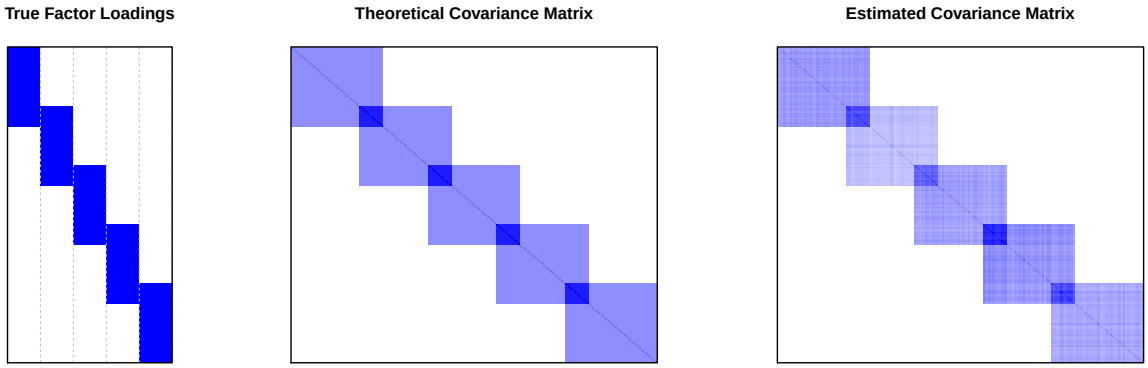
4 The Potential of PXL-EM: A Synthetic Example

4.1 Anchoring Factor Rotation

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 $\lambda_{0k} = \lambda_0$. We set $\lambda_1 = 0.001, \lambda_0 = 20, \alpha = 1/G$ and $K^\star = 20$. All the entries in $\boldsymbol{B}^{(0)}$ 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 accuracy 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 snap-

[5]Taking a sub-matrix S_1 of a symmetric positive-semi definite matrix S_2, the smallest eigenvalue satisfies $\lambda_1(S_1) \geq \lambda_1(S_2)$.



(a) $\boldsymbol{B}_{true}$ (b) $\boldsymbol{B}_{true}\boldsymbol{B}'_{true} + \mathrm{I}_G$ (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).

shots 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 with the LASSO prior in the next section.

The PXL-EM is seen to be robust against poor initializations in this example. 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. This is partly because $\boldsymbol{B}_{true}$ has a regular pattern of overlap

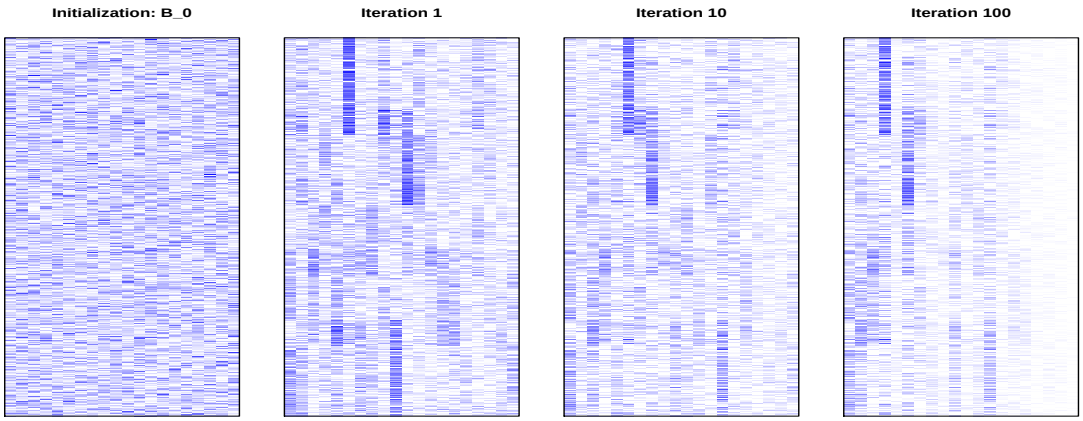


Figure 2: A trajectory of the EM algorithm, convergence *not* achieved even after 100 iterations

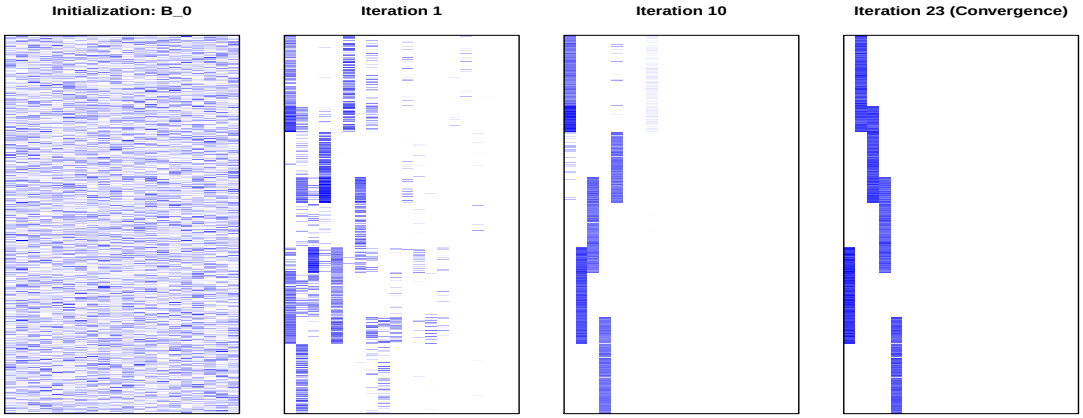
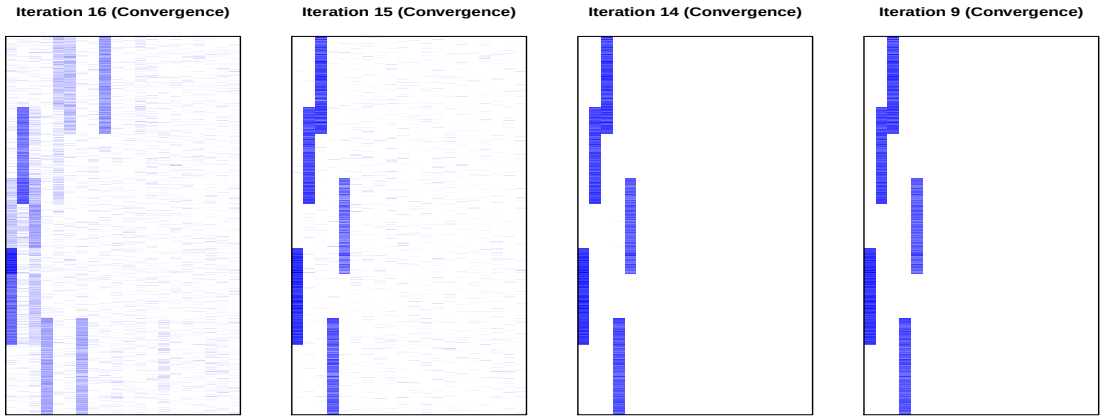


Figure 3: A trajectory of the PXL-EM algorithm, convergence achieved after 23 iterations

and considerable sparsity. In Section 6 we consider a more challenging example, where many more multiple competing sparse modes exist. PXL-EM may then output different, yet similar solutions. For such scenarios, we there propose a robustification step that further mitigates the local convergence issue. Given the vastness of the posterior with its intricate multimodality, and the arbitrariness of the initialization, the results of this experiment are very encouraging. We now turn to the lingering issue of tuning the penalty parameter λ_0.

4.2 Dynamic Posterior Exploration

The character of the posterior landscape is regulated by the two penalty parameters $\lambda_0 >> \lambda_1$. SSL-IBP priors with large differences $(\lambda_0 - \lambda_1)$ induce posteriors with many isolated sharp spikes, exacerbating the already severe multi-modality. In order to facilitate the search for good local maxima

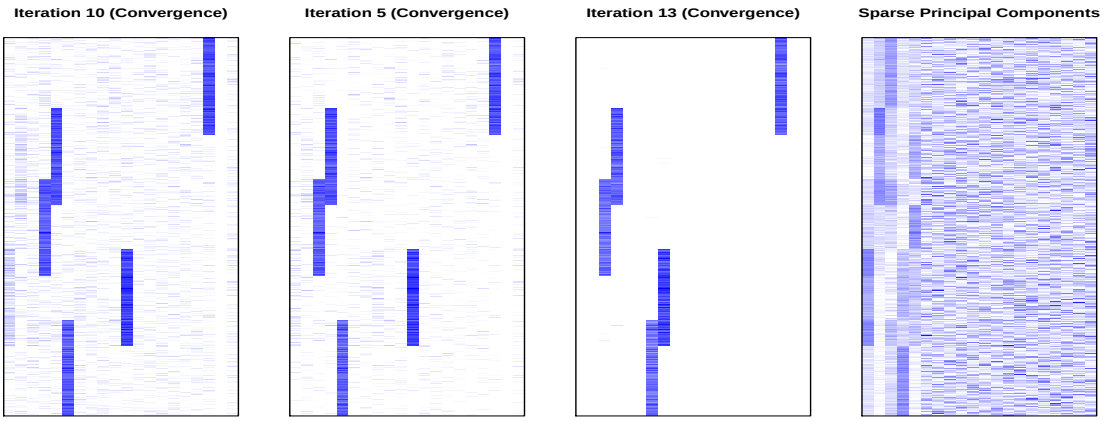


(a) $\lambda_0 = 5$ (b) $\lambda_0 = 10$ (c) $\lambda_0 = 20$ (d) $\lambda_0 = 30$

Figure 4: Recovered loading matrices of PXL-EM for different values of λ_0. The first computation ($\lambda_0 = 5$) initialized at $\boldsymbol{B}^{(0)}$ from the previous section, then reinitialized sequentially.

in the unfriendly multimodal landscape, we borrow ideas from deterministic annealing (Ueda and Nakano, 1998; Yoshida and West, 2010), which optimizes a sequence of modified posteriors indexed by a temperature parameter. Here, we implement a variant of this strategy, treating λ_0 as an inverse temperature parameter. At large temperatures (small values λ_0), the posterior is less spiky and easier to explore.

By keeping the slab penalty λ_1 steady and gradually increasing the spike penalty 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 4). The solution path 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. The stabilization of the estimated loading pattern is an indication that further increase in λ_0



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

Figure 5: PXL-EM with sequential reinitialization along the path using the LASSO prior.

may not be needed and the output is ready for interpretation.

Finally, we explored what would happen if we instead used the single LASSO prior obtained with $\lambda_0 = \lambda_1$. We performed dynamic posterior exploration with $\lambda_0 = \lambda_1$ assuming $\lambda_0 \in I = \{5, 10, 20\}$ (Figure 5(a), (b), (c)). In terms of identifying the nonzero loadings, PXL-EM did reasonably well, generating at best 45 false positives when $\lambda_0 = \lambda_1 = 20$. However, the estimate of the marginal covariance matrix was quite poor, as seen from Figure 6 which compares estimated covariances obtained with the single LASSO and the spike-and-slab LASSO priors. On this example, our PXL-EM implementation of a LASSO-penalized likelihood method dramatically boosted the sparsity recovery over an existing implementation of sparse principal component analysis (SPCA), which does not alter the factor orientation throughout the computation. Figure 5(d) shows the output of SPCA with a LASSO penalty (R package PMA of Witten et al. (2009)), with 20 principal components, using 10 fold cross-validation. Even after supplying the actual correct number of 5 principal components, the

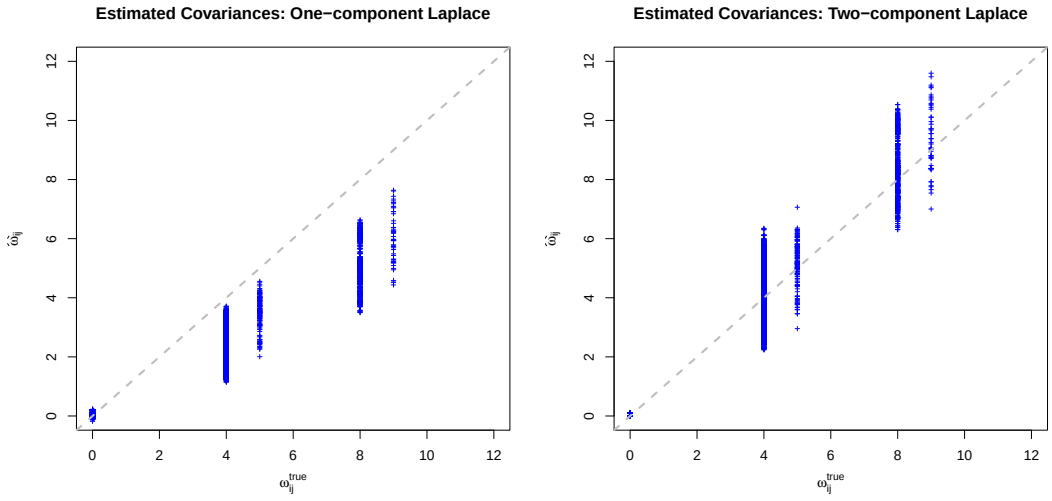


Figure 6: Estimated covariances: LASSO prior vs spike-and-slab LASSO prior

SPCA output was much farther from true sparse solution.

5 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 actual factor dimension K^+, 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 the purpose of comparison and selection from $\{\widehat{\boldsymbol{\Gamma}}_{\lambda_0} : \lambda_0 \in I\}$, a natural and appealing criterion is the posterior model probability $\pi(\boldsymbol{\Gamma} \mid \boldsymbol{Y}) \propto \pi(\boldsymbol{Y} \mid \boldsymbol{\Gamma})\pi(\boldsymbol{\Gamma})$.Whereas the continuous relaxation SSL-IBP$(\{\lambda_{0k}\}; \lambda_1; \alpha)$ was useful for model exploration, the point-mass mixture prior SSL-IBP$(\infty; \lambda_1; \alpha)$ will be more relevant for model evaluation. Unfortunately, computing the marginal likelihood $\pi(\boldsymbol{Y} \,|\, \boldsymbol{\Gamma})$ under these priors is hampered because tractable closed forms are unavailable and

Monte Carlo integration would be impractical. Instead, we replace $\pi(\boldsymbol{Y} \mid \boldsymbol{\Gamma})$by a surrogate function, motivated as an integral lower bound to the marginal likelihood (Minka, 2001).

Beginning with the integral representation $\pi(\boldsymbol{Y} \mid \boldsymbol{\Gamma}) = \int_{\boldsymbol{\Omega}} \pi(\boldsymbol{Y}, \boldsymbol{\Omega} \mid \boldsymbol{\Gamma}) \mathrm{d}\, \pi(\boldsymbol{\Omega})$, which is analytically intractable, we proceed to find an approximation to the marginal likelihood $\pi(\boldsymbol{Y} \mid \boldsymbol{\Gamma})$ by lower-bounding the integrand $\pi(\boldsymbol{Y}, \boldsymbol{\Omega} \mid \boldsymbol{\Gamma}) \geq g_{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}, \boldsymbol{\phi}), \forall (\boldsymbol{\Omega}, \boldsymbol{\phi})$, 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 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. A suitable lower bound for us is $g_{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}, \boldsymbol{\phi}) = C\, \pi(\boldsymbol{Y}, \boldsymbol{\Omega}, \boldsymbol{\phi} \mid \boldsymbol{\Gamma})$, where $\boldsymbol{\phi} = (\boldsymbol{B}, \boldsymbol{\Sigma})$ 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

$$G_{\boldsymbol{\Gamma}}(\boldsymbol{\phi}) = C\, \pi(\boldsymbol{B} \mid \boldsymbol{\Gamma}) \pi(\boldsymbol{\Sigma}) (2\pi)^{-nG/2} |\boldsymbol{\Psi}|^{n/2} \exp\left(-0.5 \sum_{i=1}^{n} \mathrm{tr}(\boldsymbol{\Psi} \boldsymbol{y}_i \boldsymbol{y}_i') \right), \tag{5.1}$$

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

By treating $G_{\boldsymbol{\Gamma}}(\boldsymbol{\phi})$ as the "complete-data" likelihood, finding $\widehat{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi}} \int_{\boldsymbol{\Omega}} \pi(\boldsymbol{Y}, \boldsymbol{\Omega}, \boldsymbol{\phi} | \boldsymbol{\Gamma}) \mathrm{d}\, \boldsymbol{\Omega}$ can be carried out with the (PXL-)EM algorithm. In particular, we can directly use the steps derived in Table 1, 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. These EM calculations are 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 regimes: (a) an exploration regime assuming $\lambda_0 < \infty$ and treating $\boldsymbol{\Gamma}$ as missing to find $\widehat{\boldsymbol{\Gamma}}$, (b) an evaluation regime assuming $\lambda_0 \approx \infty$ 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 function $G_{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\phi}})$ from (5.1) is fundamentally the height of the posterior mode $\pi(\widehat{\boldsymbol{\phi}} \mid \boldsymbol{Y}, \boldsymbol{\Gamma})$ under the point-mass prior SSL-IBP$(\infty; \lambda_1; \alpha)$, assuming $\boldsymbol{\Gamma} = \widehat{\boldsymbol{\Gamma}}$. Despite being a rather crude approximation to the posterior model probability, the function

$$\widetilde{G}(\boldsymbol{\Gamma}) = G_{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\phi}}) \pi(\boldsymbol{\Gamma}), \tag{5.2}$$

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	0.629	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 5: 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

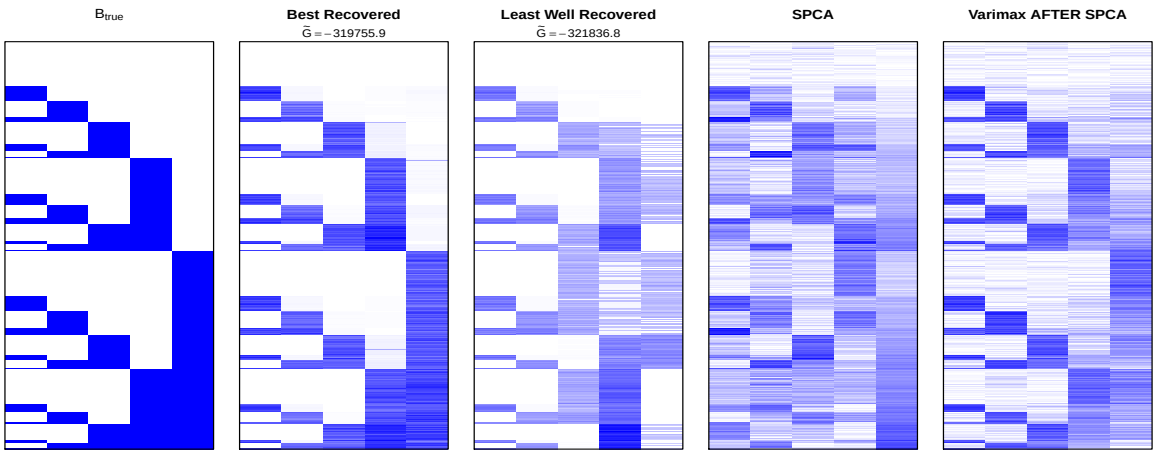
Table 2: Table summarizes the quality of the reconstruction of the marginal covariance matrix $\mathbf{\Lambda}$, namely (a) FDR, (b) FNR, (c) the estimated number of nonzero loadings, (d) the estimated effective factor cardinality, (d) the Frobenius norm $d_F(\widehat{\mathbf{\Lambda}}, \mathbf{\Lambda}_0)$ (Recovery Error).

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

We evaluated the criterion $\widetilde{G}(\mathbf{\Gamma})$ for all the models discovered with the PXL-EM algorithm in the previous section (Figure 4 and Figure 5). We also assessed the quality of the reconstructed marginal covariance matrix (Table 2). The recovery error is computed twice, once after the exploratory run ($\lambda_0 < \infty$) and then after the evaluation run ($\lambda_0 \approx \infty$). Whereas for the exploration, we used both the SSL prior (Figure 4) and the LASSO prior (Figure 5), the evaluation is run *always* with the SSL prior.

The results indicate that the criterion $\widetilde{G}(\mathbf{\Gamma})$ 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 ($\lambda_0 \approx \infty$), which reduces the reconstruction error. This is particularly evident for the single LASSO prior, which achieves good reconstruction properties (estimating the pattern of sparsity) for larger penalty values, however at the expense of the poor recovery of the coefficients (Figure 6).

Given the improved recovery, we recommend outputting the estimates after the evaluation run



(a) (b) (c) (d) (e)

Figure 7: (a) True loading matrix, (b) and (c) PXL-EM using two different random initializations, (d) sparse principal component analysis (SPCA) with $K = 5$, (e) varimax after SPCA

with $\lambda_0 = \infty$. In order to guarantee identifiability (in light of considerations in Section 2), we restrict the support of the IBP prior in the evaluation run to matrices with at least two nonzero γ_{jk} in the active columns of $\boldsymbol{\Gamma}$. With $\lambda_0 = \infty$, this guarantees the absence of singleton loadings in posterior modes.

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.

6 Varimax Robustifications

We now proceed to investigate the performance of PXL-EM in a less stylized scenario with more severe overlap and various degrees of sparsity across the columns. To this end, we generated a zero allocation pattern according to the IBP stick breaking process with $\alpha = 2$, assuming $G = 2\,000$ and $n = 100$. Confining attention to the $K^+ = 5$ strongest factors and permuting the rows and columns, we obtained a true loading matrix $\boldsymbol{B}_{true}$ with $\sum \gamma_{jk}^{true} = 3\,410$ nonzero entries, all set equal to 1 (Figure 7(a)). With considerable overlap and less regular sparsity, the detection problem here is

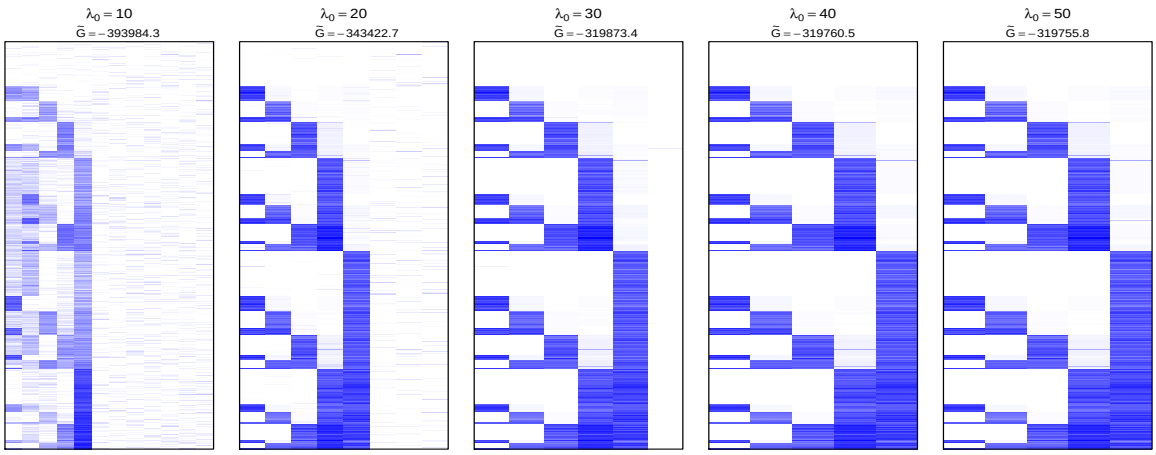


Figure 8: Dynamic posterior exploration of PXL-EM with a varimax rotation every 5 iterations. The initialization is the same as in Figure 7(c). Only nonzero columns are plotted.

far more challenging. There are more competing sparse rotations and thus more sensitivity towards initialization. Generating a dataset with $\sigma_{01}^2 = \cdots = \sigma_{0G}^2 = 1$ and setting $K^\star = 20$, we perform dynamic posterior exploration with $\lambda_0 \in I = \{10, 20, 30, 40, 50\}$ using 10 random starting matrices (generated from a matrix-valued standard Gaussian distribution).

In all 10 independent runs, PXL-EM output at $\lambda_0 = 50$ consistently identified the correct factor dimensionality. Two selected solutions[6] are depicted in Figure 7, the best reconstructed (Figure 7(b)) and the least well reconstructed loading matrix (Figure 7(c)). PXL-EM recovered the correct orientation (as in Figure 7(b)) in 3 out of the 10 runs. These three sparse orientations were rewarded with the highest values of the criterion $\widetilde{G}(\widehat{\mathbf{\Gamma}})$. The other 7 runs output somewhat less sparse variants of the true loading pattern, with two or three loading columns rotated (as in Figure 7(c)). We also compared our reconstructions with a sparse PCA method (R package PMD), performing 5-fold cross validation while setting the dimensionality equal to the oracle value $K^+ = 5$. The recovered loading matrix (Figure 7(d)) captures some of the pattern, however fares less favorably.

Interestingly, performing a varimax rotation *after* sparse PCA greatly enhanced the recovery (Figure 7(e)). Similar improvement was seen after applying varimax to the suboptimal solution in Figure

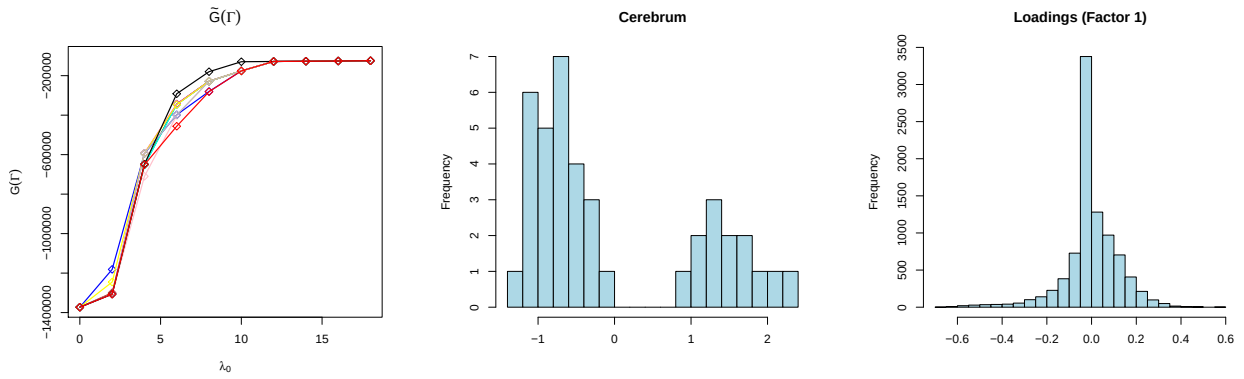
[6]All 10 solutions recovered by dynamic posterior exploration are reported in Section F of the Supplemental material.

7(c). On the other hand, the varimax rotation did not affect the solution in Figure 7(b). As discussed in Section 3.5, we consider a robustification of PXL-EM by including an occasional varimax rotation (every 5 iterations) throughout the PXL-EM computation. Such a step proved to be remarkably effective here, eliminating the local convergence issue. Applying this enhancement with the "least favorable" starting value (Figure 7(c)) yielded the solution path depicted in Figure 8. The correct rotation was identified early in the solution path. With the varimax step, convergence to the correct orientation was actually observed with every random initialization we considered.

7 The AGEMAP Data

We 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 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, as 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



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

Figure 9: (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

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 investigate the sensitivity to initialization, we consider 10 random starting matrices (standard Gaussian entries) to initialize the solution path. We use $\mathbf{\Sigma}^{(0)} = \mathrm{I}_G$, $\boldsymbol{\theta}^{(0)} = (0.5, \ldots, 0.5)'$ as initialization for every λ_0. The margin $\varepsilon = 0.01$ is used to claim convergence.

The results of dynamic posterior exploration are summarized in Table 2 of Section E of the Supplementary material. 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 9(a) and shows a remarkably similar pattern, despite the very arbitrary initializations. From both Table 1 and Figure 9(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. Based on this analysis, we would select just one factor. The best recovery, according to the value

$\widetilde{G}(\widehat{\mathbf{\Gamma}})$, yields a single latent feature (histogram on Figure 9(b)). This latent variable has a strikingly dichotomous pattern, suggesting the presence of an underlying binary hidden variable. A similar histogram was reported also by Perry and Owen (2010). Their finding was supported by a statistical test.

The representation, despite sparse in terms of the number of factors, is not sparse in terms of factor loadings. The single factor loaded on the majority of considered genes (78%). The histogram of estimated loadings (Figure 9(c)) suggests that there are a few very active genes that could potentially be interpreted as leading genes for the factor. We note that the concise representation with a single latent factor could not obtained using, for instance, sparse principal components, which smear the signal across multiple factors when the factor dimension is overfitted.

We further demonstrate the usefulness of our method with the familiar Kendall applicant dataset in Section D of the Supplemental material.

8 Discussion

We have presented a new Bayesian strategy for the discovery of interpretable latent factor models through automatic rotations to sparsity. These rotations are introduced via parameter expansion within a PXL-EM algorithm that iterates between soft-thresholding and transformations of the factor basis. Beyond its value as a method for automatic reduction to simple structure, our methodology enhances the potential for interpretability. It should be emphasized, however, that any such interpretations will ultimately only be meaningful in relation to the scientific context under consideration.

The EM acceleration with parameter expansion is related to 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. 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). An interesting future avenue would be implementing a marginal augmentation variant of our approach in the context of posterior simulation.

By deploying the IBP process, we have avoided the need for fixing the factor dimensionality in advance. By providing a posterior tail bound on the number of factors, we have shown that our posterior distribution reflects the true underlying sparse dimensionality. This result constitutes an essential first step towards establishing posterior concentration rate results for covariance matrix estimation (similar as in Pati et al (2014)). As the SSL prior itself yields rate-optimal posterior concentration in orthogonal regression designs (Ročková, 2015) and high-dimensional regression (Ročková and George, 2015), the SSL-IBP prior is en a promising route towards similarly well-behaved posteriors.

Although full posterior inference is unavailable with our approach, local uncertainty estimates can still be obtained. For example, by conditioning on a selected sparsity pattern $\widehat{\boldsymbol{\Gamma}}$, MCMC can be used to simulate from $\pi(\boldsymbol{B}, \boldsymbol{\Sigma} \mid \widehat{\boldsymbol{\Gamma}}, \boldsymbol{Y})$, focusing only on the nonzero entries in $\widehat{\boldsymbol{B}}$. Conditional credibility intervals for these nonzero values under the point-mass spike-and-slab prior could then be efficiently obtained whenever $\widehat{\boldsymbol{B}}$ is reasonably sparse. Alternatively, the inverse covariance matrix can be estimated by the observed information matrix in (4.12), again confining attention to the nonzero entries in $\widehat{\boldsymbol{B}}$. To this end, the supplemented EM algorithm (Meng and Rubin, 1991) could be deployed to obtain a numerically stable estimate of the asymptotic covariance matrix of the EM-computed estimate.

Our approach can be directly extended to non-Gaussian or mixed outcome latent variable models using data augmentation with hidden continuous responses. For instance, probit/logistic augmentations (Albert and Chib, 1993; Polson et al., 2013) can be deployed to implement a variant of factor analysis for binary responses (Klami, 2014). A generalization of the PXL-EM for this setup requires only one more step, a closed form update of the hidden continuous data. The implementation of this step is readily available. Potentially vast improvement can be obtained using extra parameter expansion, introducing additional working variance parameters of the hidden data (as in the probit

regression Example 4.3 of Liu et al. (1998)). Our methodology can be further extended to canonical correlation analysis or to latent factor augmentations of multivariate regression.

Acknowledgments

The authors would like to thank the Associate Editor and the anonymous referees for their insightful comments and useful suggestions. We would also like to thank Art Owen for kindly providing the AGEMAP dataset. The work was supported by NSF grant DMS-1406563 and AHRQ Grant R21-HS021854.

References

Albert, J. H. and Chib, S. (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, 88, 669–679.

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.

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.

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.

Geweke, J. and Zhou, G. (1996), "Measuring the pricing error of the arbitrage pricing theory," *The review of financial studies*, 9, 557–587.

Ghosh, J. and Dunson, D. (2009), "Default prior distributions and efficient posterior computation in Bayesian factor analysis," *Journal of Computational and Graphical Statistics*, 18, 306Ű320.

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.

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

Kaiser, H. (1958), "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, 23, 187–200.

Klami, A. (2014), "Pólya-gamma augmentations for factor models," in "JMLR: Workshop and Conference Proceedings 29," .

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.

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

Meng, X. and Rubin, D. (1991), "Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm," *Journal of the American Statistical Association*, 86, 899–909.

Meng, X. and van Dyk, D. (1999), "Seeking efficient data augmentation schemes via conditional and marginal augmentation," *Biometrika*, 86, 301–320.

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

Paisley, B. and Carin, L. (2009), "Nonparametric factor analysis with beta process priors," in "26th International Conference on Machine Learning," .

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.

Polson, N., Scott, J., and Windle, J. (2013), "Bayesian inference for logistic models using Pólya-gamma latent variables," *Journal of the American Statistical Association*, 108, 1339–1349.

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," .

Ročková, V. (2015), "Bayesian estimation of sparse signals with a continuous spike-and-slab prior," *Submitted manuscript*.

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

Ročková, V. and George, E. (2015), "The Spike-and-Slab LASSO," *Submitted manuscript*.

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

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

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.