Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity

Supplemental Materials

A The EM Approach to Bayesian Factor Analysis

We here provide the full development of the vanilla EM algorithm outlined in Section 3.1. We remind the reader that $\boldsymbol{B}$ now denotes the truncated approximation $\boldsymbol{B}^{K^\star}$, for some pre-specified $K^\star$, $\boldsymbol{\theta} = (\theta_{(1)}, \ldots, \theta_{(K^\star)})'$ and $\lambda_{0k} = \lambda_0$ for $k = 1, \ldots, K^\star$.

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

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

$$Q(\boldsymbol{\Delta}) = \mathsf{E}_{\boldsymbol{\Gamma},\boldsymbol{\Omega} \mid \boldsymbol{Y},\boldsymbol{\Delta}^{(m)}} \left[\log \pi(\boldsymbol{\Delta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega} \mid \boldsymbol{Y})\right], \tag{A.1}$$

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. Note that we have parametrized our posterior in terms of the ordered inclusion probabilities $\boldsymbol{\theta}$ rather than the breaking fractions $\boldsymbol{\nu}$. These can be recovered using the stick-breaking relationship $\nu_k = \theta_{(k)}/\theta_{(k-1)}$. This parametrization yields a feasible M-step, as will

be seen below.

We now take a closer look at the objective function (A.1). For notational convenience, let $\langle X \rangle$ denote the conditional expectation $\mathsf{E}_{\boldsymbol{\Gamma},\boldsymbol{\Omega} \mid \boldsymbol{Y},\boldsymbol{\Delta}^{(m)}}(X)$. As a consequence of the hierarchical separation of model parameters, $(\boldsymbol{B}, \boldsymbol{\Sigma})$ and $\boldsymbol{\theta}$ are conditionally independent given $(\boldsymbol{\Omega}, \boldsymbol{\Gamma})$. Thereby

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

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

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

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

A.1 The E-step

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

$$\langle \boldsymbol{\Omega}' \rangle = \left(\boldsymbol{B}^{(m)'} \boldsymbol{\Sigma}^{(m)^{-1}} \boldsymbol{B}^{(m)} + \mathrm{I}_{K^\star} \right)^{-1} \boldsymbol{B}^{(m)'} \boldsymbol{\Sigma}^{(m)^{-1}} \boldsymbol{Y}'. \tag{A.3}$$

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

The E-step then proceeds by updating the expectation of the binary allocations $\boldsymbol{\Gamma}$. The entries can be updated individually by noting that conditionally on $\boldsymbol{\Delta}^{(m)}$, the γ_{jk}'s are independent. The model

hierarchy separates the indicators from the data through the factor loadings so that $\pi(\boldsymbol{\Gamma} \mid \boldsymbol{\Delta}, \boldsymbol{Y}) = \pi(\boldsymbol{\Gamma} \mid \boldsymbol{\Delta})$ does not depend on $\boldsymbol{Y}$. This leads to rapidly computable updates

$$\langle \gamma_{jk} \rangle \equiv \mathsf{P}\left(\gamma_{jk} = 1 \mid \boldsymbol{\Delta}^{(m)}\right) = \frac{\theta_{(k)}^{(m)} \psi(\beta_{jk}^{(m)} | \lambda_1)}{\theta_{(k)}^{(m)} \psi(\beta_{jk}^{(m)} | \lambda_1) + (1 - \theta_{(k)}^{(m)}) \psi(\beta_{jk}^{(m)} | \lambda_0)}. \tag{A.4}$$

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

A.2 The M-step

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

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

$$Q_1(\boldsymbol{B}, \boldsymbol{\Sigma}) = \sum_{j=1}^{G} Q_j(\boldsymbol{\beta}_j, \sigma_j), \tag{A.5}$$

where

$$Q_j(\boldsymbol{\beta}_j, \sigma_j) = -\frac{1}{2\sigma_j^2} ||\widetilde{\boldsymbol{y}}^j - \widetilde{\boldsymbol{\Omega}} \boldsymbol{\beta}_j||^2 - \sum_{k=1}^{K^\star} |\beta_{jk}| \lambda_{jk} - \frac{n+1}{2} \log \sigma_j^2 - \frac{1}{2\sigma_j^2} \tag{A.6}$$

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

Proof. The statement follows by rearranging the terms in the first row of (A.2). Namely, in the likelihood term we replace the row summation by a column summation. Then, we rewrite $\frac{n}{2} \mathsf{tr}(\boldsymbol{B}' \boldsymbol{\Sigma}^{-1} \boldsymbol{B} \boldsymbol{M}) =$

$\sum_{j=1}^{G} \frac{n}{2\sigma_j^2} \boldsymbol{\beta}_j' \boldsymbol{M} \boldsymbol{\beta}_j$. This quadratic penalty can be embedded within the likelihood term by augmenting the data rows as stated in the theorem. □

Remark A.1. *The proof of Theorem A.1 explains why $\boldsymbol{M}$ can be regarded as a kernel of a Markov Random Field smoothing prior, penalizing linear combinations of loadings associated with correlated factors.*

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

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

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

$$Q_2^\star(\boldsymbol{\nu}) = \sum_{j=1}^{G} \sum_{k=1}^{K^\star} \left\{ \langle \gamma_{jk} \rangle \sum_{l=1}^{k} \log \nu_l + (1 - \langle \gamma_{jk} \rangle) \log \left(1 - \prod_{l=1}^{k} \nu_l \right) \right\} + (\alpha - 1) \sum_{k=1}^{K^\star} \log(\nu_k), \quad \text{(A.7)}$$

a nonlinear function that is difficult to optimize. Instead, we use the stick-breaking law and plug

$\nu_k = \theta_{(k)}/\theta_{(k-1)}$ into (A.7). The objective function then becomes

$$Q_2(\boldsymbol{\theta}) = \sum_{j=1}^{G}\sum_{k=1}^{K^\star}\left[\langle\gamma_{jk}\rangle \log\theta_{(k)} + (1-\langle\gamma_{jk}\rangle)\log(1-\theta_{(k)})\right] + (\alpha-1)\log\theta_{(K^\star)}, \tag{A.8}$$

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

$$\theta_{(k)} - \theta_{(k-1)} \leq 0,\ k = 2,\ldots,K^\star,$$
$$0 \leq \theta_{(k)} \leq 1,\ k = 1,\ldots,K^\star.$$

This step can be enhanced by performing a permutation of the columns (sorting the factors by their binary numbers) prior to solving the constrained optimization. Had we assumed the finite beta-Bernoulli prior (2.2), the update of the (unordered) occurrence probabilities would simply become $\theta_k^{(m+1)} = \frac{\sum_{j=1}^{G}\langle\gamma_{jk}\rangle + a - 1}{a+b+G-2}$.

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

B PXL E-Step: Example with a Non-Diagonal A

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

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

where $\boldsymbol{\theta} = (0.5,\ldots,0.5)'$. The lower ordered inclusion indicators are likely to be higher since they aggregate more coefficients, a consequence of the lower-triangular form $\boldsymbol{A}_L$. As a result, lower

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

C Efficiency of PXL-EM

C.1 Convergence Speed: Proof of Theorem 5.1

To establish the speed of convergence of PXL-EM relative to EM, we need to change slightly the notation. Let $\boldsymbol{\Delta}^\star = (\boldsymbol{\Delta}, \boldsymbol{A}) = (\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\theta}, \boldsymbol{A})$ be the parameters in the expanded space, where we use $\boldsymbol{B} = \boldsymbol{B}^\star\boldsymbol{A}_L$, which has already been rotated, in place of $\boldsymbol{B}^\star$ in $\boldsymbol{\Delta}^\star$ from before. Clearly $\widehat{\boldsymbol{\Delta}}^\star = (\widehat{\boldsymbol{\Delta}}, \boldsymbol{A}_0)$ is a fixed point of PXL-EM. Similarly as Liu et al. (1998), to obtain the speed matrix near the fixed point $\widehat{\boldsymbol{\Delta}}^\star$ we define[1]

$$I^X_{obs} = -\frac{\partial^2 \log \pi(\boldsymbol{\Delta} \,|\, \boldsymbol{Y})}{\partial\boldsymbol{\Delta}^\star\partial\boldsymbol{\Delta}^{\star\prime}}\Big|_{\boldsymbol{\Delta}^\star=\widehat{\boldsymbol{\Delta}}^\star}, \quad I^X_{aug} = -\frac{\partial^2 \log Q^X(\boldsymbol{\Delta}^\star \,|\, \boldsymbol{\Delta}^\star)}{\partial\boldsymbol{\Delta}^\star\partial\boldsymbol{\Delta}^{\star\prime}}\Big|_{\boldsymbol{\Delta}^\star=\widehat{\boldsymbol{\Delta}}^\star}. \tag{C.1}$$

Splitting the Jacobian into sub-matrices associated with $\boldsymbol{\Delta} = (\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\theta})$ and $\boldsymbol{A}$, we can relate I^X_{obs} and I^X_{aug} to I_{obs} and I_{aug} from (3.9) as follows:

$$I^X_{obs} = \begin{pmatrix} I_{obs} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{and} \quad I^X_{aug} = \begin{pmatrix} I_{aug} & I_{\boldsymbol{\Delta}^\star,\boldsymbol{A}} \\ I_{\boldsymbol{A},\boldsymbol{\Delta}^\star} & I_{\boldsymbol{A},\boldsymbol{A}} \end{pmatrix}. \tag{C.2}$$

Whereas obtaining I^X_{obs} is rather obvious, I^X_{aug} needs a bit of clarification. It can be verified that

$$-\frac{\partial^2 \log Q^X(\boldsymbol{\Delta}^\star \,|\, \boldsymbol{\Delta}^\star)}{\partial\boldsymbol{\Delta}\partial\boldsymbol{\Delta}'}\Big|_{\boldsymbol{\Delta}^\star=\widehat{\boldsymbol{\Delta}}^\star} = I_{aug} - \frac{\partial^2}{\partial\boldsymbol{\Delta}\partial\boldsymbol{\Delta}'}\left\{\mathsf{E}_{\boldsymbol{\Gamma}\,|\,\boldsymbol{\Delta}^\star}\log\left[\frac{\pi(\boldsymbol{B}\boldsymbol{A}_L^{-1}, \boldsymbol{\Gamma})}{\pi(\boldsymbol{B}, \boldsymbol{\Gamma})}\right]\right\}\Big|_{(\boldsymbol{B},\boldsymbol{\theta},\boldsymbol{A})=(\widehat{\boldsymbol{B}},\widehat{\boldsymbol{\theta}},\boldsymbol{A}_0)} = I_{aug}$$

An implication of (C.2) is that the convergence of $\boldsymbol{\Delta}$ determines the convergence of PXL-EM (by arguments analogous to Liu et al. (1998) in Section 3.3). Therefore $\widehat{\boldsymbol{A}} \approx \boldsymbol{A}_0$ near convergence when

[1] The derivatives in (C.1) are taken only with respect to nonzero directions in $\boldsymbol{B}$!

$\boldsymbol{\Delta} \approx \widehat{\boldsymbol{\Delta}}$. Let $\left(I_{aug}^{X}\right)^{-1} = \begin{pmatrix} i_{11} & i_{12} \\ i_{21} & i_{22} \end{pmatrix}$ where $i_{11} = I_{aug}^{-1} + i_{12}\, i_{22}\, i_{21}$. Then $S_X = i_{11} I_{obs}$ is the speed matrix of PXL-EM. The smallest eigenvalue of $\lambda_1(S_X)$ is at least as large as the smallest eigenvalue of $\lambda_1(S)$. Thus, we have an analog of Theorem 2 of Liu et al. (1998).

C.2 Computational Complexity: EM versus PXL-EM

In addition to fast convergence, the implementation of PXL-EM is very efficient. Here, we perform a complexity analysis of both EM and PXL-EM, assuming $K^\star < n < G$.

The computational cost of the E-step of the EM/PXL-EM algorithm (Section 3.1) is dominated by the featurization (3.3). Obtaining the covariance matrix $\boldsymbol{M}$ requires $\mathcal{O}(K^{\star 2}G)$ operations. An additional $\mathcal{O}(K^\star Gn)$ operations are needed to compute $\langle \boldsymbol{\Omega} \rangle$. Thus, the total computational cost[2] of (3.3) is $\mathcal{O}(K^\star Gn)$. The most expensive operation in the M-step (Section 3.2) is the update $\boldsymbol{B}^{(m+1)}$. Each of the G individual LASSO solutions costs at most $\mathcal{O}[K^{\star 2}(n + K^\star)]$ operations (Meinshausen, 2007). The total complexity of one M-step, and thus of one iteration of the EM algorithm, is $\mathcal{O}[GK^{\star 2}(n + K^\star)]$.

In addition, the PXL M-step computes $\boldsymbol{A}_L^{(m+1)}$, obtainable with at most $\mathcal{O}(n^2 K^\star) + \mathcal{O}(K^{\star 3})$ operations. Thus, the complexity of one iteration of PXL-EM is still $\mathcal{O}[GK^{\star 2}(n+K^\star)]$. However, the back-substitution (3.7) approximates $\boldsymbol{B}^{(m+1)}$ at the expense of at most $\mathcal{O}[GK^\star(K^\star + 1)]$ operations. Thus, one iteration of PXL-EM with the approximate M-step can be computed with at most $\mathcal{O}(K^\star Gn)$ operations.

[2] Assuming $\boldsymbol{B}$ has at most S nonzero entries in each column, the cost is only $\mathcal{O}(K^\star n \max\{n, S\})$.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	$\widehat{\sigma}_j$
Application	0.89	**-1.3**	**2.08**	0	0	0.16
Appearance	**1.01**	-0.02	0.14	**-1.6**	0	0.18
Academic ability	0.2	-0.5	-0.3	-0.2	-0.1	1.84
Likability	1.41	-0.03	0.46	-0.35	**2.29**	0.16
Self-confidence	**2.02**	0	0	0	0	1.24
Lucidity	**2.81**	0	0	0	0	1.28
Honesty	0	0	0	0	0	2.49
Salesmanship	**3.13**	0	0	0	0	1.30
Experience	0.87	**-3.11**	0	0	0	0.24
Drive	**2.51**	0	0	0	0	1.42
Ambition	**2.61**	0	0	0	0	1.20
Grasp	**2.72**	0	0	0	0	1.20
Potential	**2.79**	0	0	0	0	1.39
Keenness	**1.67**	0	0	0	0	2.00
Suitability	**1.85**	**-1.91**	0	0	0	1.84

Table 1: Kendall's data: estimated loading matrix and residual variance parameters, in bold are loading estimates that are greater than one in absolute value

D Kendall's Applicant Data

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

After centering the scores, we run our PXL-EM algorithm assuming $K^\star = 10$, $\lambda_1 = 0.001, \alpha = 1/15$. For dynamic posterior exploration we consider 10 random starting matrices (standard Gaussian entries) and a tempering schedule $\lambda_0 \in I = \{1, 2, \ldots, 50\}$. We use a convergence margin $\varepsilon = 0.01$. We deploy PXL-EM also with an intermediate varimax rotation every 5 iterations (as discussed

in Section 7). The best solution, according to the criterion (5.2), was obtained with the varimax adjustment, yielding $\widehat{K}^+ = 5$ factors. The associated loading matrix together with estimated residual variances is displayed in Table 1. With the varimax step, all 10 initializations lead to the same rotation.

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

E Details of the AGEMAP Analysis

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

Table 2: Evolutions of the $\widetilde{G}(\widehat{\mathbf{\Gamma}})$ function together with estimated factor dimension $\widehat{K^+}$ and estimated number of model parameters $\sum_{jk}\widehat{\gamma}_{jk}$ in dynamic posterior exploration using 10 random initializations.

F Simulated Example from Section 6

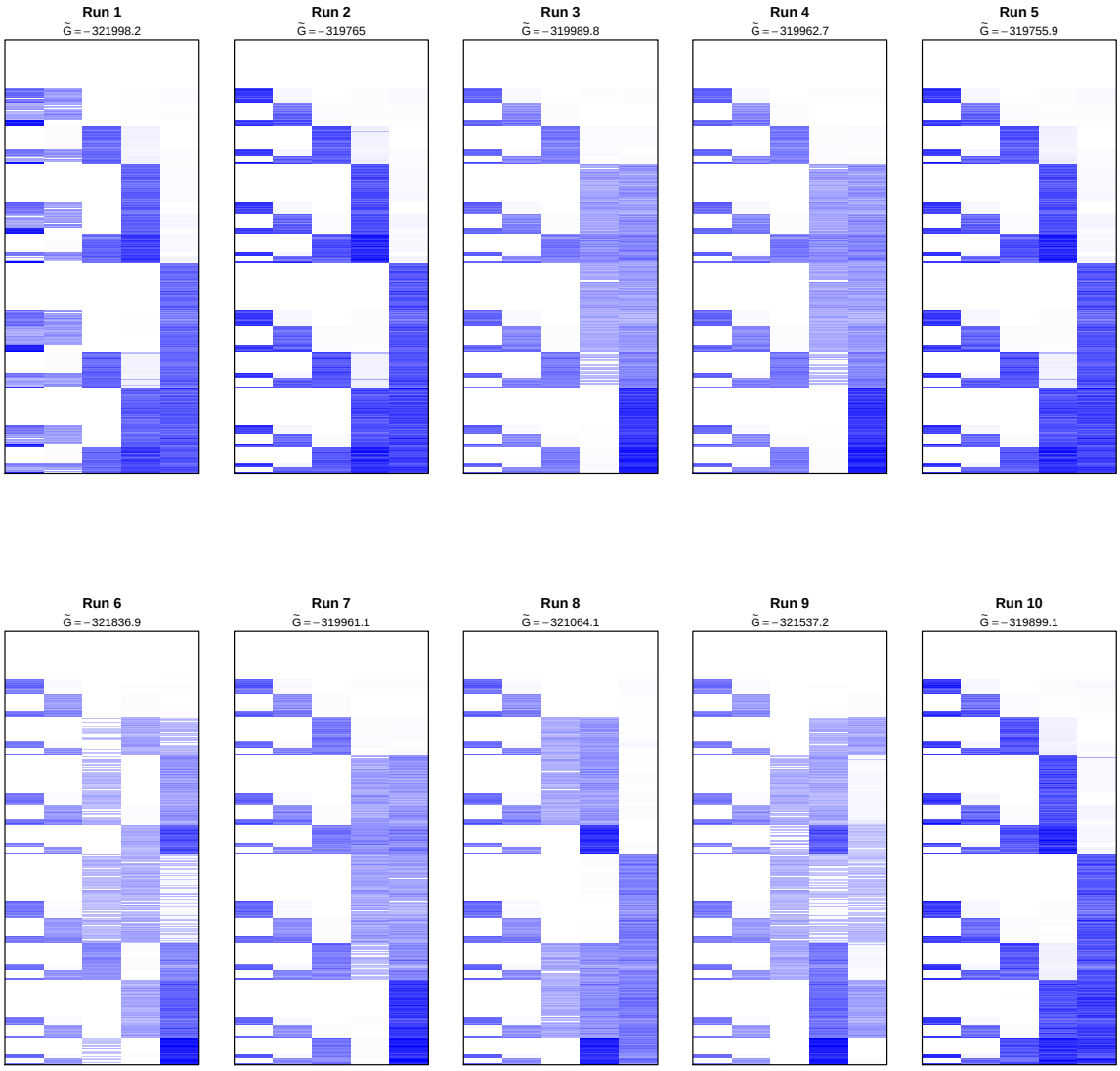

Figure 1: Recovered loading matrices of PXL-EM from 10 independent random initializations. Figures show a heat-map of absolute loadings obtained by dynamic posterior exploration at $\lambda_0 = 50$ together with a value $\widetilde{G}(\widehat{\mathbf{\Gamma}})$.

G Theory of Section 2

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

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

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

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

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

where

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

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

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

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

G.1 Proof of Theorem 2.2

We begin with an auxiliary lemma, showing that the IBP prior implies an exponentially decaying distribution on the actual dimensionality K^+.

Theorem G.1. *Let $[\mathbf{\Gamma}]$ be distributed according to the IBP prior (2.3) with an intensity $0 < \alpha \leq 1$. Let K^+ denote the actual factor dimension, that is the largest index $K^+ \in \mathbb{N}$, so that $\gamma_{jk} = 0$ for all $k > K^+, j = 1, \ldots, G$. Then*

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

Proof. We can write

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

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

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

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

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

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

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

Next, we need the following simple lemma.

Lemma G.2. *Assume $\beta \sim$ SSL(λ_0, λ_1) with $\mathsf{P}(\gamma = 1) = \theta \in (0, 1/2)$, $\lambda_1 \leq \mathrm{e}^{-2}$ and $\lambda_0 > 1$. Then*

$$\mathsf{P}[|\beta| > \delta(\lambda_0, \lambda_1, \theta)] < \theta.$$

Proof. Denote by $\delta_\theta = \delta(\lambda_0, \lambda_1, \theta)$. Using the fact $\theta\,\phi_1(\delta_\theta) = (1-\theta)\,\phi_0(\delta_\theta)$, we have $\mathsf{P}(|\beta| > \delta_\theta) = \theta\exp[-\delta_\theta\lambda_1]\left(1 + \frac{\lambda_1}{\lambda_0}\right)$. Because $\lambda_1 \leq \mathrm{e}^{-2}$ and $\lambda_0(1-\theta)/\theta > 1$, we have $\log\left[\frac{\lambda_0(1-\theta)}{\lambda_1\theta}\right] > 2$. Thereby

$$\exp(-\delta_\theta\lambda_1) = \exp\left\{-\frac{\lambda_1}{\lambda_0-\lambda_1}\log\left[\frac{\lambda_0(1-\theta)}{\lambda_1\theta}\right]\right\} \leq \exp\left(-\frac{2\lambda_1}{\lambda_0}\right) \leq \left(1 - \frac{\lambda_1}{\lambda_0}\right)$$

Altogether, we obtain

$$\mathsf{P}(|\beta| > \delta_\theta) \leq \theta \left(1 - \frac{\lambda_1^2}{\lambda_0^2}\right) < \theta. \quad \square$$

To continue with the proof of Theorem 2.2 we denote by $\pi_k = \mathsf{P}[|\beta_{jk}| > \delta(\lambda_{0k}, \lambda_1, \theta_{(k)})]$. By Lemma G.2, we have $\pi_k < \theta_{(k)}$ and thus

$$\mathsf{P}[K(\boldsymbol{B}) \leq k \,|\, \boldsymbol{\theta}] = \prod_{l>k}(1-\pi_l)^G > \prod_{l>k}(1-\theta_{(l)})^G = \mathsf{P}(K^+ \leq k \,|\, \boldsymbol{\theta}),$$

where K^+ is the actual dimension of the IBP process. Thereby, with $\alpha = c/G$ we can use Theorem G.1 to obtain

$$\mathsf{P}[K(\boldsymbol{B}) > k] \leq \mathsf{P}(K^+ > k) \leq 2\left[G(\alpha+1) + \frac{4}{3}\right]\left(\frac{\alpha}{\alpha+1}\right)^{k+1} < D\,\mathrm{e}^{-k\log(1+G/c)} \quad \square$$

G.2 Proof of Theorem 2.3

We follow the general recipe of Pati et al. (2014) and suitably modify their approach for the infinite-dimensional SSL-IBP prior. Denote by $|\boldsymbol{\beta}|$ the l_1 vector norm and by $|\boldsymbol{B}|$ the $_1$ matrix norm. Without loss of generality, we will show the Theorem assuming $c = 1$ and $S_{1n} = \cdots = S_{K_{0n}n} = S_n < G_n/2$. Let s_{max} (resp. s_{min}) denote the largest (resp. smallest) eigenvalue of $\boldsymbol{\Lambda}_0$, the true covariance matrix, and let $\rho_n = 2\, s_{max}/s_{min}$. We will need a couple of auxiliary lemmata. For simplicity of notation, we assume that $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots)'$ is the vector of *ordered* inclusion probabilities.

Lemma G.3. *Assume $\beta \sim$ SSL(λ_0, λ_1) with $\theta \in (0,1)$. For $\varepsilon > 0$ and $n > 0$, we have*

$$\mathsf{P}\left(|\beta| \leq \frac{\varepsilon}{\sqrt{n}} \,|\, \theta\right) > (1-\theta)\left(1 - \frac{1}{1+\lambda_0^2\varepsilon^2/n}\right).$$

Proof. Using $\exp(-x) < \frac{1}{1+x^2}$ for $x > 0$, we obtain

$$\begin{aligned}
\mathsf{P}\left(|\beta| \leq \frac{\varepsilon}{\sqrt{n}} \,|\, \theta\right) &= 1 - \left[\theta \exp\left(-\frac{\varepsilon}{\sqrt{n}}\lambda_1\right) + (1-\theta)\exp\left(-\frac{\varepsilon}{\sqrt{n}}\lambda_0\right)\right] \\
&> (1-\theta)\left[1 - \exp\left(-\lambda_0\,\varepsilon/\sqrt{n}\right)\right] \\
&> (1-\theta)\left(1 - \frac{1}{1+\lambda_0^2\varepsilon^2/n}\right). \qquad \square
\end{aligned}$$

Lemma G.4. *Suppose* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ *and* $\beta_i \overset{\text{ind}}{\sim}$ *SSL*(λ_0, λ_1) *with* $\mathsf{P}(\gamma_i = 1) = \theta$ *for* $i = 1, \ldots, p$. *Assume* $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ *with a nonzero support* $S_0 \subset \{1, \ldots, p\}$, *where* $0 < |S_0| = s < p/2$. *Let* $\varepsilon > 0$ *and assume* $\lambda_0^2 > 2p^2k^a/(s\,\varepsilon^2)$ *for some* $k \in \mathbb{N}$ *and* $a \geq 0$. *Then*

$$\mathsf{P}(||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 < \varepsilon \,|\, \theta) > \theta^s(1-\theta)^{p-s}\mathrm{e}^{-\lambda_1|\boldsymbol{\beta}_0|}\mathrm{e}^{-\lambda_1\varepsilon/\sqrt{2}}\left(\frac{\lambda_1\varepsilon}{4\,\mathrm{e}\,s}\right)^s \mathrm{e}^{-2s/k^a}.$$

Proof. Let $\boldsymbol{\beta}_{S_0}$ denote the s-dimensional sub-vector of $\boldsymbol{\beta}$ containing only entries in S_0. We have

$$\mathsf{P}\left(||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 < \varepsilon \,|\, \theta\right) \geq \left[\mathsf{P}\left(|\beta_1| < \frac{\varepsilon}{\sqrt{2\,p}} \,|\, \theta\right)\right]^{p-s} \mathsf{P}(||\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}||_2 < \varepsilon/\sqrt{2} \,|\, \theta). \tag{G.4}$$

Focusing on the first term in (G.4), we deploy Lemma G.3 to obtain for $s < p/2$

$$\mathsf{P}\left(|\beta_1| < \frac{\varepsilon}{\sqrt{2\,p}} \,|\, \theta\right) > (1-\theta)\left(1 - \frac{1}{1 + \lambda_0^2\varepsilon^2/(2p)}\right) > (1-\theta)\left(1 - \frac{s}{pk^a}\right) > (1-\theta)\mathrm{e}^{-2\,s/(p\,k^a)}. \tag{G.5}$$

We now focus on the second term in (G.4). Denote by $\widetilde{\Pi}(\boldsymbol{\beta}_{S_0}) = \left(\frac{\lambda_1}{2}\right)^s \mathrm{e}^{-\lambda_1|\boldsymbol{\beta}_{S_0}|}$ the Laplace product prior with a penalty λ_1. Conditionally on θ, the SSL prior density $\Pi(\boldsymbol{\beta}_{S_0} \,|\, \theta)$ satisfies $\Pi(\boldsymbol{\beta}_{S_0} \,|\, \theta) \geq \theta^s\,\widetilde{\Pi}(\boldsymbol{\beta}_{S_0})$. Thus we can write

$$\mathsf{P}(||\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}||_2 < \varepsilon/\sqrt{2} \,|\, \theta) \geq \theta^s\,\widetilde{\mathsf{P}}(||\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}||_2 < \varepsilon/\sqrt{2}),$$

where $\widetilde{\mathsf{P}}(\cdot)$ denotes the probability measure wrt $\widetilde{\Pi}(\cdot)$. As in Lemma 2 of Castillo et al. (2015), we deploy a change of variables $\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0} \to \boldsymbol{b}$ to obtain

$$\widetilde{\mathsf{P}}(||\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}||_2 < \varepsilon/\sqrt{2}) \geq \mathrm{e}^{-\lambda_1|\boldsymbol{\beta}_0|}\widetilde{\mathsf{P}}(|\boldsymbol{b}| \leq \varepsilon/\sqrt{2}) \geq \mathrm{e}^{-\lambda_1|\boldsymbol{\beta}_0|}\mathrm{e}^{-\lambda_1\varepsilon/\sqrt{2}}\left(\frac{\lambda_1\varepsilon}{\sqrt{2}}\right)^s \frac{1}{s!}. \tag{G.6}$$

The last inequality follows from observing that $\widetilde{\mathsf{P}}(||\boldsymbol{b}|| \leq \varepsilon/\sqrt{2})$ is the probability that first s events of a Poisson process of intensity λ_1 occur before time $\varepsilon/\sqrt{2}$ (Castillo et al. (2015), proof of Lemma 2). The statement of the Lemma then follows from (G.4) by combining displays (G.5) and (G.6). □

Lemma G.5. *Assume* $\boldsymbol{B} \sim$ *SSL-IBP* $(\{\lambda_{0k}\}, \lambda_1, \alpha_n)$ *with* $\alpha_n = 1/G_n$, $\lambda_1 < \mathrm{e}^{-2}$ *and* $\lambda_{0k}^2 \geq 2G_n^2k^4/(S_n\,\varepsilon^2)$ *for some* $0 < \varepsilon < 1$. *Then*

$$\mathsf{P}(||\boldsymbol{B} - \boldsymbol{B}_0||_F < \varepsilon) > \exp\left\{-C_1 - \lambda_1|\boldsymbol{B}_0| - C_2(K_0S_n + 1)[1 + \log(K_0G_n/\varepsilon^2)]\right\}$$

Proof. Because $\varepsilon^2 = \frac{6}{\pi^2}\sum_{k=1}^{\infty}\frac{\varepsilon^2}{k^2}$ we have

$$\mathsf{P}(||\boldsymbol{B}-\boldsymbol{B}_0||_F < \varepsilon \,|\, \boldsymbol{\theta}) > \prod_{k=1}^{\infty}\mathsf{P}\left(||\boldsymbol{\beta}^k-\boldsymbol{\beta}_0^k||^2 < \frac{6\,\varepsilon^2}{\pi^2 k^2}\,\middle|\,\theta_k\right).$$

Now, for each $1 \le k \le K_{0n}$, we can invoke Lemma G.4 to obtain

$$\mathsf{P}\left(||\boldsymbol{\beta}^k-\boldsymbol{\beta}_0^k||_2 < \frac{\sqrt{6}\,\varepsilon}{\pi k}\,\middle|\,\theta_k\right) > \theta_k^{S_n}(1-\theta_k)^{G_n-S_n}\mathrm{e}^{-\lambda_1|\boldsymbol{\beta}_0^k|}\mathrm{e}^{-\lambda_1\sqrt{3}\varepsilon/(\pi k)}\left(\frac{\lambda_1\sqrt{6}\,\varepsilon}{\pi k\,4\mathrm{e}S_n}\right)^{S_n}\mathrm{e}^{-2\,S_n/k^4} \tag{G.7}$$

Similarly as in (G.5), for each $k > K_{0n}$ we obtain

$$\mathsf{P}\left(||\boldsymbol{\beta}^k||_2 < \frac{\sqrt{6}\,\varepsilon}{\pi k}\,\middle|\,\theta_k\right) > (1-\theta_k)^{G_n}\mathrm{e}^{-2\,S_n/k^2}$$

Altogether, we obtain

$$\begin{aligned}\mathsf{P}(||\boldsymbol{B}-\boldsymbol{B}_0||_F < \varepsilon) > &\exp\left(-\lambda_1|\boldsymbol{B}_0| - \lambda_1\varepsilon H_n - K_{0n}S_n\left[2+\log\left(\frac{\pi\,4\,\mathrm{e}}{\sqrt{6}\lambda_1}\right)\right] - 2S_n\sum_{k>K_{0n}}\frac{1}{k^2}\right)\times\\ &\times\left(\frac{\varepsilon}{K_{0n}S_n}\right)^{K_{0n}S_n}\int\prod_{k=1}^{K_{0n}}\theta_k^{S_n}(1-\theta_k)^{G_n-S_n}\prod_{k>K_{0n}}(1-\theta_k)^{G_n}\mathrm{d}\,\Pi(\boldsymbol{\theta}),\end{aligned} \tag{G.8}$$

where H_n is the n^{th} harmonic number. To cope with the integral in (G.8), note that $(1-\theta_k) > (1-\theta_k/\theta_{K_{0n}})$ for $k > K_{0n}$. Since $\theta_1 > \theta_2 > \cdots$ we can lower-bound the integrand by

$$\theta_{K_{0n}}^{K_{0n}S_n}(1-\theta_1)^{K_{0n}G_n-K_{0n}S_n}\prod_{k>K_{0n}}\left(1-\frac{\theta_k}{\theta_{K_{0n}}}\right)^{G_n} \tag{G.9}$$

For each $k > K_{0n}$, we use the stick-breaking representation $\theta_k = \prod_{l=1}^{k}\nu_l$ and note that $\left(1-\frac{\theta_k}{\theta_{K_{0n}}}\right) = \left(1-\prod_{l=K_{0n}+1}^{k}\nu_l\right)$ does not depend on ν_l for $l \le K_{0n}$. Since ν_l's are iid, we can regard $\{\theta_k/\theta_{K_{0n}}\}$ as a lagged stick-breaking process, with the same distribution as $\{\theta_{k-K_{0n}}\}$. Using (G.9), the integral in (G.8) can be bounded from below by

$$\left[\int\nu_1^{K_{0n}S_n}\mathrm{d}\,\pi(\nu_1)\right]^{K_{0n}-1}\left[\int\nu_1^{K_{0n}S_n}(1-\nu_1)^{K_{0n}(G_n-S_n)}\mathrm{d}\,\pi(\nu_1)\right]\mathsf{E}\left[\prod_{k=1}^{\infty}(1-\theta_k)^{G_n}\right] \tag{G.10}$$

Now, with $\alpha_n = 1/G_n < 1$, we have $\int\nu_1^{K_{0n}S_n}\mathrm{d}\,\pi(\nu_1) > \frac{1}{G_n}\int_0^1\nu_1^{K_{0n}S_n}\mathrm{d}\,\nu_1 = \frac{1}{G_n(K_{0n}S_n+1)}$ and

$$\int\nu_1^{K_{0n}S_n}(1-\nu_1)^{K_{0n}(G_n-S_n)}\mathrm{d}\,\pi(\nu_1) > \frac{1}{G}\int_0^{S_n/G_n}\nu_1^{K_{0n}S_n}(1-\nu_1)^{K_{0n}(G_n-S_n)}\mathrm{d}\,\nu_1 \tag{G.11}$$

$$> \frac{\mathrm{e}^{-2K_{0n}S_n}}{G_n(K_{0n}S_n+1)}\left(\frac{S_n}{G_n}\right)^{K_{0n}S_n+1}. \tag{G.12}$$

Now, from the proof of Theorem 2.2 we obtain for $\alpha = 1/G$

$$\mathsf{E}\left[\prod_{k=1}^{\infty}(1-\theta_k)^{G_n}\right] > \exp\left\{-2\left[G_n(\alpha+1)-\frac{4}{3}\right]\left(\frac{\alpha}{\alpha+1}\right)\right\} > \mathrm{e}^{-2}. \tag{G.13}$$

Altogether, we have

$$\begin{aligned}\mathsf{P}(||\boldsymbol{B}-\boldsymbol{B}_0||_F < \varepsilon) > &\exp\left(-2-\lambda_1|\boldsymbol{B}_0|-\lambda_1\varepsilon H_n - K_{0n}S_n\left[8+\log\left(\frac{\pi}{\sqrt{6}\lambda_1}\right)\right]\right)\times \\ &\times\left(\frac{\varepsilon}{K_{0n}S_n}\right)^{K_{0n}S_n}\left(\frac{S_n}{G_n}\right)^{K_{0n}S_n+1}\left[\frac{1}{G(K_{0n}S_n+1)}\right]^{K_{0n}} \\ &> \exp\left\{-C_1-\lambda_1|\boldsymbol{B}_0|-C_2(K_{0n}S_n+1)[1+\log(K_{0n}G_n/\varepsilon)]\right\}.\end{aligned}$$

□

Denote by $\mathcal{B} = \{\boldsymbol{B} : K(\boldsymbol{B}) > D\, S_0 K_{0n}\}$. The posterior probability assigned to $\mathcal{B}$ is given by

$$P(\mathcal{B} \mid \boldsymbol{Y}^{(n)}) = \frac{\int_{\mathcal{B}}\prod_{i=1}^n f_\Sigma(y_i)/f_{\Sigma_0}(y_i)\mathrm{d}\,\Pi(\Sigma)}{\int\prod_{i=1}^n f_\Sigma(y_i)/f_{\Sigma_0}(y_i)\mathrm{d}\,\Pi(\Sigma)} \equiv \frac{N_n}{D_n}$$

where f_Σ denotes the p_n-dimensional $\mathcal{N}(0,\Sigma)$ distribution. Following Pati et al. (2014), we confine attention to the set $\mathcal{A}_n \in \sigma(\boldsymbol{y})$, on which

$$D_n \geq \mathrm{e}^{-C\,K_{0n}S_n\log(\rho_n)/s_{min}^2}\mathsf{P}(\Sigma : ||\Sigma-\Sigma_0||_F < \sqrt{S_nK_{0n}/n}),$$

Under the assumption $\sqrt{S_nK_{0n}}/s_{min} \to \infty$ and $\sqrt{S_nK_{0n}/n}/s_{min} \to 0$, Pati et al. (2014) (Lemma 9.1) show that $\mathsf{P}_{\Sigma_0}(\mathcal{A}^c) \to 0$. Now, we can write

$$\mathsf{E}_{\Sigma_0}\left[P(\mathcal{B} \mid \boldsymbol{Y}^{(n)})\mathbb{I}_{\mathcal{A}}\right] \leq \frac{\mathsf{P}(\mathcal{B})}{\mathrm{e}^{-C\,K_{0n}S_n\log(\rho_n)/s_{min}^2}\mathsf{P}(\Sigma : ||\Sigma-\Sigma_0||_F < \sqrt{S_nK_{0n}/n})} \tag{G.14}$$

To bound the numerator in (G.14), we use Theorem 2.2 to find $\mathsf{P}(\mathcal{B}) \preceq \mathrm{e}^{-D\,K_{0n}S_n\log(G+1)}$ Next, we need to find a lower-bound to the denominator in (G.14). Now, we use the following fact (as in the proof of Lemma 9.2. of Pati et al. (2014))

$$||\boldsymbol{B}-\boldsymbol{B}_0||_F < \varepsilon \Longrightarrow ||\boldsymbol{B}\boldsymbol{B}'-\boldsymbol{B}_0\boldsymbol{B}_0||_F < (2||\boldsymbol{B}_0||_2+\varepsilon)\,\varepsilon$$

With $\varepsilon = \frac{1}{4}\sqrt{K_0/n} < 1$ and assuming $||\boldsymbol{B}_0||_2 < \sqrt{S_n}$ we have $(2||\boldsymbol{B}_0||_2+\varepsilon)\,\varepsilon < \sqrt{S_nK_0/n}$. Thus

$$\mathsf{P}(\Sigma : ||\Sigma-\Sigma_0||_F < \sqrt{S_nK_{0n}/n}) > P(\boldsymbol{B} : ||\boldsymbol{B}-\boldsymbol{B}_0||_F < \sqrt{K_{0n}/n}/4).$$

Because $|\boldsymbol{B}_0| \leq \sqrt{S_n}||\boldsymbol{B}_0||_F \leq S_n K_{0n}$ by assumption (C), with Lemma G.5 for $n < G$ we find that $P(\boldsymbol{B} : ||\boldsymbol{B} - \boldsymbol{B}_0||_F < \sqrt{K_{0n}/n}/4) \succeq \mathrm{e}^{-C(K_{0n}S_n+1)[1+\log(G_n)]}$. Thus, (G.14) can be bounded as follows

$$\mathsf{E}_{\Sigma_0}\left[P(\mathcal{B} \mid \boldsymbol{Y}^{(n)})\mathbb{I}_{\mathcal{A}}\right] \preceq \frac{\mathrm{e}^{-DK_{0n}S_n\log(G_n+1)}}{\mathrm{e}^{-C\,K_{0n}S_n\log(\rho_n)/s_{min}^2}\mathrm{e}^{-C(K_{0n}S_n+1)[1+\log(G_n)]}} \tag{G.15}$$

By assumption (B) $\frac{\sqrt{S_nK_{0n}}}{s_{min}\sqrt{n}} < 1$ for n large enough. Therefore $\rho_n = 2\frac{s_{max}\sqrt{n}}{\sqrt{S_nK_{0n}}}\frac{\sqrt{S_nK_{0n}}}{s_{min}\sqrt{n}} < 2\sqrt{n}\frac{||\boldsymbol{B}_0||_2}{\sqrt{S_nK_{0n}}} < 2\sqrt{n/K_0}$. Thereby, $K_{0n}S_n\log(\rho_n)/s_{min}^2 < n\log(2\sqrt{n}) \preceq S_nK_{0n}\log(G_n+1)$ by assumption (A). In conclusion, for large enough $D > 0$, (G.15) goes to zero as $n \to \infty$. □

References

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015), "Bayesian linear regression with sparse priors," *Annals of Statistics (to appear)*.

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

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

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

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

Meinshausen, N. (2007), "Relaxed lasso," *Computational Statistics and Data Analysis*, pages 374–393.

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

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

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

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

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