Bayesian Analysis (2022) **17**, Number 1, pp. 33–66

Heterogeneous Large Datasets Integration Using Bayesian Factor Regression

Alejandra Avalos-Pacheco[*,†,¶], David Rossell[‡,‖], and Richard S. Savage[§]

Abstract. Two key challenges in modern statistical applications are the large amount of information recorded per individual, and that such data are often not collected all at once but in batches. These batch effects can be complex, causing distortions in both mean and variance. We propose a novel sparse latent factor regression model to integrate such heterogeneous data. The model provides a tool for data exploration via dimensionality reduction and sparse low-rank covariance estimation while correcting for a range of batch effects. We study the use of several sparse priors (local and non-local) to learn the dimension of the latent factors. We provide a flexible methodology for sparse factor regression which is not limited to data with batch effects. Our model is fitted in a deterministic fashion by means of an EM algorithm for which we derive closed-form updates, contributing a novel scalable algorithm for non-local priors of interest beyond the immediate scope of this paper. We present several examples, with a focus on bioinformatics applications. Our results show an increase in the accuracy of the dimensionality reduction, with non-local priors substantially improving the reconstruction of factor cardinality. The results of our analyses illustrate how failing to properly account for batch effects can result in unreliable inference. Our model provides a novel approach to latent factor regression that balances sparsity with sensitivity in scenarios both with and without batch effects and is highly computationally efficient.

Keywords: Bayesian factor analysis, EM, non-local priors, shrinkage.

MSC2020 subject classifications: Primary 62-07, 62F15; secondary 62P10.

1 Introduction

A first important task when dealing with large datasets is to conduct an exploratory analysis. Dimensionality reduction techniques have proven a highly popular tool for this purpose. Those techniques provide a lower-dimensional representation that can give insights into the underlying structure to visualise, denoise or extract meaningful features from the data. See Johnson and Wichern (1988, chap. 3) or Hastie et al. (2001,

[*] Harvard-MIT Center for Regulatory Science, Harvard Medical School, 210 Longwood Av, Boston, MA, USA, aavalos@jimmy.harvard.edu

[†] Dept. of Statistics, University of Warwick, Coventry CV4 7AL, UK

[‡] Dept. of Business and Economics, Universitat Pompeu Fabra, Carrer de Ramon Trias Fargas, 25-27, 08005 Barcelona, Spain, david.rossell@upf.edu

[§] Dept. of Statistics, University of Warwick, Coventry CV4 7AL, UK, r.s.savage@warwick.ac.uk

[¶] Supported by *the Mexican National Council of Science and Technology (CONACYT) grant no. CVU5464444* and the Harvard-MIT Center for Regulatory Science.

[‖] Partially funded by the *NIH grant R01 CA158113-01, RyC-2015-18544, Ayudas Fundación BBVA a equipos de investigación científica 2017* and *Spanish Plan Estatal grant PGC2018-101643-B-I00*.

https://doi.org/10.1214/20-BA1240

chap. 14) for a gentle introduction and Burges (2010); Cunningham and Ghahramani (2015) for more recent reviews.

Large datasets are common in modern statistical applications. For instance, technological advances in bioinformatics such as high-throughput sequencing, microarrays, mass spectrometry and single cell genomics allow the gathering of a vast amount of biological data, enabling researchers to create models to explain the complex processes and interactions of biological systems (see Bersanelli et al. (2016) for a recent review). Cancer is a prominent example. Large-scale projects such as The Cancer Genome Atlas (TCGA), Cancer Genome Project (CGP) and the International Cancer Genome Consortium (ICGC), as well as many individual laboratories are generating extensive amounts of biological data (e.g. gene expression, mutation annotation, DNA methylation profiles, copy number changes) in addition to recording other covariates (e.g. gender, tumour stage, medical treatment and patient history). These projects aim to give a better understanding of the disease and improve prognosis, prevention and treatment. However, the large and heterogeneous nature of the data make the analyses and interpretations challenging. Furthermore, such data are often generated under different experimental conditions, when new samples are incrementally added to existing samples, or in analyses coming from different projects, laboratories, or platforms; collecting data in this matter often produces batch effects (Rhodes et al., 2004). These, unless properly adjusted for, may lead to incorrect conclusions (Leek et al., 2010; Goh et al., 2017). In the context of bioinformatics, several approaches have been developed for removing batch effects (see Scherer (2009) for a review and examples). These include data "normalization" methods using control metrics or regression methods (Schadt et al., 2001; Yang et al., 2002), matrix factorisation (Alter et al., 2000; Benito et al., 2004) and location-scale methods (Leek and Storey, 2007; Johnson and Li, 2009; Parker et al., 2014; Hornung et al., 2016). Strategies for batch effect correction include data preprocessing, for example via the so-called ComBat empirical Bayes approach (Johnson et al., 2007) or via singular value decomposition (SVD) (Leek and Storey, 2007). As shown in our examples applying standard dimension reduction methods on such normalised data can produce unreliable results. Intuitively this is due to using a two-step rather than a joint inference procedure on batch effects and dimension reduction. Our examples focus on cancer-related gene expression; nonetheless, batch effects are also present in many other settings, e.g. structural magnetic resonance imaging (MRI) data from Alzheimer's disease (Shinohara et al., 2014; Fortin et al., 2016), multiple sclerosis (Shah et al., 2011), attention deficit hyperactivity disorder (Olivetti et al., 2012) or even different tissues of marine mussels (Avio et al., 2015).

We address dimensionality reduction via a model-based framework relying on Bayesian factor analysis and latent factor regression. Our model builds on the approaches introduced by Lopes and West (2004); Lucas et al. (2006); Carvalho et al. (2008) and Ročková and George (2017). An important practical extension of these works is to increase the flexibility to account for systematic biases or sources of variation that do not reflect any underlying patterns of interest, i.e. batch effects. Our main contribution is to provide a model-based approach for tackling dimensionality reduction and batch effect correction simultaneously, avoiding the use of two-step procedures. Another important contribution is to develop a scalable non-local prior based formulation to induce sparsity and learn the underlying number of factors; for this we provide a prior parameter elicitation, of practical importance to increase power to detect non-zero loadings. A strategy related to ours is to use factor models to learn, on the one hand, the biological patterns via common factors shared across the different data sources and, on the other hand, the non-common sources of variation via data-specific factors (De Vito et al., 2018b,a). However, such a strategy is not designed for batch effects and requires MCMC estimation, making the inference slower. Another related approach is to regress the covariance on batches and other explanatory variables, either parametrically or non-parametrically (Hoff and Niu, 2012; Fox and Dunson, 2015). While useful, this method is not focused on dimension reduction and does not lead to sparse factor loadings that facilitate interpretation and, as shown in our examples, can improve inference.

We model observations with a regression on latent factors with sparse loadings, observed covariates, and batch effects that can alter the mean and intrinsic variance structures. Model fitting is done via a novel Expectation-Maximisation (EM) algorithm to obtain maximum posterior mode parameter estimates in a computationally efficient manner. We focus on three different continuous prior formulations for the loadings: flat, Normal-spike-and-slab (George and McCulloch, 1993) and a novel Normal-spike-and-MOM-slab, based on a continuous relaxation of the non-local prior configuration by Johnson and Rossell (2010, 2012). We also discuss non-local Laplace-tailed extensions, along the lines of Ročková and George (2017). Spike-and-slab priors provide sparse loadings, effectively performing model selection on the number of required factors and non-zero loadings. We obtain closed-form EM updates, a novel contribution to the non-local prior literature. As we will discuss later, the main advantage of non-local priors in this setting is to help achieve a better balance between sparsity and sensitivity in inferring non-zero loadings. To our knowledge, this is the first adaptation of non-local priors to factor models. See also Bar et al. (2018) who argued for improved sensitivity via 3-component mixture priors that resemble non-local priors in generalised linear models, and Shi et al. (2019) for an application to linear regression via Gibbs sampling.

Our work is meant to contribute to substantive applied aspects in data analysis that we show via examples to be of practical relevance. Firstly, as far as dimension reduction is concerned, we provide sparse solutions that increase the quality of our estimations (see Section 5.1). Secondly, with regard to batch effect correction, we model dependence structure across batches. As a motivating example, Figure 5 (top row) displays systematic differences in mean and variance, thus revealing the problem of not accounting for batch effects. These batches represent comparable patients in terms of disease characteristics, hence we view them as exchangeable realisations from a common distribution and as such they should show the same distribution in any low dimensional projection. After two-step procedures most of these differences are corrected, but distinct covariances are still present across batches (see rows 2 and 3). Thirdly, as a prediction tool for survival analysis in ovarian, lung and colon cancer datasets, we provide competitive concordance indexes (see Section 5.3). Furthermore, our method could be of use for downstream analyses. Finally, from the computational viewpoint, via our model-based approach we study high-dimensional sparse models that facilitate the use of non-local priors to applications.

The outline of this paper is as follows. Section 2 reviews latent factor regression and introduces our extension, which includes a variance batch effect adjustment. Section 3 proposes prior formulations including non-local priors on the loadings and important aspects related to prior parameter elicitation. Section 4 describes several EM algorithms for model fitting, parameter initialisation and post-processing steps required for effective model selection and dimension reduction. Section 5 presents applications on simulations and on cancer datasets under unsupervised and supervised settings. Section 6 concludes. The supplementary material contains the derivation of the EM algorithm and additional results. Software implementing our methodology is available at https://github.com/AleAviP/BFR.BE.

2 Latent factor regression with batch effects

Consider vectors $\mathtt{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p$, observed for $i = 1, \ldots, n$ individuals. The factor regression model defines $\mathtt{x}_i$ as a regression on p_v observed covariates denoted by $\mathtt{v}_i \in \mathbb{R}^{p_v}$, and q low-dimensional latent variables denoted $\mathtt{z}_i \in \mathbb{R}^q$, also known as latent coordinates or factors. Let X be the $n \times p$ matrix with the i^{th} row equal to $\mathtt{x}_i^\top$, V the $n \times p_v$ matrix of known covariates with the i^{th} row equal to $\mathtt{v}_i^\top$ and Z the $n \times q$ matrix of latent coordinates, containing $\mathtt{z}_i^\top$ in the i^{th} row. The standard factor regression model is

$$\mathtt{x}_i = \theta \mathtt{v}_i + M\mathtt{z}_i + \mathtt{e}_i, \tag{2.1}$$

where $\theta \in \mathbb{R}^{p \times p_v}$ is the matrix of regression coefficients, $M \in \mathbb{R}^{p \times q}$ is the matrix of factor loadings and $\mathtt{e}_i \in \mathbb{R}^p$ is the error, distributed as $\mathtt{e}_i \sim N(0, \Sigma)$ independently across $i = 1, \ldots, n$, where Σ is a diagonal matrix. Factors are assumed to be standard normal, $\mathtt{z}_i \sim N(0, \mathbf{I})$, independent across $i = 1, \ldots, n$ and also independent of $\mathtt{e}_i$.

Equation (2.1) regresses the observed data X on known covariates and on a latent factor structure. In particular, it allows additive batch effects to be accounted for by incorporating the variables recording the batches into $\mathtt{v}_i$. However, in practice one often observes more complex batch effects; specifically in bioinformatics it is common to observe multiplicative effects on the variance (Johnson et al., 2007). We will later describe an example of this, shown in Figure 5. Such artefacts cannot be captured by (2.1) given that Σ is assumed constant across all individuals.

To address this issue we extend (2.1) by allowing Σ to depend on i. Suppose the data were obtained in p_b batches, e.g. from different days, laboratories or instrumental calibrations, with n_l individuals in batch l, for $l = 1, \ldots, p_b$, such that $n_1 + n_2 + \cdots + n_{p_b} = n$. Let $\mathtt{b}_i$ be the indicator vector of length p_b defined as $b_{il} := 1$ if individual i is in batch l, $b_{il} := 0$ otherwise.

We incorporate batch effects by adding a mean and variance adjustment. We let

$$\mathtt{x}_i = \theta \mathtt{v}_i + M\mathtt{z}_i + \beta \mathtt{b}_i + \mathtt{e}_i, \tag{2.2}$$

where θ, $\mathtt{v}_i$, M and $\mathtt{z}_i$ are as (2.1), $\beta \in \mathbb{R}^{p \times p_b}$ captures additive batch effects and the variance of $\mathtt{e}_i$ captures multiplicative batch effects. We denote by τ_{jl}, $j = 1, \ldots, p$ and

$l = 1, \ldots, p_b$ as the j^{th} idiosyncratic precision element in batch l. Then, given $b_{il} = 1$, the errors are independently distributed as $\mathtt{e}_{ij} \sim N(0, \tau_{jl}^{-1})$. Further, denote by $\mathcal{T}$ the $p \times p_b$ matrix that has τ_{jl} as its (j,l) element.

To help interpret the practical implications of the model, suppose that one has orthonormal factor loadings $M^\top M = \mathbf{I}$. Then (2.2) implies

$$\mathtt{z}_i = M^\top (\mathtt{x}_i - (\theta \mathtt{v}_i + \beta \mathtt{b}_i + \mathtt{e}_i)) \tag{2.3}$$

and thus, $\mathbb{E}(\mathtt{z}_i \mid \mathtt{x}_i, \mathtt{v}_i, \mathtt{b}_i, M, \theta, \beta) = M^\top \mathtt{x}_i - M^\top \theta \mathtt{v}_i - M^\top \beta \mathtt{b}_i$. That is, the mean of the latent coordinates is the projection $M^\top \mathtt{x}_i$ plus a translation given by the batch effect adjustment and (potentially) the observed covariates. An interesting observation is that their covariance $\mathrm{Cov}(\mathtt{z}_i \mid \mathtt{x}_i, \mathtt{v}_i, \mathtt{b}_i, M, \theta, \beta, \mathcal{T}) = M^\top \mathcal{T}_{\mathtt{b}_i}^{-1} M$ depends on the batch b_i. As an example, the top-left panel in Figure 5 shows the first two factors of an ovarian dataset that contains two batches. The latent coordinates of these batches exhibit a different mean and variance. The middle-left panels show the results after applying the method ComBat to standardise the mean and variance across batches. After ComBat correction, the latent coordinates exhibit the same mean and variance across batches (as expected) but a very different covariance structure. Given that patients in both batches are believed to be roughly exchangeable, such a difference in covariance is likely due to technical artifacts and to a two-step procedure that fits the factor model separately from the adjustment step. The bottom-left panel illustrates that estimating $\mathcal{T}$ jointly with (M, θ, β) via our proposed methods addresses this issue; in particular, both batches exhibited similar mean, variance and covariance structure.

Model (2.2) can be represented in matrix notation as

$$X = V\theta^\top + ZM^\top + B\beta^\top + E, \tag{2.4}$$

where $E \in \mathbb{R}^{n \times p}$ is the matrix of errors.

The latent factor model is non-identifiable up to orthogonal transformations, of the form $M^{*\top} = A^\top M^\top$ and $Z^* = ZA$, where A is any orthogonal $q \times q$ matrix. Thus, the factor model in (2.4) can equivalently be rewritten as $X = V\theta^\top + Z^* M^{*\top} + B\beta^\top + E$. To obtain unique point estimates of M and Z, several alternative prior specifications have been developed. One option is restricting the parameter space. Seber (1984) constrained M such that $M^\top \Omega M$ is diagonal. Lopes and West (2004) restricted M to be lower-triangular with a strictly positive diagonal, $m_{jj} > 0$, and assumed M to be full-rank. More recently, Frühwirth-Schnatter and Lopes (2018) suggested a factor reordering via a Generalized Lower Triangular loading matrix. However, under this approach the interpretation of M depends on the arbitrary ordering of the columns in X, and it gives special roles to the first factors. Another option is to encourage sparsity in M, e.g. the classical varimax solution (Kaiser, 1958) maximises the variance in the squared rotated loadings. A more modern strategy is to favour sparse solutions containing exact zero loadings, e.g. Ročková and George (2017) proposed an EM algorithm that seeks rotations based on a so-called Parameter Expansion (PX) that aims to avoid local suboptimal regions. We adopt a similar strategy where sparse solutions are preferred by the introduced non-local penalties.

In this paper, the main reasons for considering a sparse M, rather than addressing identifiability, are the following: First, a sparse M facilitates interpretation of factors as linear combinations of a smaller set of variables. This motivation goes back to classical work on varimax rotations and more recent work on Ročková and George's rotations. These sparsity considerations are important in many applications to help assign a meaning to the latent coordinates. Second, the sparsity assumption, when warranted, has significant potential gains for estimation accuracy. For illustration, the sparse scenarios in Table 1 show that MOM-SS attains low errors when estimating both the covariance and the expected value of the response variable, and requires significantly less factors than our competitors. Of course, if the sparsity assumption is not warranted then there is the risk of over-enforcing sparsity, but one may overcome this issue by carefully eliciting the prior parameters, as our results show. For instance, in our ovarian cancer example MOM-SS chose a sparse M and achieved a concordance index (CI) similar to ComBat-MLE but with considerably less factors (4 instead of 101). On the other hand, in the lung and colon cancer settings, MOM-SS selected a non-sparse M (74 and 53 factors respectively), achieving competitive CI for lung and the highest CI for colon cancer. Overall, MOM-SS provided a more stable performance than competing methods in balancing sparsity and prediction accuracy. Third, inducing sparsity allows one to work with large q and let the data learn how many factors are needed. Namely, since in practice the value of q is unknown, one might consider large q, in which case the dimension of Z and M could be substantial. Thus, considering sparse solutions acts as a model selection tool to learn the number of factors from data.

3 Prior formulation

To complete Model (2.2) we set priors for the loadings M, precisions τ_{jl}, and regression parameters (θ, β). Through our proposed default prior formulation we assume that the columns in X have been centred to zero mean and unit variance. For the idiosyncratic precisions τ_{jl} we set

$$\tau_{jl} \mid \eta, \xi \sim \text{Gamma}(\eta/2, \eta\xi/2) \tag{3.1}$$

independently across $j = 1, \ldots, p$ and $l = 1, \ldots, p_b$. By default in our examples we set the fairly informative values $\eta = \xi = 1$, leading to diffuse though proper priors.

For the regression parameters we set

$$(\theta_j, \beta_j) \sim N(0, \psi \mathbf{I}), \quad j = 1, \ldots, p, \tag{3.2}$$

where ψ is a user-defined prior dispersion that in our examples by default we set to $\psi = 1$. The choice of $\psi = 1$ assigns the same marginal prior variances to elements in (θ_j, β_j) as the unit information prior often adopted as a default for linear regression (Schwarz, 1978).

We remark that this prior does not encourage sparsity in the regression parameters (θ, β) or factor loadings, which we view as reasonable provided the number of variables p_v and batches p_b are moderate. For large p_v or p_b, a direct extension of our prior on the loadings M could be adopted.

The loadings matrix M plays an important role in improving shrinkage and simplifying interpretation. Some recent strategies include a LASSO-based method (Witten et al., 2009), horseshoe priors (Carvalho et al., 2009), an Indian buffet process (Knowles and Ghahramani, 2011), an infinite factor model (Dunson and Bhattacharya, 2011) among others. In this paper, we consider three priors on the loadings: an improper flat prior $\mathsf{p}(M) \propto 1$, a Normal spike-and-slab and a novel non-local pMoM spike-and-slab. The local and non-local spike-and-slab prior formulations are detailed below, along with Laplace-based extensions. These build on the approach by Ročková and George (2014, 2017), our main contribution being the introduction of non-local-based variations.

3.1 Local spike-and-slab prior

A traditional Bayesian approach to variable selection is the spike-and-slab prior, a two-component mixture prior (Mitchell and Beauchamp, 1988; George and McCulloch, 1993). This prior aims to discriminate those loadings that warrant inclusion, modelled by the slab component, from those that should be excluded, modelled by the spike component.

Specifically, a spike-and-slab prior density for the loadings M has the form

$$\mathsf{p}(M \mid \gamma, \lambda_0, \lambda_1) = \prod_{j=1}^{p}\prod_{k=1}^{q}(1-\gamma_{jk})\mathsf{p}(m_{jk} \mid \lambda_0, \gamma_{jk}=0) + \gamma_{jk}\mathsf{p}(m_{jk} \mid \lambda_1, \gamma_{jk}=1), \quad (3.3)$$

where $\mathsf{p}(m_{jk} \mid \lambda_0, \gamma_{jk}=0)$ is a continuous density, λ_0 is a given dispersion parameter of the spike component and $\lambda_1 > \lambda_0$ is that of the slab component. The indicators $\gamma_{jk} \in \{0,1\}$ signal which m_{jk} were generated by each component, and serve as a proxy for which loadings are significantly non-zero. We take as a base formulation the Normal-spike-and-slab prior by George and McCulloch (1993) were the spike is a Normal density with a small variance λ_0 and the slab a Normal distribution with large variance λ_1. Although Laplace-Spike-and-Slab priors have been shown to possess better properties for sparse inference (Ročková and George, 2018), as discussed below the introduction of non-local penalties improves certain undesirable features of the Normal-based prior. The elicitation of λ_0 and λ_1 is an important aspect of the formulation and will be discussed in Section 3.3. Specifically, the Normal-spike-and-slab is

$$\mathsf{p}(m_{jk} \mid \gamma_{jk}=l, \lambda_l) = N(m_{jk}; 0, \lambda_l). \quad (3.4)$$

The continuity of the spike distribution gives closed form expressions for the EM algorithm, making it computationally appealing. We refer to (3.4) as Normal-SS.

We complete the model specification with a hierarchical prior over the latent indicator $\gamma = \{\gamma_{jk}, j=1,\dots,p, k=1,\dots,q\}$ as follows,

$$\begin{aligned}\gamma_{jk} \mid \zeta_k &\sim \text{Bernoulli}(\zeta_k),\\ \zeta_k \mid a_\zeta, b_\zeta &\sim \text{Beta}\left(\frac{a_\zeta}{k}, b_\zeta\right),\end{aligned} \quad (3.5)$$

with independence across (j,k) where $a_\zeta > 0$ and $b_\zeta > 0$ are given prior parameters. By default we set $a_\zeta = b_\zeta = 1$, which leads to a uniform prior for the first factor $(k=1)$,

$\zeta_k \mid a_\zeta, b_\zeta \sim \mathrm{U}(0,1)$. Furthermore, note that $\frac{a_\zeta}{k}$ encourages increasingly sparse solutions in subsequent factors. That is, related to our earlier discussion of non-identifiability (Section 2), we encourage loadings where the first factors have larger importance, leading to solutions that are sparse both in the rank of M and its non-zero entries.

A potential concern with Normal-SS is that the slab density assigns non-negligible probability to regions of the parameter space that are also consistent with the spike, namely when m_{jk} lies close to zero. We will address this via non-local priors and show that these, by enforcing separation between two components, help increase sensitivity.

3.2 Non-local spike-and-slab prior

Non-local priors (NLPs) are a family of distributions that assign vanishing prior density to a neighbourhood of the null hypothesis (Johnson and Rossell, 2010). Definition 3.1 is an adaptation of the definition in Johnson and Rossell (2010) to (3.3).

Definition 3.1. An absolutely continuous measure with density $\mathsf{p}(m_{jk}|\gamma_{jk} = 1)$ is a non-local prior if $\lim_{m_{jk}\to 0} \mathsf{p}(m_{jk}|\gamma_{jk} = 1) = 0$.

We call any prior not satisfying Definition 3.1 a local prior. Non-local priors possess appealing properties for Bayesian model selection. They discard spurious parameters faster as the sample size n grows, but preserve exponential rates to detect important coefficients (Johnson and Rossell, 2010; Fúquene et al., 2018) and can lead to improved parameter estimation shrinkage (Rossell and Telesca, 2017). To illustrate the motivation for NLPs in our setting consider Figure 1. Normal-SS assigns positive probability to $m_{jk} = 0$. Correspondingly, the conditional inclusion probability $\mathsf{p}(\gamma_{jk} = 1 \mid m_{jk})$ remains non-negligible, even when $m_{jk} = 0$ (lower left panel).

As an alternative, we consider a product moment (pMOM) prior (Johnson and Rossell, 2012).

$$\begin{aligned}\mathsf{p}(m_{jk} \mid \gamma_{jk} = 0, \tilde{\lambda}_0) &= \mathrm{N}(m_{jk}; 0, \tilde{\lambda}_0),\\ \mathsf{p}(m_{jk} \mid \gamma_{jk} = 1, \tilde{\lambda}_1) &= \frac{m_{jk}^2}{\tilde{\lambda}_1}\mathrm{N}(m_{jk}; 0, \tilde{\lambda}_1).\end{aligned} \tag{3.6}$$

We denote (3.6) as MOM-SS. This prior assigns zero density to $m_{jk} = 0$ given $\gamma_{jk} = 1$, which implies $\mathsf{p}(\gamma_{jk} = 1 \mid m_{jk} = 0) = 0$ (Figure 1). Prior elicitation for $\tilde{\lambda}_0$ and $\tilde{\lambda}_1$ is discussed in Section 3.3. From a computational point of view, the EM algorithm can accommodate this extension by using a trivial extra gradient evaluation at negligible additional cost relative to the Normal-SS. Parameter estimation and algebraic details are described in Section 4. The prior on the inclusion indicators is set as in (3.5).

Beyond (3.4) and (3.6), another natural extension is to use Laplace-based priors based on the Spike-and-Slab LASSO by Ročková and George (2018)

$$\mathsf{p}(m_{jk} \mid \gamma_{jk}, \lambda_0, \lambda_1) = (1-\gamma_{jk})\mathrm{Laplace}(m_{jk}; 0, \lambda_0) + \gamma_{jk}\mathrm{Laplace}(m_{jk}; 0, \lambda_1), \tag{3.7}$$

with a slab component with variance $2\lambda_0^2$, and a spike component with $2\lambda_1^2$, where $\mathrm{Laplace}(m_{jk}; 0, \lambda) = \frac{1}{2\lambda}\exp\left(\frac{-|m_{jk}|}{\lambda}\right)$. We refer to (3.7) as Laplace-SS. As illustrated in

Figure 1 (right panels) this prior can help encourage sparsity, setting $\mathsf{p}(\gamma_{jk} = 1 \mid m_{jk} = 0)$ to smaller values (though still non-zero) than the Normal-SS.

As an extension, akin to (3.6), one could set a moment penalty on the Laplace density.

$$\begin{aligned} \mathsf{p}(m_{jk} \mid \gamma_{jk} = 0, \tilde{\lambda}_0) &= \text{Laplace}(m_{jk}; 0, \tilde{\lambda}_0), \\ \mathsf{p}(m_{jk} \mid \gamma_{jk} = 1, \tilde{\lambda}_1) &= \frac{m_{jk}^2}{2\tilde{\lambda}_1^2}\text{Laplace}(m_{jk}; 0, \tilde{\lambda}_1). \end{aligned} \tag{3.8}$$

We denote (3.8) as Laplace-MOM-SS. Relative to (3.6), as illustrated in Figure 1, Laplace-MOM-SS leads to lower $\mathsf{p}(\gamma_{jk} = 1 \mid m_{jk} = 0)$ and higher $\mathsf{p}(\gamma_{jk} = 1 \mid m_{jk})$ for moderately large m_{jk}.

We discuss prior elicitation for Laplace-MOM-SS in Section 3.3 and derive an EM algorithm in Section 4.2 but in our examples we focus on the MOM-SS for simplicity. However, the Laplace-based (3.8) can also be shown to lead to closed-form EM updates.

3.3 Prior elicitation for the variance of the spike-and-slab priors

A crucial aspect in a spike-and-slab prior is the choice of the prior scale parameters. It is common to fix the variance of the spike distribution λ_0 to a value close to zero. Regarding λ_1, one option is to set a hyper-prior or to try to estimate it from the data (George and McCulloch, 1993, 1997; Ročková and George, 2014, 2018). Setting a hyper-prior does not bypass prior elicitation, as one then needs to set the hyper-prior parameters, whereas estimating λ_1 from the data increases the cost of computations. Instead, we capitalise on the fact that factor loadings have a natural interpretation in terms of the fraction of explained variance in X. Thus, we propose default values that dictate which coefficients are considered as meaningfully different from zero. These defaults are guidelines in the absence of a priori knowledge. A convenient feature of such an elicitation is that it can be easily extended to local priors and other non-Gaussian spike-and-slab priors.

Our goal is to find values $\tilde{\lambda}_0$ and $\tilde{\lambda}_1$ for the MOM-SS that distinguish practically relevant factors. In the absence of covariates, the factor model decomposes the total variance in variable j as $\text{Var}(\mathtt{x}_{ij}) = \sum_{k=1}^{q} m_{jk}^2 + \tau_{jj}^{-1}$, hence m_{jk}^2 is the proportion of variance in variable j explained by factor k. We take $m_{jk}^2 > 0.1$ as a threshold for practical relevance. Specifically, we set $\tilde{\lambda}_0$ such that $\mathsf{p}(|m_{jk}| \leq \sqrt{0.1} \mid \tilde{\lambda}_0) = 0.95$, that is $\tilde{\lambda}_0 = \frac{0.1}{(\Phi^{-1}(0.025))^2} \approx 0.026$, where Φ^{-1} denotes the standard normal quantile function. Likewise we set $\mathsf{p}(|m_{jk}| \geq \sqrt{0.1} \mid \tilde{\lambda}_1) = 0.95$ under the MOM-SS, obtaining the default $\tilde{\lambda}_1 \approx 0.2842$.

Regarding the Normal-SS prior, we set $\lambda_0 = \tilde{\lambda}_0$ and λ_1 such that it is comparable to the MOM-SS in terms of informativeness, namely it matches the variance of the MOM-SS, obtaining that $\lambda_1 = 3\tilde{\lambda}_1 \approx 0.8526$.

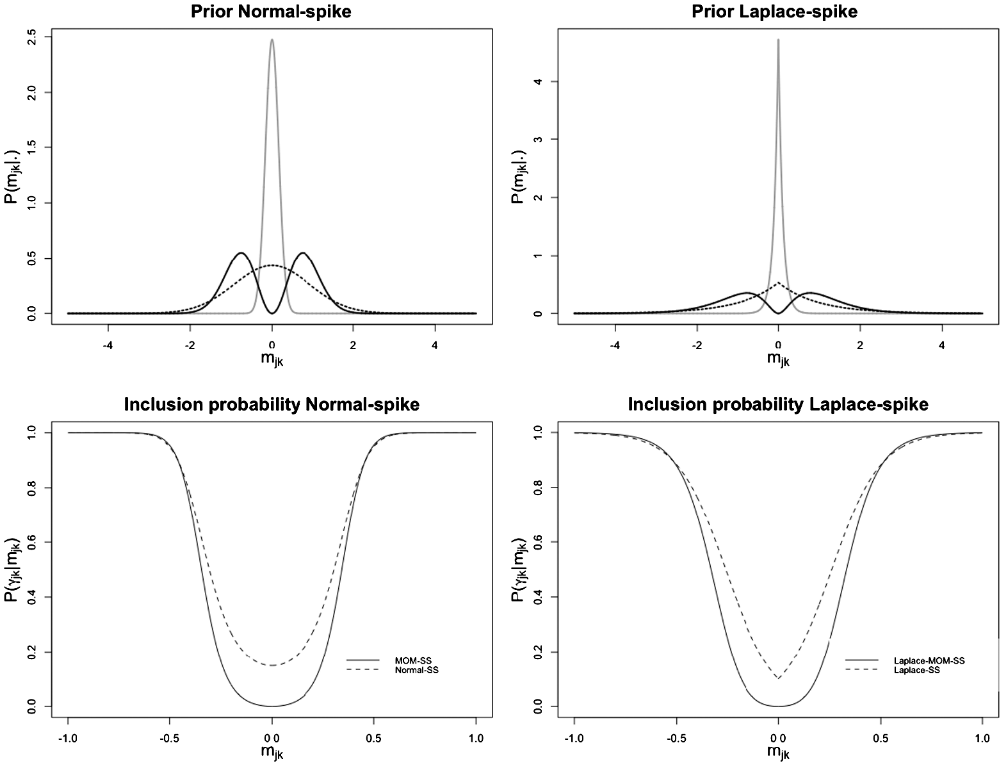


Figure 1: Prior comparison (top panels) for m_{jk} under different prior specifications and its inclusion probabilities $\mathsf{p}(\gamma_{jk} \mid m_{jk})$ (bottom panels). Comparison between Normal-based (left) and Laplaced-based (right) priors. Scales (λ_0, λ_1) are set to the defaults from Section 3.3.

In Laplace-MOM-SS, we analogously set $\tilde{\lambda}_0 = -\frac{\sqrt{0.1}}{\log(0.05)} \approx 0.1056$ so that $\mathsf{p}(|m_{jk}| \leq \sqrt{0.1} \mid \tilde{\lambda}_0)$ and $\tilde{\lambda}_1 \approx 0.3867$ such that $\mathsf{p}(|m_{jk}| \geq \sqrt{0.1} \mid \tilde{\lambda}_1) = 0.95$ for the Laplace-spike-and-MOM-slab prior. Finally for the Laplace-SS we set $\lambda_1 = \sqrt{6}\tilde{\lambda}_1 \approx 0.9473$ and $\lambda_0 = \tilde{\lambda}_0$ for the spike and slab component, respectively, matching the variances of the non-local Laplace-based priors.

The resulting priors are in Figure 1. We remark that a considerable difference can be observed between the local prior based and the non-local prior based formulations, particularly in the conditional inclusion probability around $m_{jk} = 0$. In our examples we will focus on the Normal MOM-SS. Deeper analysis of Laplace-based non-local priors, whose thicker tails might help improve estimation accuracy, is left for interesting future work.

4 Parameter estimation

Parameter estimation in factor analysis is usually conducted using Expectation-Maximisation (EM, Dempster et al. (1977)), MCMC algorithms (Lopes and West, 2004) or approximated via variational inference (Ghahramani and Beal, 2000). At the core of these algorithms is the fact that, conditional on the data and all other model parameters, we can set $\tilde{\mathtt{x}}_i = \mathtt{x}_i - \theta \mathtt{v}_i - \beta \mathtt{b}_i$ and express the model in (2.2) as a linear regression $\tilde{\mathtt{x}}_i = M\mathtt{z}_i + \mathtt{e}_i$, where M and Σ are fixed at their current values of each MCMC iteration or maximisation step (West, 2003; Carvalho et al., 2008). We develop a deterministic optimisation along the lines of the EM algorithm of Ročková and George (2017). Section 4.1 provides two EM algorithms to obtain posterior modes for our factor regression with batch effect correction with and without sparse formulation. Section 4.2 outlines an algorithm separately for Normal-SS, MOM-SS, Laplace-SS and Laplace-MOM-SS priors. Section 4.3 discusses parameter initialisation and Section 4.4 how to post-process the fitted model to obtain sparse solutions and variance-adjusted dimensionality reduction.

4.1 EM algorithm under a uniform prior

We outline an EM algorithm to fit Model (2.2) under a uniform prior $\mathtt{p}(M) \propto 1$ on the loadings via maximum a posteriori (MAP) estimation. The algorithm maximises the log-posterior by treating the latent factors Z as missing data and setting them to their expectation (conditional on all other parameters) in the E-step. Then, the remaining parameters $\Delta = (M, \theta, \beta, \mathcal{T})$ are optimised in the M-step. In other words, the EM algorithm obtains a local mode of the log-posterior $\mathtt{p}(M, \theta, \beta, \mathcal{T} \mid X)$ by maximising the expected complete-data log-posterior $\mathtt{p}(M, \theta, \beta, \mathcal{T} \mid X, Z)$ iteratively. For convenience we denote by $\mathcal{T}_{\mathtt{b}_i}$ the idiosyncratic precision matrix in batch l, i.e. if $b_{il} = 1$ by τ_{jl}, then the errors are distributed as $\mathtt{e}_i \sim N(0, \mathcal{T}_{\mathtt{b}_i}^{-1})$. We also denote with $\hat{\Delta} = (\hat{M}, \hat{\theta}, \hat{\beta}, \hat{\mathcal{T}})$ the current value of the parameters We briefly describe the algorithm; see Supplementary Avalos-Pacheco et al. (2020) Section 4 for its full derivation.

The E-step takes the expectation of $\log \mathtt{p}(M, \theta, \beta, \mathcal{T} \mid X, Z)$ with respect to $\mathtt{p}(Z \mid \hat{\Delta}, X)$ Specifically, let

$$\begin{aligned} Q(\Delta) &= \mathbb{E}_{z|\hat{\Delta},X}\left[\log \mathtt{p}(M, \theta, \beta, \mathcal{T} \mid X, Z)\right] \\ &= C - \frac{1}{2}\sum_{i=1}^{n}\Big[(\mathtt{x}_i - \theta\mathtt{v}_i - \beta\mathtt{b}_i)^\top \mathcal{T}_{\mathtt{b}_i}(\mathtt{x}_i - \theta\mathtt{v}_i - \beta\mathtt{b}_i) \\ &\quad - 2(\mathtt{x}_i - \theta\mathtt{v}_i - \beta\mathtt{b}_i)^\top \mathcal{T}_{\mathtt{b}_i} M \mathbb{E}[\mathtt{z}_i \mid \hat{\Delta}, X] + \operatorname{tr}\left(M^\top \mathcal{T}_{\mathtt{b}_i} M \mathbb{E}[\mathtt{z}_i\mathtt{z}_i^\top \mid \hat{\Delta}, X]\right)\Big] \\ &\quad + \sum_{l=1}^{p_b} \frac{n_l + \eta - 2}{2} \log \mid \mathcal{T}_l \mid - \sum_{l=1}^{p_b} \frac{\eta\xi}{2} \operatorname{tr}(\mathcal{T}_l) - \frac{1}{2}\sum_{j=1}^{p} (\theta_j^\top, \beta_j^\top) \frac{1}{\psi} \mathbf{I}(\theta_j, \beta_j), \end{aligned} \tag{4.1}$$

where C is a constant. Expression (4.1) only depends on Z through the conditional posterior mean

$$\mathbb{E}[\mathtt{z}_i|\hat{\Delta}, X] = (\mathbf{I}_q + \hat{M}^\top \hat{\mathcal{T}}_{\mathtt{b}_i} \hat{M})^{-1} \hat{M}^\top \hat{\mathcal{T}}_{\mathtt{b}_i}(\mathtt{x}_i - \hat{\theta}\mathtt{v}_i - \hat{\beta}\mathtt{b}_i) \tag{4.2}$$

and the conditional second moments

$$\mathbb{E}[\mathtt{z}_i \mathtt{z}_i^\top \mid \hat{\Delta}, X] = (\mathbf{I}_q + \hat{M}^\top \hat{\mathcal{T}}_{\mathtt{b}_i} \hat{M})^{-1} + \mathbb{E}[\mathtt{z}_i \mid \hat{\Delta}, X] \mathbb{E}[\mathtt{z}_i \mid \hat{\Delta}, X]^\top, \tag{4.3}$$

where $(\mathbf{I}_q + \hat{M}^\top \hat{\mathcal{T}}_{\mathtt{b}_i} \hat{M})^{-1} = \mathrm{Cov}[\mathtt{z}_i | \hat{\Delta}, X]$ is the conditional covariance matrix of the latent factors. We emphasise that (4.2) and (4.3) depend on batch-specific precisions $\mathcal{T}_{\mathtt{b}_i}$.

The M-step maximises $Q(\Delta)$ with respect to $M, \theta, \beta, \mathcal{T}$. Setting its partial derivatives to 0 gives the updates

$$\hat{m}_j = \left[\sum_{i=1}^n \left(\hat{\tau}_j^\top \mathtt{b}_i \tilde{x}_{ij} \mathbb{E}[\mathtt{z}_i^\top \mid \hat{\Delta}, X] \right) \right] \left[\sum_{i=1}^n \left(\hat{\tau}_j^\top \mathtt{b}_i \mathbb{E}[\mathtt{z}_i \mathtt{z}_i^\top \mid \hat{\Delta}, X] \right) \right]^{-1}, \tag{4.4}$$

$$\begin{aligned} \hat{\mathcal{T}}_l^{-1} = \frac{1}{n_l + \eta - 2} \mathrm{diag} \Bigg\{ & \sum_{i:\, b_{il} = 1} \left(\tilde{\mathtt{x}}_i \tilde{\mathtt{x}}_i^\top - 2 \tilde{\mathtt{x}}_i \mathbb{E}[\mathtt{z}_i \mid \hat{\Delta}, X]^\top \hat{M}^\top + \hat{M} \mathbb{E}[\mathtt{z}_i \mathtt{z}_i^\top \mid \hat{\Delta}, X] \hat{M}^\top \right) \\ & + \eta \xi \mathbf{I}_p \Bigg\}, \end{aligned} \tag{4.5}$$

where $\tilde{\mathtt{x}}_i = \mathtt{x}_i - \hat{\theta} \mathtt{v}_i - \hat{\beta} \mathtt{b}_i$ and $\tilde{x}_{ij} = x_{ij} - \hat{\theta} v_{ij} - \hat{\beta} b_{ij}$.

The updates for (θ_j, β_j) are

$$\begin{aligned} (\hat{\theta}_j^\top, \hat{\beta}_j^\top) = & \sum_{i=1}^n \left[\hat{\tau}_j^\top \mathtt{b}_i (x_{ij} - \hat{m}_j^\top \mathbb{E}[\mathtt{z}_i \mid \hat{\Delta}, X]) (\mathtt{v}_i, \mathtt{b}_i)^\top \right] \\ & \times \left[\sum_{i=1}^n \left[\hat{\tau}_j^\top \mathtt{b}_i (\mathtt{v}_i, \mathtt{b}_i) (\mathtt{v}_i, \mathtt{b}_i)^\top \right] + \frac{1}{\psi} \mathbf{I} \right]^{-1}. \end{aligned} \tag{4.6}$$

Equation (4.6) has the form of a ridge regression estimator with penalty ψ.

Algorithm 1 summarises the EM algorithm. The stopping criteria is reaching a tolerance ϵ^* in the log-posterior change, a maximum number of iterations T or a change ϵ_M^* on the loadings. By default we set $\epsilon^* = 0.001$, $T = 100$ and $\epsilon_M^* = 0.05$. Parameter initialisation is an important aspect that helps obtain better local modes and reduce computational time; its discussion is deferred to Section 4.3.

4.2 EM algorithm for spike-and-slab priors

The algorithm is derived analogously to Section 4.1. The expected complete-data log-posterior can be split into $Q(\Delta) = C + Q_1(\theta, M, \beta, \mathcal{T}) + Q_2(\zeta)$, where

$$\begin{aligned} Q_1(\theta, M, \beta, \mathcal{T}) = -\frac{1}{2} \sum_{i=1}^n & \left[(\mathtt{x}_i - \theta \mathtt{v}_i - \beta \mathtt{b}_i)^\top \mathcal{T}_{\mathtt{b}_i} (\mathtt{x}_i - \theta \mathtt{v}_i - \beta \mathtt{b}_i) \right. \\ & - 2 (\mathtt{x}_i - \theta \mathtt{v}_i - \beta \mathtt{b}_i)^\top \mathcal{T}_{\mathtt{b}_i} M \mathbb{E}[\mathtt{z}_i \mid \hat{\Delta}, X] \end{aligned}$$

Algorithm 1: EM algorithm for factor regression model with uniform $\mathsf{p}(M)$.

```
initialise $\hat{M} = M^{(0)}$, $\hat{\theta} = \theta^{(0)}$, $\hat{\beta} = \beta^{(0)}$, $\hat{\mathcal{T}}_{\mathtt{b}_i} = \mathcal{T}^{(0)}_{\mathtt{b}_i}$
while $\epsilon > \epsilon^*$, $\epsilon_M > \epsilon^*_M$ and $t < T$ do
    E-step:
     Latent factors: $\mathbb{E}[\mathbf{z}_i|\hat{\Delta}, X] = (\mathbf{I}_q + \hat{M}^\top \hat{\mathcal{T}}_{\mathtt{b}_i} \hat{M})^{-1} \hat{M}^\top \hat{\mathcal{T}}_{\mathtt{b}_i} (\mathtt{x}_i - \hat{\theta}\mathtt{v}_i - \hat{\beta}\mathtt{b}_i)$
    M-step:
        Loadings: $\hat{m}_j = \left[\sum_{i=1}^n \left(\hat{\tau}_j^\top \mathtt{b}_i \tilde{x}_{ij} \mathbb{E}[\mathbf{z}_i^\top \mid \hat{\Delta}, X]\right)\right] \left[\sum_{i=1}^n \left(\hat{\tau}_j^\top \mathtt{b}_i \mathbb{E}[\mathbf{z}_i \mathbf{z}_i^\top \mid \hat{\Delta}, X]\right)\right]^{-1}$
        Variances: $\hat{\mathcal{T}}_l^{-1} = \frac{1}{n_l + \eta - 2} \mathrm{diag}\left\{\sum_{i:\, b_{il}=1} \left(\tilde{\mathtt{x}}_i \tilde{\mathtt{x}}_i^\top - 2\tilde{\mathtt{x}}_i \mathbb{E}[\mathtt{z}_i \mid \hat{\Delta}, X]^\top \hat{M}^\top + \hat{M}\mathbb{E}[\mathtt{z}_i \mathtt{z}_i^\top \mid \hat{\Delta}, X]\hat{M}^\top\right) + \eta\xi\mathbf{I}_p\right\}$
      Coefficients: $(\hat{\theta}_j^\top, \hat{\beta}_j^\top) = \sum_{i=1}^n \left[\hat{\tau}_j^\top \mathtt{b}_i (x_{ij} - \hat{m}_j^\top \mathbb{E}[\mathtt{z}_i \mid \hat{\Delta}, X])(\mathtt{v}_i, \mathtt{b}_i)^\top\right] \left[\sum_{i=1}^n \left[\hat{\tau}_j^\top \mathtt{b}_i (\mathtt{v}_i, \mathtt{b}_i)(\mathtt{v}_i, \mathtt{b}_i)^\top\right] + \frac{1}{\psi}\mathbf{I}\right]^{-1}$
    set $\Delta^{(t+1)} = \hat{\Delta}$ and $M^{(t+1)} = \hat{M}$
    compute $\epsilon = Q(\Delta^{t+1}) - Q(\Delta^t)$, $\epsilon_M = \max|m_{jk}^{(t+1)} - m_{jk}^{(t)}|$ and $t = t + 1$
end
```

$$
\begin{aligned}
&+ \mathrm{tr}\left(M^\top \mathcal{T}_{\mathtt{b}_i} M \mathbb{E}[\mathbf{z}_i \mathbf{z}_i^\top \mid \hat{\Delta}, X]\right)\Big] \\
&+ \sum_{l=1}^{p_b} \frac{n_l + \eta - 2}{2} \log \mid \mathcal{T}_l \mid - \sum_{l=1}^{p_b} \frac{\eta\xi}{2} \mathrm{tr}(\mathcal{T}_l) \\
&- \frac{1}{2} \sum_{j=1}^{p} (\theta_j, \beta_j)^\top \frac{1}{\psi} \mathbf{I}(\theta_j, \beta_j) + \sum_{j=1}^{p} \sum_{k=1}^{q} \mathbb{E}_{\gamma|\hat{\Delta}} \left[\log \mathsf{p}(m_{jk} \mid \gamma_{jk}, \lambda_0, \lambda_1)\right],
\end{aligned} \tag{4.7}
$$

$$
\begin{aligned}
Q_2(\zeta) = &\sum_{j=1}^{p} \sum_{k=1}^{q} \log\left(\frac{\zeta_k}{1 - \zeta_k}\right) \mathbb{E}[\gamma_{jk} \mid \hat{\Delta}] \\
&+ \sum_{k=1}^{q} \left((\frac{a_\zeta}{k} - 1) \log(\zeta_k) + (p + b_\zeta - 1) \log(1 - \zeta_k)\right),
\end{aligned} \tag{4.8}
$$

with C a constant and $\mathbb{E}[\mathbf{z}_i \mid \hat{\Delta}, X]$ and $\mathbb{E}[\mathbf{z}_i \mathbf{z}_i^\top \mid \hat{\Delta}, X]$ as in (4.2) and (4.3).

$Q_1(\theta, M, \beta, \mathcal{T})$ resembles the E-step for the flat prior in Section 4.1, plus an extra conditional expectation $\mathbb{E}_{\gamma|\hat{\Delta}} [\log \mathsf{p}(m_{jk} \mid \gamma_{jk}, \lambda_0, \lambda_1)]$. $Q_2(\zeta)$ arises from the Beta-Binomial prior on γ_{jk} and the $\mathbb{E}[\gamma_{jk} \mid \cdot]$ are straightforward to compute. In the M-step we maximise Q_1 w.r.t. $(\theta, M, \beta, \mathcal{T})$, this can be done in a completely independent fashion from optimising Q_2 w.r.t. ζ.

Further the conditional expectation of $\mathbb{E}[\gamma_{jk} \mid \hat{\Delta}] = \hat{p}_{jk}$ is

$$
\hat{p}_{jk} = \frac{\mathsf{p}(\hat{m}_{jk} \mid \gamma_{jk} = 1, \lambda_0, \lambda_1)\mathsf{p}(\gamma_{jk} = 1)}{\mathsf{p}(\hat{m}_{jk} \mid \gamma_{jk} = 0, \lambda_0, \lambda_1)\mathsf{p}(\gamma_{jk} = 0) + \mathsf{p}(\hat{m}_{jk} \mid \gamma_{jk} = 1, \lambda_0, \lambda_1)\mathsf{p}(\gamma_{jk} = 1)}. \tag{4.9}
$$

For the Normal-SS prior, (4.9) is

$$
\hat{p}_{jk} = \left[1 + \sqrt{\frac{\lambda_1}{\lambda_0}} \exp\left(-\frac{1}{2}\hat{m}_{jk}^2 \left(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right)\right) \frac{1 - \mathbb{E}[\zeta_j]}{\mathbb{E}[\zeta_j]}\right]^{-1}, \tag{4.10}
$$

for the MOM-SS

$$\hat{p}_{jk} = \left[1 + \frac{\tilde{\lambda}_1}{\hat{m}_{jk}^2}\sqrt{\frac{\tilde{\lambda}_1}{\tilde{\lambda}_0}} \exp\left(-\frac{1}{2}\hat{m}_{jk}^2\left(\frac{1}{\tilde{\lambda}_0} - \frac{1}{\tilde{\lambda}_1}\right)\right)\frac{1-\mathbb{E}[\zeta_j]}{\mathbb{E}[\zeta_j]}\right]^{-1}, \tag{4.11}$$

for the Laplace-SS

$$\hat{p}_{jk} = \left[1 + \frac{\lambda_1}{\lambda_0} \exp\left(- \mid \hat{m}_{jk} \mid \left(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right)\right)\frac{1-\mathbb{E}[\zeta_j]}{\mathbb{E}[\zeta_j]}\right]^{-1}, \tag{4.12}$$

and for the Laplace-MOM-SS

$$\hat{p}_{jk} = \left[1 + \frac{2\tilde{\lambda}_1^2}{\hat{m}_{jk}^2}\frac{\tilde{\lambda}_1}{\tilde{\lambda}_0} \exp\left(- \mid \hat{m}_{jk} \mid \left(\frac{1}{\tilde{\lambda}_0} - \frac{1}{\tilde{\lambda}_1}\right)\right)\frac{1-\mathbb{E}[\zeta_j]}{\mathbb{E}[\zeta_j]}\right]^{-1}. \tag{4.13}$$

Equations (4.10) and (4.12) are analogous to the EM posterior update for m_{jk} in a two-component Gaussian or Laplace mixture (Ročková and George, 2014). Equations (4.11) and (4.13) are similar to their local counterparts, but incorporate a penalty for small m_{jk}^2.

The main difference between the local and non-local priors lies in updating the loadings and the idiosyncratic variances. We discuss these separately for each prior later in this section.

The updates for the precision $\mathcal{T}_l$ and the regression parameters (θ, β) are given in (4.5) and (4.6) respectively.

Maximising $Q_2(\zeta)$ with respect to ζ_k gives

$$\hat{\zeta}_k = \frac{\sum_{j=1}^p \hat{p}_{jk} + \frac{a_\zeta}{k} - 1}{\frac{a_\zeta}{k} + b_\zeta + p - 1}, \tag{4.14}$$

for $k = 1, \dots, q$.

Algorithm 2 summarises the algorithm. It is initialised with the two-stage least-squares method described in Section 4.3 and $\zeta_k = 0.5$ for $k = 1, \dots, q$. The stopping criteria are as in Algorithm 1. The different updates for M are outlined below, separately for each prior specification.

Let $d_{jk} = [(1-\gamma_{jk})\lambda_0 + \gamma_{jk}\lambda_1]^{-1}$. In Expression (4.7), under a Normal-SS prior

$$\mathbb{E}_{\gamma \mid \hat{\Delta}}\left[\log \mathsf{p}(m_{jk} \mid \gamma_{jk}, \lambda_0, \lambda_1)\right] \propto -\frac{1}{2}\hat{m}_{jk}^2 \mathbb{E}\left[d_{jk} \mid \hat{\Delta}\right] = -\frac{1}{2}\hat{m}_{jk}^2\left[\frac{1-\hat{p}_{jk}}{\lambda_0} + \frac{\hat{p}_{jk}}{\lambda_1}\right], \tag{4.15}$$

where $\hat{p}_{jk}$ is as in (4.10).

Algorithm 2: EM algorithm for factor regression model with spike-and-slab $\mathsf{p}(M)$.

```
initialise $\hat{M} = M^{(0)}$, $\hat{\theta} = \theta^{(0)}$, $\hat{\beta} = \beta^{(0)}$, $\hat{\mathcal{T}}_{\mathtt{b}_i} = \mathcal{T}_{\mathtt{b}_i}^{(0)}$, $\hat{\zeta} = \zeta^{(0)}$
while $\epsilon > \epsilon^*$, $\epsilon_M > \epsilon_M^*$ and $t < T$ do
    E-step:
        Latent factors: $\mathbb{E}[\mathtt{z}_i|\hat{\Delta}, X] = (\mathbf{I}_q + \hat{M}^\top \hat{\mathcal{T}}_{\mathtt{b}_i} \hat{M})^{-1} \hat{M}^\top \hat{\mathcal{T}}_{\mathtt{b}_i} (\mathtt{x}_i - \hat{\theta}\mathtt{v}_i - \hat{\beta}\mathtt{b}_i)$
        Latent indicators+: $\mathbb{E}[\gamma_{jk} \mid \hat{\Delta}] = \hat{p}_{jk}$
    M-step:
        Loadings+: $\hat{m}_{jk} = \arg\max_{m_{jk}} Q_1(\hat{\Delta})$
        Variances: $\hat{\mathcal{T}}_l^{-1} = \frac{1}{n_l+\eta-2} \mathrm{diag}\left\{\sum_{i:\, b_{il}=1} \left(\tilde{\mathtt{x}}_i \tilde{\mathtt{x}}_i^\top - 2\tilde{\mathtt{x}}_i \mathbb{E}[\mathtt{z}_i \mid \hat{\Delta}, X]^\top \hat{M}^\top + \hat{M}\mathbb{E}[\mathtt{z}_i \mathtt{z}_i^\top \mid \hat{\Delta}, X]\hat{M}^\top\right) + \eta\xi\mathbf{I}_p\right\}$
        Coefficients: $(\hat{\theta}_j^\top, \hat{\beta}_j^\top) = \sum_{i=1}^n \left[\hat{\tau}_j^\top \mathtt{b}_i (x_{ij} - \hat{m}_j^\top \mathbb{E}[\mathtt{z}_i \mid \hat{\Delta}, X])(\mathtt{v}_i, \mathtt{b}_i)^\top\right] \left[\sum_{i=1}^n \left[\hat{\tau}_j^\top \mathtt{b}_i (\mathtt{v}_i, \mathtt{b}_i)(\mathtt{v}_i, \mathtt{b}_i)^\top\right] + \frac{1}{\psi}\mathbf{I}\right]^{-1}$
        Weights: $\hat{\zeta}_k = \frac{\sum_{j=1}^p \hat{p}_{jk} + \frac{a_\zeta}{k} - 1}{\frac{a_\zeta}{k} + b_\zeta + p - 1}$
    set $\Delta^{(t+1)} = \hat{\Delta}$ and $M^{(t+1)} = \hat{M}$
    compute $\epsilon = Q(\Delta^{t+1}) - Q(\Delta^t)$, $\epsilon_M = \max|m_{jk}^{(t+1)} - m_{jk}^{(t)}|$ and $t = t + 1$
end
```

+ see Section 4.2, Supplementary Avalos-Pacheco et al. (2020) Sections 5 and 6 for details.

Thus, the EM update for the j^{th} row of matrix M is,

$$\hat{m}_j = \left[\sum_{i=1}^n \left(\hat{\tau}_j^\top \mathtt{b}_i \tilde{x}_{ij} \mathbb{E}[\mathtt{z}_i^\top \mid \hat{\Delta}, X]\right)\right] \times \left[\mathrm{diag}\{\mathbb{E}[d_{j1} \mid \hat{\Delta}], \ldots, \mathbb{E}[d_{jq} \mid \hat{\Delta}]\} + \sum_{i=1}^n \left(\hat{\tau}_j^\top \mathtt{b}_i \mathbb{E}[\mathtt{z}_i \mathtt{z}_i^\top \mid \hat{\Delta}, X]\right)\right]^{-1}, \qquad (4.16)$$

for $j = 1, \ldots, p$, where $\tilde{x}_{ij} = x_{ij} - \theta v_{ij} - \beta b_{ij}$. A full derivation is given in Supplementary Avalos-Pacheco et al. (2020) Section 5.

For the MOM-SS

$$\mathbb{E}_{\gamma|\hat{\Delta}}\left[\log \mathsf{p}(m_{jk} \mid \gamma_{jk}, \tilde{\lambda}_0, \tilde{\lambda}_1)\right] \propto -\frac{1}{2} m_{jk}^2 \left[\frac{1 - \hat{p}_{jk}}{\tilde{\lambda}_0} + \frac{\hat{p}_{jk}}{\tilde{\lambda}_1}\right] + \hat{p}_{jk} \log(m_{jk}^2), \qquad (4.17)$$

where $\hat{p}_{jk}$ is given in (4.11).

For the M-step, we use a coordinate descent algorithm (CDA) that performs successive univariate optimisation on (4.7) with respect to each m_{jk}. An advantage is that the updates have a closed-form that is computationally inexpensive. As a potential drawback it could require a larger number of iterations to converge relative to performing joint optimisation with respect to multiple elements in M. However, we have not found this to be a practical problem in our examples.

Viewed as a function of only m_{jk}, it is possible to express $Q_1(m_{jk})$ as

$$Q_1(m_{jk}) = a m_{jk}^2 + b m_{jk} + c \log(m_{jk}^2), \qquad (4.18)$$

where

$$\begin{aligned}
a &= -\frac{1}{2}\left(\left[\frac{1-\hat{p}_{jk}}{\tilde{\lambda}_0}+\frac{\hat{p}_{jk}}{\tilde{\lambda}_1}\right]+\sum_{i=1}^{n}\hat{\tau}_j^{\top}\mathtt{b}_i\mathbb{E}[z_{ik}z_{ik}^{\top}\mid\hat{\Delta},X]\right),\\
b &= \sum_{i=1}^{n}\left[\hat{\tau}_j^{\top}\mathtt{b}_i(x_{ij}-\hat{\theta}v_{ij}-\hat{\beta}b_{ij})\mathbb{E}[z_{ik}\mid\hat{\Delta},X]-\sum_{r\neq k}^{q}\hat{m}_{jr}\hat{\tau}_j^{\top}\mathtt{b}_i\mathbb{E}[z_{ir}z_{ik}^{\top}\mid\hat{\Delta},X]\right],\\
c &= \hat{p}_{jk}.
\end{aligned} \tag{4.19}$$

See Supplementary Avalos-Pacheco et al. (2020) Section 6. The global maximum of (4.18) is summarised in Lemma 1.

Lemma 1. Let $f(m_{jk}) = am_{jk}^2 + bm_{jk} + c\log(m_{jk}^2)$, where $a < 0$ and $c > 0$. Define $\underline{m}_{jk} = \frac{-b-\sqrt{b^2-16ac}}{4a}$ and $\bar{m}_{jk} = \frac{-b+\sqrt{b^2-16ac}}{4a}$.

If $b > 0$, then $\underline{m}_{jk} = \arg\max_{m_{jk}} f(m_{jk})$. If $b < 0$, then $\bar{m}_{jk} = \arg\max_{m_{jk}} f(m_{jk})$. If $b = 0$, then $\bar{m}_{jk} = \underline{m}_{jk} = \arg\max_{m_{jk}} f(m_{jk})$.

Akin to the MOM-SS, we can express $Q_1(m_{jk})$ as function of m_{jk} for the Laplace-based priors as:

$$Q_1(m_{jk}) = am_{jk}^2 + bm_{jk} + c|m_{jk}| + d\log(m_{jk}^2),$$

$$\begin{aligned}
a &= -\frac{1}{2}\sum_{i=1}^{n}\hat{\tau}_j^{\top}\mathtt{b}_i\mathbb{E}[z_{ik}z_{ik}^{\top}\mid\hat{\Delta},X],\\
b &= \sum_{i=1}^{n}\left[\hat{\tau}_j^{\top}\mathtt{b}_i(x_{ij}-\hat{\theta}v_{ij}-\hat{\beta}b_{ij})\mathbb{E}[z_{ik}\mid\hat{\Delta},X]-\sum_{r\neq k}^{q}\hat{m}_{jr}\hat{\tau}_j^{\top}\mathtt{b}_i\mathbb{E}[z_{ir}z_{ik}^{\top}\mid\hat{\Delta},X]\right],\\
c &= -\left[\frac{1-\hat{p}_{jk}}{\lambda_0}+\frac{\hat{p}_{jk}}{\lambda_1}\right],\\
d &= \begin{cases}0 & \text{for Laplace-SS}\\ \hat{p}_{jk} & \text{for Lapace-MOM-SS,}\end{cases}
\end{aligned} \tag{4.20}$$

for $j = 1,\dots,p$ and where $\hat{p}_{jk}$ is as in (4.12) and (4.13) for Laplace-SS and Laplace-MOM-SS respectively.

Lemma 2 summarises the global maximum for Laplace-SS.

Lemma 2. Let $f(m_{jk}) = am_{jk}^2 + bm_{jk} + c|m_{jk}|$, where $a < 0$ and $c < 0$. Define $m_{jk}^{+} = \frac{-(b+c)}{2a}$ and $m_{jk}^{-} = \frac{-(b-c)}{2a}$.

If $b > -c$, then $m_{jk}^{+} = \arg\max_{m_{jk}} f(m_{jk})$. If $b < c$, then $m_{jk}^{-} = \arg\max_{m_{jk}} f(m_{jk})$. If $c \leq b \leq -c$, then $0 = \arg\max_{m_{jk}} f(m_{jk})$.

Finally for the Laplace-MOM-SS, we emphasise that when $m_{jk} = 0$, $Q_1(m_{jk} = 0) = -\infty$. Thus the solution for m_{jk} is given by setting $\frac{\partial Q_1}{\partial m_{jk}} = 0$ as given in Lemma 3.

Lemma 3. Let $f(m_{jk}) = am_{jk}^2 + bm_{jk} + c|m_{jk}| + d\log(m_{jk}^2)$, where $a < 0$, $c < 0$ and $d > 0$. Define $m_{jk}^+ = \frac{-(b+c)-\sqrt{(b+c)^2-16ad}}{4a}$ and $m_{jk}^- = \frac{-(b-c)+\sqrt{(b-c)^2-16ad}}{4a}$.

If $b > 0$, then $m_{jk}^+ = \arg\max_{m_{jk}} f(m_{jk})$. If $b < 0$, then $m_{jk}^- = \arg\max_{m_{jk}} f(m_{jk})$. If $b = 0$, then $m_{jk}^+ = m_{jk}^- = \arg\max_{m_{jk}} f(m_{jk})$.

We remark that if either $\mathtt{x}_i$ or $\mathtt{v}_i$ are continuous, the event of $b = 0$ has zero probability. If both $\mathtt{x}_i$ and $\mathtt{v}_i$ are discrete and in presence of the rare event of $b = 0$, then the sign of the update for m_{jk} is set to the previous one.

4.3 Initialisation of parameters

The EM algorithm can be sensitive to parameter initialisation. We propose two different strategies: least-squares and least-squares with rotation.

The first option is a simple two-step least-squares that is computationally efficient and performs well in many of our examples.

Step 1: initialise $(\theta^{(0)}, \beta^{(0)}) = [(V, B)^\top (V, B)]^{-1}(V, B)^\top X$.

Step 2: Let $\hat{E} = X - (V\theta^{(0)\top} + B\beta^{(0)\top})$. Consider the eigendecomposition of $\frac{1}{n}\hat{E}^\top \hat{E}$ where $l_1 \geq l_2 \geq \cdots \geq l_q$ are the eigenvalues and $u_1, \ldots, u_q$ the eigenvectors. Set $M^{(0)} = [\sqrt{l_1}u_1 \mid \cdots \mid \sqrt{l_q}u_q]$ and $\mathcal{T}_l^{(0)} = [\text{diag}\{\frac{1}{n}\hat{E}^\top \hat{E} - M^{(0)}M^{(0)\top}\}]^{-1}$ for $l = 1, \ldots, p_b$.

The rotated least-squares adds an extra step.

Step 3: varimax rotation for the loadings obtained in Step 2.

The reason for this extra step is to help escape local modes. The EM algorithm does not guarantee convergence to a global maximum, but it increases the log-posterior at each iteration. This local maxima issue is intensified by the non-identifiability of the factor model through the rotational ambiguity of the likelihood and the strong association between the updates of loadings and factors.

4.4 Post-processing for model selection and dimensionality reduction

The EM algorithm gives point estimates $(\hat{M}, \hat{\theta}, \hat{\mathcal{T}}, \hat{\zeta})$. Under Laplace-SS one can obtain exact sparsity via $\hat{m}_{jk} = 0$, however this is not the case for our other priors. To address this, we define $\hat{\gamma}$ as the solution of the following optimisation problem

$$\hat{\gamma} = \text{argmax}_\gamma \mathsf{p}(\gamma \mid X, \hat{M}, \hat{\theta}, \hat{\mathcal{T}}, \hat{\zeta}) = \text{argmax}_\gamma \prod_{jk} \mathsf{p}(\gamma_{jk}|\hat{m}_{jk}, \hat{\zeta}_k), \tag{4.21}$$

where the right-hand side follows from the assumed conditional independence of m_{jk}. That is, we set $\hat{\gamma}_{jk} = 1$ if $\mathsf{p}(\gamma_{jk} = 1|\hat{m}_{jk}, \hat{\zeta}_k) > 0.5$ and $\gamma_{jk} = 0$ otherwise. When $\hat{\gamma}_{jk} = 0$ we set $\hat{m}_{jk} = 0$ effectively selecting the number of factors and the non-zero loadings within each factor.

As an alternative post-processing step we consider that in some applications one may want to select only the number of factors. We then consider to setting $\tilde{\gamma}_{jk} = 1$ if $\sum_{j=1}^{p} \hat{\gamma}_{jk} \neq 0$ and $\gamma_{jk} = 0$ otherwise.

The combination of the two initialisation alternatives and two different post-processing options gives four possible solutions for $\hat{M}$. To choose which is best in our examples, we use weighted 10-fold cross-validation, where the weights reflect that batches with higher variance should receive lower weight, selecting the model with smallest weighted cross validation reconstruction error (see Supplementary Avalos-Pacheco et al. (2020) Section 9 for details).

Finally we re-order of the factors so that $\sum_{j=1}^{p} \gamma_{jk}$ is decreasing in k, which under our prior (3.5) is guaranteed to increase the log-posterior. This is the so-called left-ordered inclusion matrix of Griffiths and Ghahramani (2011). This facilitates the interpretation of latent factors.

Latent factors are also post-processed for data visualisation purposes. The aim of this is to obtain new standardised factors $\tilde{\mathbf{z}}_i = [\mathrm{Cov}(\mathbf{z}_i \mid \hat{\Delta}, X)]^{-1}\mathbb{E}[\mathbf{z}_i \mid \hat{\Delta}, X]$, with $\mathrm{Cov}(\mathbf{z}_i \mid \hat{\Delta}, X) = (\mathbf{I}_q + \hat{M}^\top \hat{\mathcal{T}}_{\mathtt{b}_i} \hat{M})^{-1}$, whose covariance does not depend on their batch.

5 Results

We assess our approach on simulated and experimental datasets. In Section 5.1 we evaluate the accuracy of our prior in obtaining sparse factor loadings and in estimating the covariance and low-dimensional representations, by comparing its performance to other methods in a setting where there are no batch effects. In Section 5.2 we show the importance of accounting for batch effects in simulated data. In Section 5.3 we do the same analysis for three cancer datasets, assessing the ability of our dimension reduction to predict survival outcomes.

Sections 5.1 and 5.2 study simulations under two different loading matrices M (truly sparse and dense) and two different scenarios (without and with batch effects). We compare our methods with the Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity (FastBFA) of Ročková and George (2017) and the Penalized Likelihood Factor Analysis with a LASSO penalty (LASSO-BIC) of Hirose and Yamamoto (2015). We also use the ComBat empirical Bayes batch effect correction of Johnson et al. (2007) for scenarios with batch effects, doing an MLE estimation of the factor analysis model (ComBat-MLE). In Section 5.3 we analyse a high-dimensional gene expression data under a supervised and an unsupervised framework. We use the clinically annotated data for the ovarian cancer transcriptome from **R** package `curatedOvarianData 1.16.0` (Ganzfried et al., 2013), the lung cancer data from The Cancer Genome Atlas (TCGA) from **R** package `TCGA2STAT 1.2` (Wan et al., 2015) and gene expression data from the colon cancer datasets of Calon et al. (2012).

The **R** package for our model is available at `https://github.com/AleAviP/BFR.BE`. The software sets all tuning parameters to the default values suggested in this paper. We used **R** function FACTOR_ROTATE of Ročková and George (2017) for FastBFA, the **R** package fanc 2.2 for LASSO-BIC (Hirose et al., 2016) and package `sva 3.26.0` for ComBat (Leek et al., 2017). Hyper-parameters for the Normal-SS and MOM-SS were set as in Section 3.3, the hyper-parameters for FastBFA were set via Dynamic Posterior Exploration as in Ročková and George (2017) with $1/\lambda_0 = 0.001$ and $1/\lambda_1 \in \{5, 10, 20, 30\}$ and using varimax robustifications. For the LASSO-BIC we selected the model with smallest BIC to set the regularisation parameter. Finally, for scenarios with batch effects, we adjusted the data via a ComBat correction and performed a Factor Analysis via EM algorithm to maximise likelihood with the `fa.em` function in the `cate` package (Wang and Zhao, 2015).

5.1 No batch effect

To assess the precision of the parameter estimates returned by the EM algorithm, we simulated data from two different data-generating truths: truly sparse and dense for the loadings M. In both, the truth was set to $q^* = 10$ factors. The dense loadings matrix has a grid of elements set uniformly between $(-1, 1)$, whereas the truly sparse M has a banded-diagonal structure with $m_{jk} = 1$ for the non-zero elements, as shown in Figure 2.

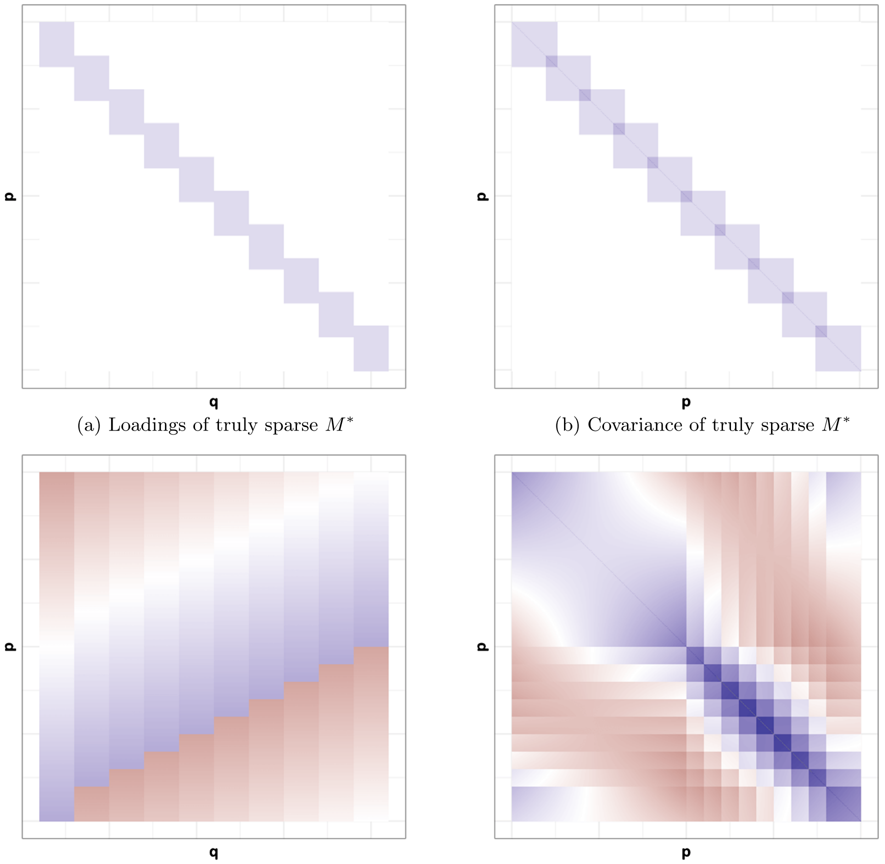
Some visual representations of our findings are display in Figure 3 and in the Supplementary Avalos-Pacheco et al. (2020) Figures 3–8.

We simulated $n = 100$ observations from $\mathtt{x}_i = M^*\mathtt{z}_i, +\mathtt{e}_i$, with growing $p = 1{,}000$ and 1,500, where the factors $\mathtt{z}_i \sim N(0, \mathbf{I}_q)$, the errors $\mathtt{e}_i \sim N(0, \mathcal{T}^{-1})$ with $\mathcal{T}^{-1} = \mathbf{I}_p$, and the loadings M^* are set as dense or sparse as in Figure 2. Our simulations have been normalised to zero mean and unit variance. For comparison, FastBFA was initialised as our models via two-step least-squares (Section 4.3).

Table 1 shows the selected number of factors $\hat{q}$, the number of estimated non-zero loadings $|\hat{M}|_0 = \sum_{j,k} \mathbb{1}(\hat{m}_{jk} \neq 0)$, the Frobenius norm (F.N.) between the true expected value and its reconstruction $||E[X] - \hat{E}[X]||_F = ||ZM^\top - \mathbb{E}[Z \mid \hat{\Delta}, X]\hat{M}^\top||_F$ and between the true and reconstructed covariances $||\text{Cov}[x_i] - \widehat{\text{Cov}}[x_i]||_F = ||(MM^\top + \mathcal{T}^{-1}) - (\hat{M}\hat{M}^\top + \hat{\mathcal{T}}^{-1})||_F$, the number of iterations until convergence, and the computation time in minutes. The mean across 100 different simulations is displayed.

We first considered the unrealistic scenario where M is dense and one guessed correctly the true number of factors $q = q^* = 10$. The aim of this setting was to investigate if MOM-SS shrinkage provided a poor estimation when the factors were not truly sparse. MOM-SS and Normal-SS performed similarly as p grew, and competitively relative to the flat prior. To extend our example, we then set $q = 100$ to illustrate the performance when there is sparsity in terms of the number of factors, but not within factors. LASSO-BIC had the best reconstruction for the mean but performed poorly on the covariance, whereas FastBFA outperformed all the models to estimate the covariance but performed poorly for the mean. However, MOM-SS had a good balance in terms of estimating the expected value and the covariance, being the second best in both cases but requiring only 6 factors instead of 100.

We further illustrate our model under the arguably more interesting case of truly sparse loadings. First we set $q = 10$ the true cardinality. In this scenario MOM-SS



(a) Loadings of truly sparse M^*

(b) Covariance of truly sparse M^*

(c) Loadings of dense M^*

(d) Covariance of dense M^*

Figure 2: Synthetic data. Heatmaps of data-generating loadings and covariance with red highly negative, blue highly positive and white zero values.

and Normal-SS presented the best results both for mean and covariance. This example reflects the advantages of shrinkage and the varimax rotation for the initialisation in the loadings, leading to good sparse solutions. Finally we considered the same scenario with $q = 100$. LASSO-BIC was best to estimate the mean at the cost of reduced precision in the covariance reconstruction. MOM-SS displayed the lowest error for the mean covariance and the first or second lowest error for the covariance, showing a good balance between those metrics.

In general, MOM-SS achieved a good balance between estimating the mean, which is useful for dimensionality reduction, and sparse covariance estimation. Recall that we used a coordinate descent algorithm for the non-local prior, which as a potential drawback could require a larger number of iterations than performing jointly optimising multiple elements in M. However, Table 1 showed that MOM-SS required roughly the

	$p = 1{,}000$						$p = 1{,}500$					
Model	$\hat{q}$	$\lvert\hat{M}\rvert_0$	$\lvert\lvert\mathbb{E}[X] - \hat{\mathbb{E}}[X]\rvert\rvert_F$	$\lvert\lvert\text{Cov}[x_i] - \widehat{\text{Cov}}[x_i]\rvert\rvert_F$	it	time	$\hat{q}$	$\lvert\hat{M}\rvert_0$	$\lvert\lvert\mathbb{E}[X] - \hat{\mathbb{E}}[X]\rvert\rvert_F$	$\lvert\lvert\text{Cov}[x_i] - \widehat{\text{Cov}}[x_i]\rvert\rvert_F$	it	time
	Dense M, $q = 10$											
Flat	10.0	10000.0	59.5	569.4	2.0	0.4	10.0	10000.0	73.0	917.4	2.0	1.4
Normal-SS	10.0	1048.5	54.5	604.8	4.1	1.0	10.0	1467.6	67.4	957.7	4.0	2.9
MOM-SS	10.0	986.6	54.4	607.9	4.4	1.9	10.0	1389.2	67.2	957.5	4.0	4.5
FastBFA	9.3	959.0	78.9	594.0	11.1	0.5	9.3	1447.9	96.2	891.1	11.0	0.8
LASSO-BIC	10.0	5322.8	49.5	794.2	NA	2.1	10.0	8600.8	60.3	1196.9	NA	4.5
	Dense M, $q = 100$											
Flat	100.0	100000.0	163.9	576.2	3.0	4.1	100.0	100000.0	204.1	926.8	3.0	10.9
Normal-SS	7.3	1745.3	72.1	589.8	4.0	8.7	28.6	1875.5	125.2	951.2	4.0	18.7
MOM-SS	6.4	1385.6	80.8	577.1	5.0	10.4	5.9	1682.5	107.9	935.2	4.1	22.4
FastBFA	70.7	1516.9	150.1	478.6	13.2	1.5	77.3	2231.1	191.2	703.6	12.5	2.1
LASSO-BIC	11.0	4830.2	45.0	794.4	NA	61.5	11.2	7844.7	55.4	1197.0	NA	128.3
	Sparse M, $q = 10$											
Flat	10.0	10000.0	73.5	125.3	2.0	0.5	10.0	10000.0	89.4	203.7	2.0	1.1
Normal-SS	10.0	1298.6	43.9	89.1	3.0	1.3	10.0	1931.4	54.2	180.7	3.0	2.7
MOM-SS	10.0	1296.6	43.5	80.7	3.1	2.2	10.0	1919.3	56.2	169.4	3.0	4.3
FastBFA	9.9	778.1	60.3	165.0	11.2	0.6	9.9	1157.8	72.8	247.7	11.3	0.7
LASSO-BIC	10.0	5288.7	54.9	270.2	NA	0.7	10.0	8414.6	67.2	408.4	NA	0.8
	Sparse M, $q = 100$											
Flat	100.0	100000.0	209.5	185.7	3.0	5.1	100.0	100000.0	259.2	280.2	3.0	7.1
Normal-SS	31.0	1228.6	109.0	144.6	4.3	9.9	56.4	1568.2	181.3	231.9	4.0	13.8
MOM-SS	9.7	856.8	79.4	143.3	4.8	10.8	9.2	745.4	105.0	245.6	4.0	17.3
FastBFA	83.6	1389.9	198.1	141.9	12.0	1.7	87.2	1763.9	208.2	211.3	6.5	1.0
LASSO-BIC	10.0	4787.3	54.1	271.4	NA	19.6	10.0	7976.6	66.1	409.3	NA	31.6

Table 1: Synthetic data without batch effects for $n = 100$, $q^* = 10$, $p = 1{,}000$ or 1,500 parameters, truly sparse and dense loadings M^*.

same number of iterations and the same time to converge as the Normal-SS. We can see that MOM-SS and LASSO-BIC estimated $\hat{q}$ accurately in most scenarios. Note that in general FastBFA had the highest estimated latent cardinality $\hat{q}$, due to the fat tails of the Laplace priors, which adds some columns of M that contain very few non-zero loadings after the tenth factor, as shown in Supplementary Avalos-Pacheco et al. (2020) Sections 10 and 11. Nonetheless, this model displayed a mean number of non-zero loadings closer to the ground truth (1,300 and 1,940 for the $p = 1{,}000$ and $p = 1{,}500$ respectively under sparse M).

5.2 Batch effects

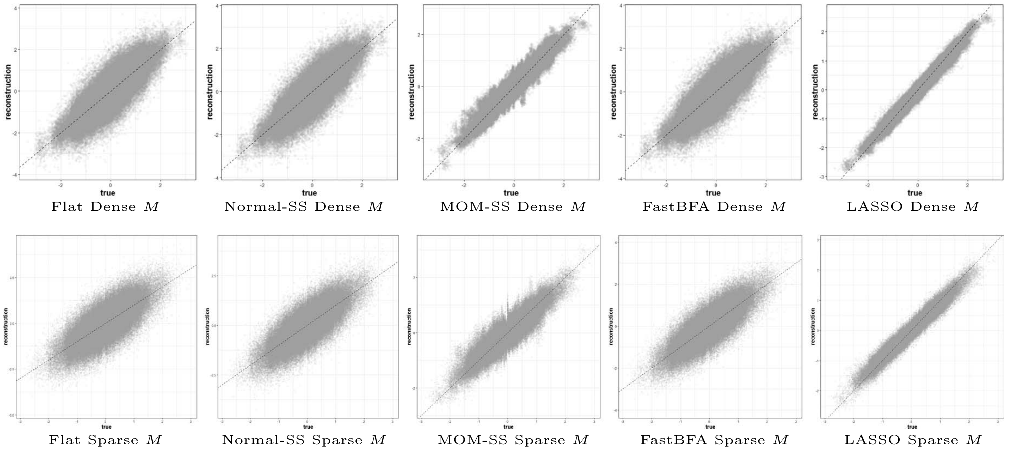
We evaluate our method in our main setting of interest where there are mean and variance batch effects. We emphasise that, the competing methods are not designed to account for batch effects; thus, this is not a fair comparison but rather an illustration of how much inference can suffer when not properly accounting for batches. Also, since Flat-SS, Normal-SS and MOM-SS do incorporate batches, comparing them illustrates the advantages of NLP-based sparsity, e.g. see the bottom row in Figure 4.

We simulated data with a mean and variance batch effect, $\mathtt{x}_i = \theta^* \mathtt{v}_i + M^* \mathtt{z}_i + \beta^* \mathtt{b}_i + \mathtt{e}_i$, sample size $n = 200$ and growing $p = 250$ or $p = 500$. We set $q^* = 10$, $p_v = 1$ and $p_b = 2$ batches and considered the truly sparse and dense loadings M^* in Figure 2. Factors $\mathtt{z}_i$ were drawn from $N(0, \mathbf{I}_q)$, errors $\mathtt{e}_i$ from $N(0, \mathcal{T}_{\mathtt{b}_i}^{-1})$, where $\tau_{j1}^{-1} = 0.5$ and $\tau_{j2}^{-1} = 1.5\tau_{j1}^{-1}$ for $j = 1, \ldots, p$; $\mathtt{v}_i$ from a continuous Uniform(0,3) and $\mathtt{b}_i$ from a discrete Uniform{0,1}. We set the first $p/2$ values of $\theta^* \in \mathbb{R}^p$ to -2 and the other $p/2$ to 2 and $\beta^*_{j1} = 0$, $\beta^*_{j2} = 2$ for $j = 1, \ldots, p$ we fixed to 2 for the first batch and 0 for the second. We then normalised our simulations to zero mean and unit variance. We compared our models with FastBFA and LASSO-BIC without batch effect correction for illustration of the importance of a proper mean and variance batch effect adjustment; and with empirical Bayes batch effect correction, ComBat, followed with an MLE estimation of the parameters ComBat-MLE. Table 2 shows the results. The following plots show the comparison between the true $ZM^\top$ against their reconstruction $\mathbb{E}[Z \mid \hat{\Delta}, X]$ in the scenario with sparsity with factors $q = 100$.

Firstly, we considered the scenario when one correctly guesses $q = 10$ and loadings are truly dense and sparse solutions could provide poor estimations. MOM-SS and Normal-SS achieved similar performance as the case without batch effect and similar results were observed for the $q = 100$.

Secondly, we studied the scenario with sparse factors. MOM-SS and Normal-SS achieved a small estimation error for the mean and were effective in estimating $q^* = 10$. LASSO-BIC had a small estimation error of the mean, although solutions were generally less sparse in the number of non-zero loadings.

It is important to highlight that even though FastBFA and LASSO-BIC achieved a precise reconstruction of $\mathbb{E}[X]$ for purposes of dimensionality reduction, the estimates of $ZM^\top$ are less precise, as shown in Table 2 and Figure 4 (right panels), clearly suggesting two clusters. We remark that for FastBFA and LASSO-BIC these results mainly highlight that one should take into account batch effects. For Combat-MLE



Flat Dense M Normal-SS Dense M MOM-SS Dense M FastBFA Dense M LASSO Dense M

Flat Sparse M Normal-SS Sparse M MOM-SS Sparse M FastBFA Sparse M LASSO Sparse M

Figure 3: Scatterplots comparing $ZM^\top$ vs. $\mathbb{E}[Z \mid \hat{\Delta}, X]\hat{M}^\top$ between the different models under dense (top) and truly sparse (bottom) loadings M with $q = 100$ in simulations without batch effect.

they highlight the limitations of using two-step procedures relative to a joint estimation of the factor model and batch effects. The reconstruction of $ZM^\top$ for Combat-MLE corrected the batch effects but was less precise than our suggested models.

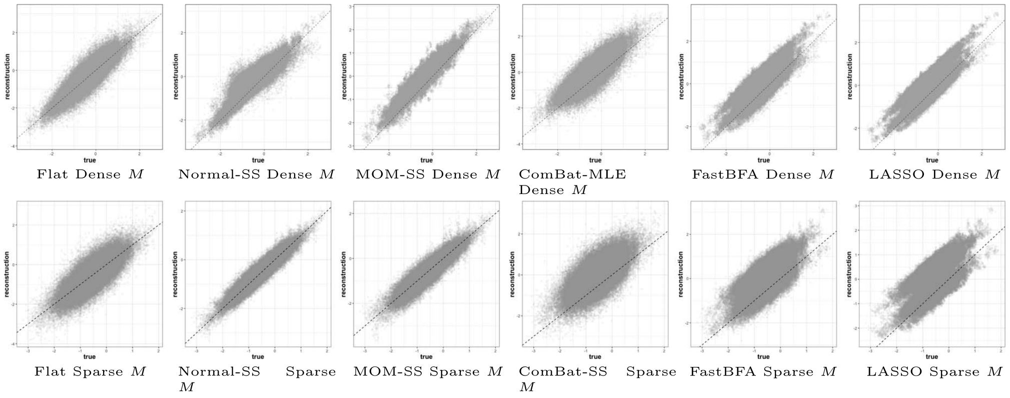
Finally, Supplements Avalos-Pacheco et al. (2020) 14 and 15 present the results of our methods without standardising the data to zero mean and unit variance, to assess the sensitivity of the various methods. The results showed that the performance of the two-step method COMBAT-MLE varied significantly, suggesting that such a method may not be robust to re-scaling the data. MOM-SS showed a robust performance across the two scenarios.

5.3 Applications to cancer datasets

We applied our method to three high-dimensional cancer datasets, related to ovarian, lung and colorectal cancer. For the ovarian cancer we combined information from two datasets from the package `curatedOvarianData 1.16.0`. The first was the Illumina Human microRNA array expression dataset `E.MTAB.386`, formed by Angiogenic mRNA and microRNA gene expression signature with $n_1 = 129$ patients (Bentink et al., 2012). The second was the NCI-60 GEO dataset `GSE30161` and consisted of multi-gene expression predictors of single drug responses to adjuvant chemotherapy in ovarian carcinoma for $n_2 = 52$ patients (Ferriss et al., 2012). For the lung cancer, we used microarray and mRNA-array, data from two different high-throughput platforms: Affymetrix Human Genome `U133A 2.0` Array with $n_1 = 133$ patients and Affymetrix Human `Exon 1.0` ST Array with $n_2 = 112$ (Wan et al., 2016). For the colorectal cancer we integrated the two colon cancer datasets used by Calon et al. (2012): The Omnibus gene expression dataset `GSE17538` with $n_1 = 232$ patients and the gene expression data of $n_2 = 101$ patients from the Australi hospital `GSE14333`.

	$p = 250$						$p = 500$					
Model	$\hat{q}$	$\lvert\hat{M}\rvert_0$	$\lvert\lvert\mathbb{E}[X] - \hat{\mathbb{E}}[X]\rvert\rvert_F$	$\lvert\lvert ZM^\top - \mathbb{E}[Z \mid \hat{\Delta}, X]\hat{M}^\top\rvert\rvert_F$	it	time	$\hat{q}$	$\lvert\hat{M}\rvert_0$	$\lvert\lvert\mathbb{E}[X] - \hat{\mathbb{E}}[X]\rvert\rvert_F$	$\lvert\lvert ZM^\top - \mathbb{E}[Z \mid \hat{\Delta}, X]\hat{M}^\top\rvert\rvert_F$	it	time
	Dense M, $q = 10$											
Flat	10.0	2500.0	34.1	104.3	7.0	0.3	10.0	2500.0	44.3	145.8	6.9	0.8
Normal-SS	10.0	511.1	32.4	103.1	8.3	0.3	10.0	873.4	41.6	143.4	5.7	0.8
MOM-SS	10.0	395.3	33.3	103.3	14.6	0.6	8.7	1005.5	53.9	144.3	12.2	4.0
ComBat-MLE	10.0	2500.0	103.6	107.8	2.0	0.0	10.0	2500.0	144.3	151.8	1.0	0.0
FastBFA	7.9	271.7	38.0	143.6	9.2	0.0	8.3	514.5	49.8	202.7	6.9	0.1
LASSO-BIC	10.0	1250.7	34.0	143.3	0.0	0.6	10.0	2460.8	43.3	202.5	0.0	1.2
	Dense M, $q = 100$											
Flat	100.0	25000.0	79.8	86.2	4.5	1.4	100.0	25000.0	122.5	131.2	4.0	2.8
Normal-SS	8.6	682.0	84.0	104.8	5.0	1.7	17.7	677.8	114.4	139.9	4.9	3.5
MOM-SS	7.3	602.3	151.1	103.4	6.9	8.1	8.1	1093.4	211.4	140.5	6.6	24.5
ComBat-MLE	100.0	25000.0	142.1	145.1	8.3	0.0	100.0	25000.0	194.2	199.5	5.6	0.0
FastBFA	9.9	368.1	33.9	143.3	17.6	0.4	11.0	728.6	47.0	203.4	16.5	0.5
LASSO-BIC	11.2	1109.8	28.4	141.5	0.0	9.1	11.0	2292.3	33.0	200.0	0.0	20.1
	Sparse M, $q = 10$											
Flat	10.0	2500.0	42.7	52.0	4.2	0.2	10.0	2500.0	54.8	68.2	4.0	0.4
Normal-SS	10.0	330.0	39.7	53.7	5.5	0.3	10.0	650.0	51.2	68.1	4.0	0.7
MOM-SS	10.0	330.0	39.2	61.3	8.8	0.5	10.0	650.0	49.6	86.1	9.3	1.5
ComBat-MLE	10.0	2500.0	127.2	143.3	2.8	0.0	10.0	2500.0	177.9	200.8	2.0	0.0
FastBFA	10.0	173.1	53.7	166.8	5.2	0.0	10.0	376.0	71.3	235.4	5.0	0.1
LASSO-BIC	10.0	1441.3	39.9	179.4	0.0	0.5	10.0	3159.1	50.0	254.2	0.0	0.8
	Sparse M, $q = 100$											
Flat	100.0	25000.0	96.8	100.6	4.4	1.3	100.0	25000.0	147.8	152.5	4.0	3.6
Normal-SS	10.0	765.8	45.7	54.8	5.0	1.7	10.6	1146.3	60.0	72.6	5.0	4.0
MOM-SS	10.0	740.4	63.8	72.4	6.0	1.5	10.0	1158.7	85.7	108.3	5.4	3.6
ComBat-MLE	100.0	25000.0	169.0	182.9	8.7	0.0	100.0	25000.0	232.7	252.4	4.9	0.0
FastBFA	10.0	337.0	51.9	168.3	12.7	0.3	11.3	681.8	75.8	247.9	11.9	0.4
LASSO-BIC	10.3	1374.0	39.6	178.9	0.0	3.5	10.3	2613.9	49.8	252.1	0.0	8.8

Table 2: Synthetic data with batch effects for $n = 200$, $q^* = 10$, $p = 250$ or 500 parameters, truly sparse and dense loadings M^*.



Flat Dense M Normal-SS Dense M MOM-SS Dense M ComBat-MLE Dense M FastBFA Dense M LASSO Dense M

Flat Sparse M Normal-SS Sparse M MOM-SS Sparse M ComBat-SS Sparse M FastBFA Sparse M LASSO Sparse M

Figure 4: Scatterplots comparing $ZM^\top$ vs. $\mathbb{E}[Z \mid \hat{\Delta}, X]\hat{M}^\top$ between the different models under dense (top) and truly sparse (bottom) loadings M with $q = 100$ in simulations with batch effect.

We considered two main tasks: to give a visual representation of the latent factors of the data, i.e. an unsupervised dimension reduction task and a supervised survival analysis using the factors obtained in our method as predictions. Prior to our analyses, for ovarian and lung datasets we selected the 10% genes with highest total variance across all samples obtaining $p = 1{,}007$ for ovarian and $p = 1{,}198$ for lung, and we included the age at initial pathologic diagnosis as a covariate. For colorectal cancer we included tumour stage as covariate and we considered the $p = 172$ genes identified in Calon et al. (2012) as they are potentially related to the Transforming Growth Factor beta (TGF-β) pathway. TGF-β signatures are essential to a better understanding of colorectal cancers, as a large proportion of such cancers display a high production of TGF-β. All data sets have been normalised to zero mean and unit variance.

Unsupervised: data visualisation

Our first goal was to demonstrate the usefulness of our method as a data visualisation tool. We remark that there are no other model-based approaches to jointly adjust for batch effects and estimate latent factors. Thus, for comparison we first corrected the data using ComBat and then estimated the latent parameters via MLE and FastBFA akin to Section 5.2. To decide the number of factors for ComBat-MLE, we carried a principal component analysis to the corrected data prior to factor analysis and chose a number of components $\hat{q}$ that explained 90% or 70% of the total variance. It is important to notice that we are doing an over-optimistic assessment of ComBat-MLE and ComBat-FastBFA as we are doing a cross-validated factor analysis over the ComBat-corrected data, as opposed to also running ComBat in an out-of-sample fashion.

Figure 5 illustrates the advantages of our method. We can appreciate the usefulness of ComBat correction (middle panels) relative to the raw uncorrected data (top panels). ComBat removed systematic differences in location and scale across the two

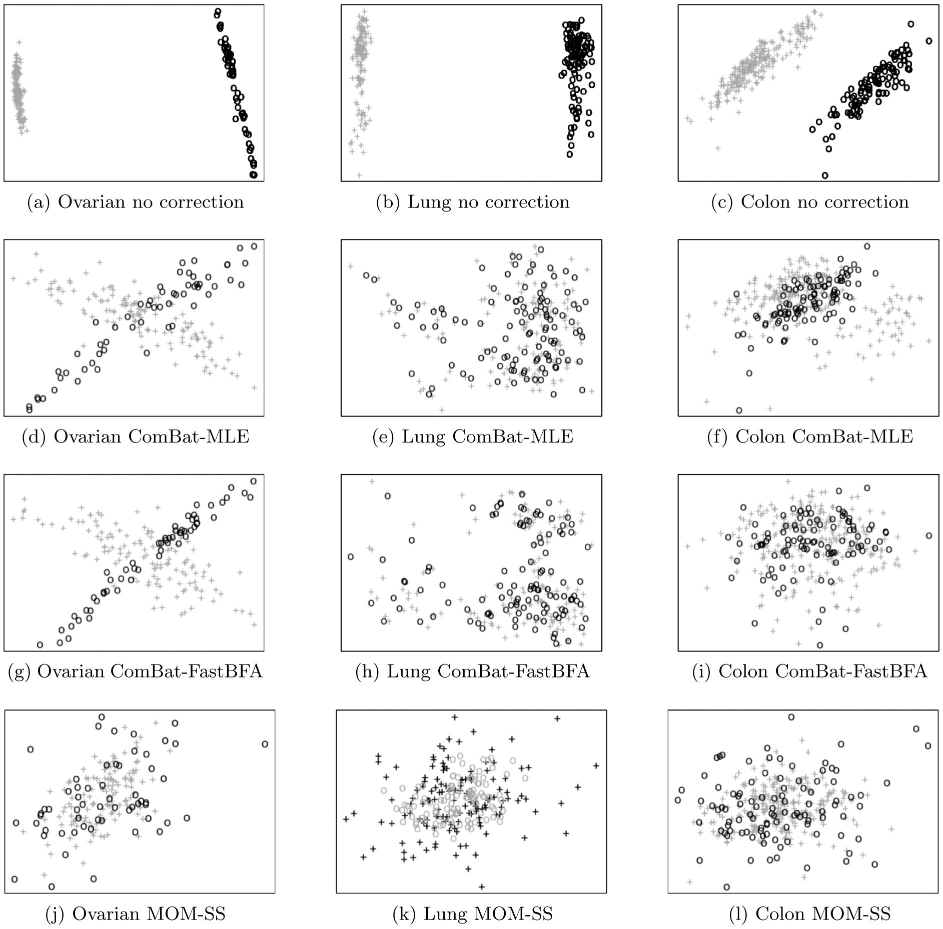

(a) Ovarian no correction
(b) Lung no correction
(c) Colon no correction
(d) Ovarian ComBat-MLE
(e) Lung ComBat-MLE
(f) Colon ComBat-MLE
(g) Ovarian ComBat-FastBFA
(h) Lung ComBat-FastBFA
(i) Colon ComBat-FastBFA
(j) Ovarian MOM-SS
(k) Lung MOM-SS
(l) Colon MOM-SS

Figure 5: Scatterplot of the first two factors of ovarian (left), lung (centre) and colon (right) datasets for the two different batches (grey pluses and black circles). Comparison between models without batch effect adjustment: ComBat-MLE, ComBat-FastBFA and MOM-SS.

batches. The latent coordinates displayed distinct covariances for the ovarian dataset which, given that patients in both batches are thought to be roughly exchangeable, is likely due to a technical artifact. Such distinct covariances were not present in the MOM-SS latent factors (bottom panels). In all panels we displayed the two factors that contributed the most to the covariance, i.e. with highest $\sum_{j=1}^{p} \hat{m}_{jk}^2$; the latent coordinates were post-processed to standardise their variance $\text{Cov}(\mathbf{z}_i \mid \hat{\Delta}, X)$, as explained in Section 4.4. For colon cancer, ComBat-MLE (5f) again showed differences in the covariance structure of the two different batches. Such differences were not present in the ComBat-FastBFA (5i) and MOM-SS (5l) latent factors, as expected due to the batch exchangeability. Differences between ComBat-MLE and ComBat-FastBE highlighted the variability of two-step methods. MOM-SS showed a robust performance across the different datasets.

Supervised: survival analysis

We also illustrate the potential of our method as a supervised tool, performing a survival analysis that aims to predict the time until death. To do that, we applied a Cox proportional hazards model (Cox, 1972) using as covariates the latent coordinates obtained in our models. We used the `coxph` function of the **R** package `survival 2.38` (Therneau, 2015). We then used the concordance index to assess the quality of our predictions. This index is a non-parametric metric to quantify the power of a prediction rule via a pair-wise comparison that measures the probability of concordance between the predicted and the observed survival time (Harrell Jr. et al., 1982). To obtain the concordance index we used the function `concordance.index` in the **R** package `survcomp` (Schröeder et al., 2011). The presented results are from 10 independent runs of 10-fold cross-validation. We initialised MOM-SS with the values obtained for the Flat model along with the other initialisations discussed in Section 4.3 and chose the one with smallest leave-one-out cross-validated concordance index.

We first performed survival analysis to each batch separately for all the cancer datasets (non-integrative analysis). We then conducted the integrative survival analysis. We finally compared the concordance index obtained with non-integrative and integrative analyses. Our results in Table 3 show that, for the ovarian cancer dataset, the MOM-SS integrative analysis displays competitive CI compared to the non-integrative survival analysis, with a sparser latent representation, requiring only 4 factors. Further, for lung and colorectal cancers, integrative MOM-SS shows a higher CI than the non-integrative survival analysis (from CI = 0.522 to CI = 0.665 for lung and from CI = 0.736 to CI = 0.764 for colorectal cancer). Our results highlight the importance of data integration and batch effect correction, as our integrative method is able to capture common latent information of different studies and provides competitive or higher CI compared to the single standard factor analysis.

For the integrative analyses of the cancer data sets, Table 3 shows that Flat-SS achieved a high concordance index, even though loadings are not sparse. Normal-SS gave sparse loading representations but displayed a concordance index lower than Flat-SS; this illustrates a lack of power to detect truly non-zero loadings. In general, MOM-SS provided sparse loadings and a good concordance index. In the ovarian cancer data, MOM-SS achieved a concordance index similar to ComBat-MLE 90% with considerably less factors (4 instead of 101) and a bit higher than Normal-SS. In the lung cancer data MOM-SS achieved a high concordance index, particularly relative to Normal-SS and ComBat-MLE 70%. In the colon cancer data the highest concordance index was obtained by MOM-SS. The competing methods generally lead to less sparse solutions and their performance fluctuates across scenarios. In the lung cancer data ComBat-MLE, despite its good performance, had a concordance index that proved to be sensitive to the number of factors (see ComBat-MLE 90% vs 70%). ComBat-FastBFA provided competitive results with a non-sparse reconstruction, recovering values in the latent loadings that were close to zero (even though not exactly zero) and smaller than the ones of the Flat-SS. MOM-SS proved to have practical advantages as a supervised tool in comparison with the two-step approaches considered here. Overall, MOM-SS provided a more stable performance that achieved a good balance between sparsity and prediction accuracy.

	Ovarian			Lung			Colon		
	$\hat{q}$	$\lvert\hat{M}\rvert_0$	CI	$\hat{q}$	$\lvert\hat{M}\rvert_0$	CI	$\hat{q}$	$\lvert\hat{M}\rvert_0$	CI
Batch 1-MLE 90%	67.1	67569.7	0.618	52.1	62415.8	0.461	52.9	9081.6	0.736
Batch 1-MLE 70%	27.0	27088.3	0.632	35.2	42169.6	0.471	17.0	2924.0	0.721
Batch 2-MLE 90%	40.4	40481.4	0.522	36.6	43607.2	0.522	48.1	8256.0	0.479
Batch 2-MLE 70%	23.4	23362.4	0.524	23.2	27913.4	0.419	23.3	4007.6	0.495
Flat	100.0	100700.0	0.634	100.0	119800.0	0.669	100.0	17200.0	0.594
Normal-SS	7.8	7854.6	0.568	11.0	13178.0	0.489	7.0	1204.0	0.621
MOM-SS	4.0	4028.0	0.588	74.0	88652.0	0.665	53.4	9184.8	0.764
ComBat-MLE 90%	101.0	101707.0	0.589	79.0	94642.0	0.688	67.0	11524.0	0.738
ComBat-MLE 70%	41.0	41287.0	0.588	30.0	35940.0	0.568	24.0	4128.0	0.734
ComBat-FastBFA	100.0	100700.0	0.527	100.0	119800.0	0.707	100.0	17200.0	0.582

Table 3: Survival analysis for ovarian (p = 1,007 genes), lung (p = 1,198 genes) and colon ($p = 172$ genes) cancer data sets.

6 Discussion

We have presented a novel model to integrate data from multiple sources using joint dimension reduction and batch effect adjustment via high-dimensional latent factor regression. We outlined three different prior configurations for the loadings and Laplace-tailed extensions whose deeper analysis remain as future work. To our knowledge this is the first time NLPs are implemented in the factor analysis context. We gave novel EM algorithms to obtain posterior modes. We showed that the use of sparse models increases the quality of our estimations even in the absence of batches. In our empirical results MOM-SS priors proved to be appealing, improving the estimation of factor cardinality and encouraging parsimony and selective shrinkage.

We illustrated the utility of our method in unsupervised and supervised frameworks. MOM-SS provided dimension reduction that corrected distinct covariance patterns present in two-stage methods that adjust variances separately from fitting the factor model. Such patterns are highly likely to be technical artefacts, since patients from different batches are believed to be exchangeable. Our model demonstrated to be useful for downstream analyses, achieving a competitive concordance indexes, in some cases with substantially less factors. It is important to notice that although our examples focus on gene expression of cancer datasets, the applications should also be useful in other settings.

We also remark that our novel MOM-SS and its closed-form EM updates can be extended to frameworks of interest beyond factor models such as: linear regression, generalised linear models as well as graphical models.

Our model assumes common factors across the datasets being integrated. An interesting extension for future research is to consider more complex settings where some of the factors differ across data sources or where one wishes to integrate datasets by adding variables (as opposed to adding individuals as we did here), or where potentially same variables were only recorded for a subset of the individuals.

Supplementary Material

Supplementary material to "Heterogeneous large datasets integration using Bayesian factor regression" (DOI: 10.1214/20-BA1240SUPP; .pdf). The supplementary materials are as follow: Proofs for Lemma 1, Lemma 2 and Lemma 3. EM algorithm under a flat, Normal-SS, MOM-SS, Laplace-SS and Laplace-MOM-SS on the loadings, a pseudo-code-algorithm for the weighted 10-fold cross-validation, and heatmaps for $\hat{M}$, $\widehat{\text{Cov}}(\mathtt{x}_i \mid \cdot)^{-1}$ and $\hat{\gamma}$ for the different simulated scenarios and setting $q = 100$. Robust analysis for simulations without batch effect and with batch effect.

References

Alter, O., Brown, P. O., and Botstein, D. (2000). "Singular value decomposition for genome-wide expression data processing and modeling." *Proceedings of the National Academy of Sciences*, 97(18): 10101–10106. URL http://www.pnas.org/content/97/18/10101.abstract 34

Avalos-Pacheco, A., Rossell, D., and Savage, R. S. (2020). "Supplement to "Heterogeneous large datasets integration using Bayesian factor regression"." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1240SUPP. 43, 47, 48, 50, 51, 54, 55

Avio, C. G., Gorbi, S., Milan, M., Benedetti, M., Fattorini, D., d'Errico, G., Pauletto, M., Bargelloni, L., and Regoli, F. (2015). "Pollutants bioavailability and toxicological risk from microplastics to marine mussels." *Environmental Pollution*, 198: 211–222. URL http://www.sciencedirect.com/science/article/pii/S0269749114005211 34

Bar, H., Booth, J., and Wells, M. T. (2018). "A scalable empirical Bayes approach to variable selection in generalized linear models." *arXiv:1803.09735*, 1–20. MR3920630. doi: https://doi.org/10.1002/wics.1455. 35

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., and Marron, J. S. (2004). "Adjustment of systematic microarray data biases." *Bioinformatics*, 20(1): 105–114. URL http://bioinformatics.oxfordjournals.org/content/20/1/105.abstract 34

Bentink, S., Haibe-Kains, B., Risch, T., Fan, J.-B., Hirsch, M. S., Holton, K., Rubio, R., April, C., Chen, J., Wickham-Garcia, E., Liu, J., Culhane, A., Drapkin, R., Quackenbush, J., and Matulonis, U. A. (2012). "Angiogenic mRNA and microRNA Gene Expression Signature Predicts a Novel Subtype of Serous Ovarian Cancer." *PLOS ONE*, 7(2): 1–9. doi: https://doi.org/10.1371/journal.pone.0030269. 55

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., and Milanesi, L. (2016). "Methods for the integration of multi-omics data: mathematical aspects." *BMC Bioinformatics*, 17(2): 167–177. doi: https://doi.org/10.1186/s12859-015-0857-9. 34

Burges, C. J. C. (2010). "Dimension Reduction: A Guided Tour." *Foundations and Trends in Machine Learning*, 2(4): 276–365. doi: https://doi.org/10.1561/2200000002. 34

Calon, A., Espinet, E., Palomo-Ponce, S., Tauriello, D. v., Iglesias, M., Céspedes, M. v., Sevillano, M., Nadal, C., Jung, P., Zhang, X. h.-F., Byrom, D., Riera, A., Rossell, D., and Mangues, R. (2012). "Dependency of Colorectal Cancer on a TGF-Beta-Driven Program in Stromal Cells for Metastasis Initiation." *Cancer Cell*, 22(5): 571–584. 50, 55, 57

Carvalho, C., Polson, N., and Scott, J. (2009). "Handling sparsity via the horseshoe." *Journal of Machine Learning Research*, 5: 73–80. 39

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics." *Journal of the American Statistical Association*, 103(484): 1438–1456. MR2655722. doi: https://doi.org/10.1198/016214508000000869. 34, 43

Cox, D. R. (1972). "Regression models and life-tables." *Journal of the Royal Statistical Society, Series B: Methodological*, 34: 187–220. MR0341758. 59

Cunningham, J. P. and Ghahramani, Z. (2015). "Linear Dimensionality Reduction: Survey, Insights, and Generalizations." *Journal of Machine Learning Research*, 16: 2859–2900. URL http://jmlr.org/papers/v16/cunningham15a.html. MR3450527. 34

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2018a). "Bayesian Multi-study Factor Analysis for High-throughput Biological Data." *arXiv:1806.09896*, 1–35. MR3953734. doi: https://doi.org/10.1111/biom.12974. 35

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2018b). "Multi-study Factor Analysis." *Biometrics*, 75: 337–346. MR3953734. doi: https://doi.org/10.1111/biom.12974. 35

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 39(1): 1–38. MR0501537. 43

Dunson, D. and Bhattacharya, A. (2011). "Sparse Bayesian infinite factor models." *Biometrika*, 98: 291–306. MR2806429. doi: https://doi.org/10.1093/biomet/asr013. 39

Ferriss, J. S., Kim, Y., Duska, L., Birrer, M., Levine, D. A., Moskaluk, C., Theodorescu, D., and Lee, J. K. (2012). "Multi-Gene Expression Predictors of Single Drug Responses to Adjuvant Chemotherapy in Ovarian Carcinoma: Predicting Platinum Resistance." *PLOS ONE*, 7(2): 1–9. doi: https://doi.org/10.1371/journal.pone.0030550. 55

Fortin, J.-P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M., and Shinohara, R. T.

(2016). "Removing inter-subject technical variability in magnetic resonance imaging studies." *NeuroImage*, 132: 198–212. 34

Fox, E. B. and Dunson, D. B. (2015). "Bayesian Nonparametric Covariance Regression." *Journal of Machine Learning Research*, 16: 2501–2542. URL http://jmlr.org/papers/v16/fox15a.html. MR3450515. 35

Frühwirth-Schnatter, S. and Lopes, H. F. (2018). "Sparse Bayesian Factor Analysis when the Number of Factors is Unknown." *arXiv:1804.04231*, 1–34. 37

Fúquene, J., Steel, M., and Rossell, D. (2018). "On choosing mixture components via non-local priors." *arXiv:1604.00314*, 1–72. MR4025398. doi: https://doi.org/10.1111/rssb.12333. 40

Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M., Parmigiani, G., Huttenhower, C., and Waldron, L. (2013). "curatedOvarianData: Clinically Annotated Data for the Ovarian Cancer Transcriptome." *Database*, 2013. URL http://database.oxfordjournals.org/content/2013/bat013.abstract 50

George, E. and McCulloch, R. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 35, 39, 41

George, E. and McCulloch, R. (1997). "Approaches for Bayesian variable selection." *Statistica Sinica*, 339–374. 41

Ghahramani, Z. and Beal, M. J. (2000). "Variational Inference for Bayesian Mixtures of Factor Analysers." In Solla, S. A., Leen, T. K., and Müller, K. (eds.), *Advances in Neural Information Processing Systems 12*, 449–455. MIT Press. URL http://papers.nips.cc/paper/1672-variational-inference-for-bayesian-mixtures-of-factor-analysers.pdf 43

Goh, W. W. B., Wang, W., and Wong, L. (2017). "Why Batch Effects Matter in Omics Data, and How to Avoid Them." *Trends in Biotechnology*, 35: 498–507. 34

Griffiths, T. L. and Ghahramani, Z. (2011). "The Indian Buffet Process: An Introduction and Review." *J. Mach. Learn. Res.*, 12: 1185–1224. MR2804598. 50

Harrell Jr., F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). "Evaluating the yield of medical tests." *JAMA*, 247(18): 2543–2546. 59

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc. MR2722294. doi: https://doi.org/10.1007/978-0-387-84858-7. 33

Hirose, K. and Yamamoto, M. (2015). "Sparse estimation via nonconcave penalized likelihood in factor analysis model." *Statistics and Computing*, 25(5): 863–875. MR3375622. doi: https://doi.org/10.1007/s11222-014-9458-0. 50

Hirose, K., Yamamoto, M., and Nagata, H. (2016). *fanc: Penalized Likelihood Factor Analysis via Nonconvex Penalty*. R package version 2.2. URL https://CRAN.R-project.org/package=fanc 50

Hoff, P. and Niu, X. (2012). "A Covariance Regression Model." *Statistica Sinica*, 22: 729–753. URL `http://www.stat.washington.edu/hoff/Code/hoff_niu_2009_ss`. MR2954359. doi: `https://doi.org/10.5705/ss.2010.051`. 35

Hornung, R., Boulesteix, A.-L., and Causeur, D. (2016). "Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment." *BMC Bioinformatics*, 17(1): 1–19. doi: `https://doi.org/10.1186/s12859-015-0870-z`. 34

Johnson, R. A. and Wichern, D. W. (eds.) (1988). *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc. MR2372475. 33

Johnson, V. E. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(2): 143–170. MR2830762. doi: `https://doi.org/10.1111/j.1467-9868.2009.00730.x`. 35, 40

Johnson, V. E. and Rossell, D. (2012). "Bayesian Model Selection in High-Dimensional Settings." *Journal of the American Statistical Association*, 107(498): 649–660. MR2980074. doi: `https://doi.org/10.1080/01621459.2012.682536`. 35, 40

Johnson, W. E. and Li, C. (2009). *Adjusting Batch Effects in Microarray Experiments with Small Sample Size Using Empirical Bayes Methods*, 113–129. John Wiley & Sons, Ltd. doi: `https://doi.org/10.1002/9780470685983.ch10`. 34

Johnson, W. E., Li, C., and Rabinovic, A. (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods." *Biostatistics (Oxford, England)*, 8(1): 118–27. URL `http://www.ncbi.nlm.nih.gov/pubmed/16632515` 34, 36, 50

Kaiser, H. F. (1958). "The varimax criterion for analytic rotation in factor analysis." *Psychometrika*, 23(3): 187–200. doi: `https://doi.org/10.1007/BF02289233`. 37

Knowles, D. A. and Ghahramani, Z. (2011). "Nonparametric Bayesian sparse factor models with application to gene expression modeling." *The Annals of Applied Statistics*, 5(2B): 1534–1552. MR2849785. doi: `https://doi.org/10.1214/10-AOAS435`. 39

Leek, J. T., Johnson, W. E., Parker, H. S., Fertig, E. J., Jaffe, A. E., Storey, J. D., Zhang, Y., and Torres, L. C. (2017). *sva: Surrogate Variable Analysis*. R package version 3.26.0. 51

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data." *Nat Rev Genet*, 11(10): 733–739. doi: `https://doi.org/10.1038/nrg2825`. 34

Leek, J. T. and Storey, J. D. (2007). "Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis." *PLoS Genet*, 3(9): 1–12. 34

Lopes, H. F. and West, M. (2004). "Bayesian model assessment in factor analysis." *Statistica Sinica*, 14: 41–67. MR2036762. 34, 37, 43

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2006). "Sparse statistical modelling in gene expression genomics." In *Bayesian Inference for Gene*

Expression and Proteomics, 155–176. Cambridge University Press. MR2655722. doi: https://doi.org/10.1198/016214508000000869. 34

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian Variable Selection in Linear Regression." *Journal of the American Statistical Association*, 83(404): 1023–1032. MR0997578. doi: https://doi.org/10.1080/01621459.1988.10478694. 39

Olivetti, E., Greiner, S., and Greiner, S. (2012). "ADHD diagnosis from multiple data sources with batch effects." *Frontiers in Systems Neuroscience*, 6: 1662–5137. 34

Parker, H. S., Corrada Bravo, H., and Leek, J. T. (2014). "Removing batch effects for prediction problems with frozen surrogate variable analysis." *PeerJ*, 2: e561. doi: https://doi.org/10.7717/peerj.561. 34

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. M. (2004). "Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression." *Proceedings of the National Academy of Sciences of the United States of America*, 101(25): 9309–9314. URL http://www.pnas.org/content/101/25/9309.abstract 34

Rossell, D. and Telesca, D. (2017). "Nonlocal Priors for High-Dimensional Estimation." *Journal of the American Statistical Association*, 112(517): 254–265. MR3646569. doi: https://doi.org/10.1080/01621459.2015.1130634. 40

Ročková, V. and George, E. I. (2014). "EMVS: The EM Approach to Bayesian Variable Selection." *Journal of the American Statistical Association*, 109(506): 828–846. MR3223753. doi: https://doi.org/10.1080/01621459.2013.869223. 39, 41, 46

Ročková, V. and George, E. I. (2017). "Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity." *Journal of the American Statistical Association*, 111(516): 1608–1622. MR3601721. doi: https://doi.org/10.1080/01621459.2015.1100620. 34, 35, 37, 38, 39, 43, 50, 51

Ročková, V. and George, E. I. (2018). "The Spike-and-Slab LASSO." *Journal of the American Statistical Association*, 113(521): 431–444. MR3803476. doi: https://doi.org/10.1080/01621459.2016.1260469. 39, 40, 41

Schadt, E. E., Li, C., Ellis, B., and Wong, W. H. (2001). "Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data." *Journal of Cellular Biochemistry*, 84(S37): 120–125. doi: https://doi.org/10.1002/jcb.10073. 34

Scherer, A. (2009). *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. Wiley Series in Probability and Statistics. Wiley. 34

Schröeder, M. S., Culhane, A., Quackenbush, J., and Haibe-Kains, B. (2011). "survcomp: an R/Bioconductor package for performance assessment and comparison of survival models." *Bioinformatics*, 27(22): 3206–3208. 59

Schwarz, G. (1978). "Estimating the Dimension of a Model." *Ann. Statist.*, 6(2): 461–464. MR0468014. doi: https://doi.org/10.1214/aos/1176344136. 38

Seber, G. (1984). *Multivariate observations*. Wiley series in probability and mathematical statistics. New York, NY: Wiley. MR0746474. doi: https://doi.org/10.1002/9780470316641. 37

Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D. L., Collins, D. L., and Arbel, T. (2011). "Evaluating intensity normalization on MRIs of human brain with multiple sclerosis." *Medical Image Analysis*, 15(2): 267–282. URL http://www.sciencedirect.com/science/article/pii/S1361841510001337 34

Shi, G., Lim, C. Y., and Maiti, T. (2019). "Model selection using mass-nonlocal prior." *Statistics & Probability Letters*, 147(C): 36–44. URL https://ideas.repec.org/a/eee/stapro/v147y2019icp36-44.html. MR3892045. doi: https://doi.org/10.1016/j.spl.2018.11.027. 35

Shinohara, R. T., Sweeney, E. M., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., Jarso, S., Pham, D. L., Reich, D. S., and Crainiceanu, C. M. (2014). "Statistical normalization techniques for magnetic resonance imaging." *NeuroImage : Clinical*, 6: 9–19. 34

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. Version 2.38. URL https://CRAN.R-project.org/package=survival 59

Wan, Y.-W., Allen, G. I., Anderson, M. L., and Liu, Z. (2015). *TCGA2STAT: Simple TCGA Data Access for Integrated Statistical Analysis in R*. R package version 1.2. URL https://CRAN.R-project.org/package=TCGA2STAT 50

Wan, Y.-W., Allen, G. I., and Liu, Z. (2016). "TCGA2STAT: simple TCGA data access for integrated statistical analysis in R." *Bioinformatics*, 32(6): 952–954. doi: https://doi.org/10.1093/bioinformatics/btv677. 55

Wang, J. and Zhao, Q. (2015). *cate: High Dimensional Factor Analysis and Confounder Adjusted Testing and Estimation*. R package version 1.0.4. URL https://CRAN.R-project.org/package=cate 51

West, M. (2003). "Bayesian factor regression models in the "large p, small n" paradigm." In *Bayesian Statistics 7*, 723–732. Oxford University Press. URL http://ftp.isds.duke.edu/WorkingPapers/02-12.html. MR2003537. 43

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis." *Biostatistics*, 10(3): 515–534. 39

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation." *Nucleic Acids Research*, 30(4): e15. URL http://nar.oxfordjournals.org/content/30/4/e15.abstract 34

Acknowledgments

We thank Chris Yau for valuable insights and Veronika Ročková for providing the FastBFA package.