# Multi-study factor regression biclustering

Margarita Grushanina[*] Leonardo Bottolo[†] Verena Zuber[‡]

Sylvia Frühwirth-Schnatter[§]

July 20, 2026

**Abstract**

High-dimensional molecular feature data is often acquired across multiple research entities under varying conditions and platforms and therefore arises from multiple studies. Existing biclustering methods largely ignore differences between studies and are not designed to explicitly separate shared biological structure from study-specific variation. We propose a multi-study factor regression biclustering (MSFRB) model, which unifies multi-study factor analysis, factor regression, and biclustering within a single Bayesian framework. The model enables the identification of subsets of features that exhibit coherent patterns in subsets of samples while explicitly separating shared and study-specific sources of variation. Latent factors are decomposed into common factors that capture shared biological structure across studies and study-specific factors that account for heterogeneity arising from batch effects, study design, or technological differences. To induce bicluster structure, we place a sparse exchangeable shrinkage process (ESP) prior on the loadings and scores of the common factors. The sparse ESP prior promotes structured sparsity while enabling automatic determination of the number of active factors. The model also incorporates covariates by including a regression component, thus allowing adjustment for observed sample characteristics. To address the identifiability challenges inherent in multi-study factor models, we derive a sufficient identification condition based on unique factor support and develop a post hoc procedure for assessing whether recovered factors satisfy this condition. For scalable inference, we derive an efficient variational algorithm. Simulation studies and an application to multi-study colorectal cancer gene expression data demonstrate improved recovery of shared latent structure compared with existing biclustering approaches.

**Keywords:** Factor analysis, biclustering, gene expression, multi-study, shrinkage, spike-and-slab, variational Bayes

## 1 Introduction

High-dimensional molecular profiling technologies play a central role in modern biomedical research. Platforms such as transcriptomics, proteomics, and metabolomics provide measurements on

[*]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom. Email: m.grushanina@imperial.ac.uk

[†]Department of Genomic Medicine, University of Cambridge, Cambridge, UK. Email: lb664@cam.ac.uk

[‡]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. Email: v.zuber@imperial.ac.uk

[§]Institute for Statistics and Mathematics, Vienna University of Economics and Business, Vienna, Austria. Email: sfruehwi@wu.ac.at

thousands of molecular features across biological samples, often collected in multiple studies or cohorts, with different experimental conditions. Understanding the latent biological mechanisms that jointly regulate subsets of these features remains a fundamental statistical challenge.

A recurring empirical phenomenon in molecular data is the presence of coordinated activity among subsets of molecular features such as genes, proteins, or metabolites that manifest itself only in specific subsets of samples. Biclustering methods are designed to capture such structure by simultaneously clustering rows and columns of the data matrix, enabling the identification of overlapping feature–sample associations linked to latent biological processes (Pontes et al., 2015).

Since the pioneering work of Hartigan (1972), who introduced simultaneous clustering of rows and columns of a data matrix, and Cheng and Church (2000), who popularised biclustering in gene expression analysis, a wide range of biclustering methods has been developed. Following Madeira and Oliveira (2004), these methods can be broadly classified into approaches based on constant submatrices (e.g. Hartigan, 1972; Prelić et al., 2006), additive effects (e.g. Cheng and Church, 2000; Lazzeroni and Owen, 2002), low-rank factorisations (e.g. Kluger et al., 2003), and pattern-based structures (e.g. Tanay et al., 2002; Li et al., 2009). While these approaches differ in their assumptions about bicluster geometry and signal generation, they share the goal of identifying subsets of features that exhibit coherent behaviour within subsets of samples. Reviews of the biclustering literature are provided by Madeira and Oliveira (2004), Eren et al. (2012), and Padilha and Campello (2017).

Among these approaches, the methods most directly related to the present work are biclustering models based on sparse factorisations, where biclusters are represented through latent factors and their associated loadings, and overlapping structure is induced through sparsity on both sides of the factorisation. A seminal example is FABIA (Hochreiter et al., 2010). More recent Bayesian formulations include BicMix (Gao et al., 2016), which places a three-layer shrinkage prior on factor scores and loadings; spike-and-slab biclustering models such as Denitto et al. (2017) and Moran et al. (2021), with the latter also incorporating an Indian Buffet Process (IBP) prior to automatically infer the number of biclusters; and empirical Bayes matrix factorisation (EBMF, Wang and Stephens, 2021), where the degree of sparsity is learned from the data.

While these methods have proved useful in single-study settings, they do not explicitly account for heterogeneity across multiple studies, nor do they incorporate observed covariate effects within a unified biclustering framework. Among them, BicMix (Gao et al., 2016) is unique in explicitly accommodating broad confounding effects by allowing latent factors to be either sparse or dense through a combination of spike-and-slab and three parameter beta (TPB) priors on factor scores and loadings. Dense factors can then absorb broad sources of variation such as batch effects, whereas sparse factors capture more localised structure. In contrast, Moran et al. (2021) focus on recovering interpretable biclustering structure and automatically inferring the number of biclusters, whereas Wang and Stephens (2021) emphasise flexible empirical Bayes learning of sparsity patterns directly from the data. Nevertheless, none of these approaches explicitly distinguish shared biological structure from study-specific variation. Besides batch effects and other technology-related sources of variation, such heterogeneity may arise from differences in experimental platforms and protocols, patient populations, sample collection procedures, or study design (see e.g. Liang et al., 2026).

In this paper, we propose a multi-study factor regression biclustering (MSFRB) model, a unified framework that simultaneously accounts for covariate effects, distinguishes between shared and study-specific sources of variation, and recovers latent bicluster structure within a Bayesian sparse factor analysis setting. To model study-specific variation, we build on the multi-study factor analysis (MSFA) framework introduced by De Vito et al. (2019) in a frequentist setting and later extended

to high-dimensional Bayesian inference (De Vito et al., 2021; Hansen et al., 2025). MSFA assumes that some latent factors are shared across studies, while others are specific to individual studies. This leads to a decomposition into common and study-specific factors, where the latter account for study-dependent variability arising from batch effects, measurement noise, and other sources of heterogeneity. To account for observed sample characteristics such as age and sex, we further extend the model by incorporating a design matrix of known covariates alongside the common and study-specific latent factors and their loadings. This extension is inspired by the multi-study factor regression (MSFR) model of De Vito and Avalos-Pacheco (2025), developed in the context of traditional factor analysis.

To recover bicluster structure, we impose sparsity on both the loadings and scores of the common factors, thereby identifying subsets of features that exhibit coordinated behaviour within subsets of samples. For this purpose, we employ a sparse version of the exchangeable shrinkage process (ESP) prior (Frühwirth-Schnatter, 2023) on the variances of factor loadings and factor scores. Similar in spirit to the hierarchical shrinkage constructions used in sparse factor models such as BicMix (Gao et al., 2016), the ESP prior combines element-wise sparsity with explicit factor-level model selection, enabling automatic determination of the number of active factors. This yields a clear distinction between active and inactive factors while allowing the sparsity structure within active factors to be learned from the data. Study-specific factor loadings are assigned similar priors, although with a simpler and less restrictive sparsity formulation, while the corresponding study-specific factor scores follow the standard factor analysis assumption of independent Gaussian distributions with zero mean and unit variance. Consequently, biclustering is induced only in the common component, whereas the study-specific component is primarily used to capture study-dependent heterogeneity.

Because MSFA decomposes variation into shared and study-specific components, it introduces an additional source of unidentifiability beyond the classical rotational indeterminacy of factor models. In particular, these models may suffer from information switching, whereby shared structure can be represented by study-specific factors and vice versa, leading to observationally indistinguishable parameterisations (Chandra et al., 2025; Bortolato and Canale, 2024). Addressing this issue typically requires additional assumptions or constraints. For example, Chandra et al. (2025) achieve identifiability by constraining study-specific factors to lie within the shared loading subspace, thereby separating shared and study-specific covariance components. In a less restrictive approach, Bortolato and Canale (2024) impose distinct activation patterns across studies, so that each factor is uniquely determined by its study-specific support. We approach this problem differently and instead exploit the sparse support patterns induced by the biclustering formulation. In particular, when a factor possesses support that is not shared by any other common or study-specific factor, the corresponding factor can be uniquely distinguished. This observation motivates a sufficient identification condition based on unique factor support. Accordingly, we assess identifiability post hoc rather than enforcing it during model fitting.

The rest of the paper is organised as follows. Section 2.4 explores biclustering with sparse factor models and defines biclustering framework with the ESP prior. Section 2 extends this framework to the multi-study factor regression setting. Identification is handled in Section 4.An efficient variational inference scheme is provided in Section 5. The model's performance is illustrated on extensive simulation studies in Section 6 and real data application in Section 7. Initialisations, hyperparameter selection and the detailed steps of the variational algorithms as well as the identification procedure are discussed in the Supplementary material. The paper concludes in Section 8.

## 2 Multi-study factor regression biclustering

In this section, we introduce the proposed multi-study factor regression biclustering (MSFRB) model. We begin with a review of sparse Bayesian factor analysis and then show how this framework can be extended to accommodate multiple studies and observed covariates. Finally, we describe how biclustering structure can be incorporated through sparsity priors on both factor loadings and factor scores.

### 2.1 Sparse Bayesian Factor Analysis

The main purpose of factor analysis is to find a lower-dimensional representation of the data by identifying latent factors that account for the observed variation. By leveraging sparsity-inducing priors, one can achieve automatic detection of the latent dimensionality, as well as a more interpretable representation of the latent factors.

Suppose we have a basic factor model:

$$\boldsymbol{Y} = \boldsymbol{\Lambda}\boldsymbol{F} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{Y}$ is a $p \times N$ matrix of observed values, with $i = 1, \ldots, p$ indexing variables and $n = 1, \ldots, N$ indexing observations. Further, $\boldsymbol{\Lambda}$ is a $p \times H$ factor loading matrix and $\boldsymbol{F}$ is an $H \times N$ matrix of latent factor scores, where $\boldsymbol{f}_n$ denotes the $n$th column of $\boldsymbol{F}$ and contains the factor scores associated with observation $n$. Finally, $\boldsymbol{\epsilon}$ is a $p \times N$ matrix of idiosyncratic errors following a normal distribution with variable-specific variances, $\boldsymbol{\epsilon}_n \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Xi})$, where $\boldsymbol{\Xi} = \mathrm{diag}(\xi_1^2, \ldots, \xi_p^2)$. The model assumes that the variability of the data can be described by $H$ latent factors, with $H \ll p$. In Bayesian factor analysis, it is typically assumed that the latent factors are mutually uncorrelated and have isotropic variance,, namely

$$\mathbf{f}_n \sim \mathcal{N}_H(\mathbf{0}, \mathbf{I}_H), \qquad \mathbf{f}_n \in \mathbb{R}^H \tag{2.2}$$

and that $\mathbf{f}_n$, $\mathbf{f}_t$, $\boldsymbol{\epsilon}_n$ and $\boldsymbol{\epsilon}_t$ are pairwise independent for all $n \neq t$. A common prior specification for the factor loadings is Gaussian $\lambda_{ih} \sim \mathcal{N}(0, \sigma_{ih}^2)$, for $i = 1, \ldots, p$ and $h = 1, \ldots, H$.

Sparsity is typically induced by assigning suitable priors to the factor loadings, $\lambda_{ih}$, or to their variances, $\sigma_{ih}^2$. In the context of factor analysis, sparsity often refers to element-wise sparsity, whereby the loadings associated with a particular factor $h$, represented by the column $\boldsymbol{\lambda}_h$ of the loading matrix $\boldsymbol{\Lambda}$, contain many zero or near-zero elements. In this case, only a subset of variables contributes to a given factor, which facilitates interpretation of the latent structure. A related but distinct notion is column-wise sparsity, which is closely linked to the automatic determination of the latent dimensionality $H$. Under an overfitted or infinite factor model, column-wise sparsity implies that only a subset of columns of $\boldsymbol{\Lambda}$ remain active, while the remaining columns are shrunk entirely to zero. Consequently, the number of active columns determines the effective dimension of the factor model.

There is a vast body of literature on sparse Bayesian factor analysis (SBFA) devoted to inducing sparsity and automatically determining the number of latent factors. A common approach is to combine element-wise and column-wise sparsity mechanisms, with the latter enabling automatic factor selection through priors that increasingly penalize higher-dimensional models. Such priors are typically assigned to variances of factor loadings, with examples including multiplicative gamma process priors (Bhattacharya and Dunson, 2011) and spike-and-slab formulations (Legramanti et al., 2020;

Kowal and Canale, 2023; Frühwirth-Schnatter, 2023). Alternative approaches impose sparsity directly on the factor loadings rather than through their variances (see, e.g. Knowles and Ghahramani (2011) and Ročková and George (2016) which employ Indian buffet process prior). Some models additionally incorporate global shrinkage parameters or hierarchical sparsity structures combining global, column-wise and local shrinkage (Gao et al., 2013; Schiavon et al., 2022; Frühwirth-Schnatter et al., 2025).

## 2.2 Extension to multi-study framework

As already mentioned in Section 1, molecular feature data are often measured across multiple cohorts, laboratories, platforms, or populations. This results in separate datasets, each of which may differ from the others in terms of the technology used, population characteristics, or laboratory conditions. Throughout the paper, we assume that these datasets correspond to different studies, indexed by $s = 1, \ldots, S$. Each study may contain a different number of observations (samples), denoted by $N_s$, but the same set of variables (features), indexed by $i = 1, \ldots, p$, is measured in every study. As illustrated in Figure 1, each dataset from study $s$ has dimension $p \times N_s$ and stacked together they form a data set $p \times N$, where $N = N_1 + \cdots + N_S$.

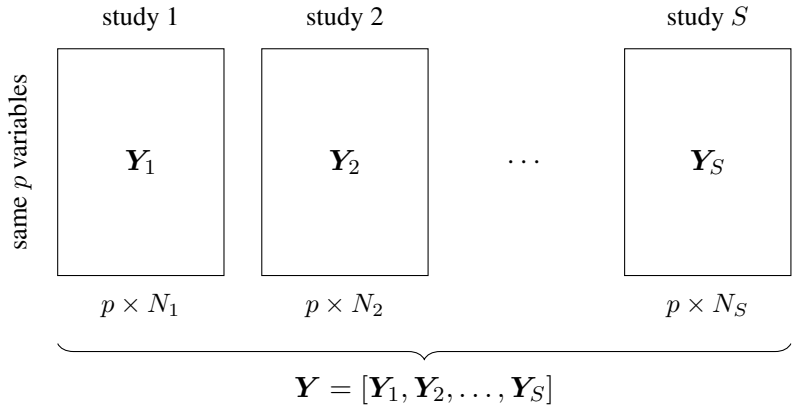



Figure 1: *Multi-study data structure: the same set of $p$ variables is measured across $S$ studies, with $N_s$ samples in study $s$.*

To separate latent structure into components that are shared across studies and components that are specific to individual studies, MSFA (De Vito et al., 2019, 2021) introduces two sets of latent factors. Common latent factors, denoted by $\boldsymbol{f}_{ns} \in \mathbb{R}^H$ for $n = 1, \ldots, N_s$ and $s = 1, \ldots, S$, capture variation shared across all studies, where $H$ is the dimension of the common latent space. In addition, study-specific latent factors, denoted by $\boldsymbol{\eta}_{ns} \in \mathbb{R}^{H_s}$ for $n = 1, \ldots, N_s$ and $s = 1, \ldots, S$, capture variation unique to study $s$, where $H_s$ is the dimension of the correspondong latent space. Each study is associated with its own loading matrix $\boldsymbol{\Lambda}_s \in \mathbb{R}^{p \times H_s}$, which links the study-specific factors to the observed variables. This decomposition allows information to be shared across studies through the common factors while simultaneously accounting for study-specific heterogeneity through the study-specific factors.

The resulting Bayesian MSFA model can then be written as (Liang et al., 2026):

$$\boldsymbol{y}_{ns} = \boldsymbol{\Lambda} \boldsymbol{f}_{ns} + \boldsymbol{\Lambda}_s \boldsymbol{\eta}_{ns} + \boldsymbol{\epsilon}_{ns}, \tag{2.3}$$

where $\boldsymbol{y}_{ns}$ denotes sample $n$ from study $s$. Since the model explicitly decomposes variation into common and study-specific components, the idiosyncratic noise term is also allowed to vary across studies. Specifically, $\boldsymbol{\epsilon}_{ns} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Xi}_s)$, where $\boldsymbol{\Xi}_s = \mathrm{diag}(\xi_{1s}^2, \ldots, \xi_{ps}^2)$.

Let $\boldsymbol{F}_s \in \mathbb{R}^{H \times N_s}$ denote the matrix of common factor scores for study $s$, whose columns correspond to the vectors $\boldsymbol{f}_{ns}$, $n = 1, \ldots, N_s$. Similarly, let $\boldsymbol{\eta}_s \in \mathbb{R}^{H_s \times N_s}$ denote the matrix of study-specific factor scores and $\boldsymbol{\epsilon}_s \in \mathbb{R}^{p \times N_s}$ the matrix of idiosyncratic errors. The model can then be written in matrix form as

$$[\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_S] = \boldsymbol{\Lambda} [\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_S] + [\boldsymbol{\Lambda}_1 \boldsymbol{\eta}_1, \boldsymbol{\Lambda}_2 \boldsymbol{\eta}_2, \ldots, \boldsymbol{\Lambda}_S \boldsymbol{\eta}_S] + [\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_S] .$$

The above discussion provides only a brief overview of Bayesian multi-study factor models. For a more comprehensive treatment, the reader is referred to De Vito et al. (2021), Hansen et al. (2025), and Liang et al. (2026).

## 2.3 Multi-study factor regression

To account for the confounding effects of observed covariates, such as age or sex of the individuals from whom the molecular samples were obtained, we include these variables as known regressors in the model. This formulation, known as a factor regression model, was introduced in the Bayesian factor analysis context by Lucas et al. (2006), further developed in Avalos-Pacheco et al. (2022), and recently extended to the multi-study setting in De Vito and Avalos-Pacheco (2025).

Let $p_X$ denote the number of observed covariates. The multi-study factor regression model can then be written as

$$\boldsymbol{y}_{ns} = \boldsymbol{\beta}^X \boldsymbol{x}_{ns} + \boldsymbol{\Lambda} \boldsymbol{f}_{ns} + \boldsymbol{\Lambda}_s \boldsymbol{\eta}_{ns} + \boldsymbol{\epsilon}_{ns}, \tag{2.4}$$

where $\boldsymbol{x}_{ns} \in \mathbb{R}^{p_X}$ denotes the covariate vector associated with sample $n$ from study $s$, and $\boldsymbol{\beta}^X \in \mathbb{R}^{p \times p_X}$ is the matrix of regression coefficients. The remaining quantities are defined as in the multi-study factor model in Equation 2.3.

## 2.4 Biclustering through common factors

The multi-study factor regression model introduced above accounts for covariate effects and separates shared and study-specific sources of variation. Our primary interest, however, lies in identifying subsets of variables and observations that participate in the shared signal. To this end, we focus on the sparsity of the common component $\boldsymbol{\Lambda F}$ and exploit the connection between sparse factor models and biclustering. Biclustering is achieved by replacing the isotropic Gaussian prior in Equation (2.2) with a sparsity-inducing prior on the factor scores, analogous to that used for the factor loadings (see, for example, Hochreiter et al. (2010), Gao et al. (2016), Moran et al. (2021), and Wang and Stephens (2021)). This mirrors the sparsity structure on both sides of the common signal, so that each factor is active only for a subset of variables and a subset of observations. Consequently, each common factor defines a bicluster, allowing simultaneous clustering of variables and observations. Since a variable or observation may be active in multiple factors, the resulting biclusters are not required to be disjoint and may overlap.

Various sparsity-inducing priors can be used to induce biclustering. While sparsity may be imposed directly on the factor loadings and factor scores (see, e.g., Moran et al., 2021), we instead adopt

a variance decomposition approach in which the variances of the factor loadings and factor scores are expressed as products of multiple sparsity components. These components correspond to different levels of shrinkage and are each assigned their own sparsity-inducing prior. This approach is closely related to the biclustering formulation of Gao et al. (2016), which also decomposes the variance into multiple sparsity levels, although the latter adopts a hierarchical construction whereas we employ a multiplicative formulation. More specifically,

$$\begin{aligned} \lambda_{ih}|\sigma^2_{ih} &\sim \mathcal{N}(0, \sigma^2_{ih}), \qquad \sigma^2_{ih} = \mathfrak{f}(\kappa, \theta_h, \phi_{ih}), \\ f_{hn}|\omega^2_{hn} &\sim \mathcal{N}(0, \omega^2_{hn}), \qquad \omega^2_{hn} = \mathfrak{f}(\gamma, \upsilon_h, \zeta_{hn}), \end{aligned} \tag{2.5}$$

where $\kappa$ and $\gamma$ represent global shrinkage parameters, $\phi_{ih}$ and $\zeta_{hn}$ control local (element-wise) sparsity, and $\theta_h$ and $\upsilon_h$ control sparsity at the latent factor level.

# 3 Biclustering with sparse exchangeable shrinkage process prior

## 3.1 Prior specification for global and local sparsity components

As discussed in Section 2.4, biclustering is achieved through sparsity in both the common factor loadings and the common factor scores. We decompose their variances into global, factor-level and local sparsity components:

$$\begin{aligned} \lambda_{ih} \,|\, \theta_h, \phi_{ih}, \kappa_\Lambda &\sim \mathcal{N}(0, \theta_h \phi_{ih} \kappa_\Lambda), \qquad i = 1, \ldots, p, \;\; h = 1, \ldots, H, \\ f_{hns} \,|\, \upsilon_h, \zeta_{hns}, \kappa_F &\sim \mathcal{N}(0, \upsilon_h \zeta_{hns} \kappa_F), \qquad n = 1, \ldots, N_s, \;\; s = 1, \ldots, S, \;\; h = 1, \ldots, H. \end{aligned} \tag{3.1}$$

Here, $\kappa_\Lambda$ and $\kappa_F$ control global shrinkage, $\phi_{ih}$ and $\zeta_{hns}$ induce local (element-wise) sparsity within the loadings and scores of factor $h$, and $\theta_h$ and $\upsilon_h$ govern factor-level sparsity by controlling which latent factors remain active.

For the global and local sparsity components, we adopt inverse Gamma priors commonly used in Bayesian factor analysis. To allow the degree of local sparsity to be learned from the data, each local sparsity parameter is assigned its own hyperprior at the variable or observation level:

$$\begin{aligned} \kappa_\Lambda &\sim \mathcal{G}^{-1}(a^\kappa, b^\kappa), \qquad \phi_{ih} \sim \mathcal{G}^{-1}(a_\phi, \beta_{i\phi}), \qquad \beta_{i\phi} \sim \mathcal{G}(c, d), \\ \kappa_F &\sim \mathcal{G}^{-1}(c^\kappa, d^\kappa), \qquad \zeta_{hns} \sim \mathcal{G}^{-1}(a_\zeta, \beta_{\zeta ns}), \qquad \beta_{\zeta ns} \sim \mathcal{G}(e, w). \end{aligned}$$

## 3.2 Exchangeable shrinkage process prior

The factor-level sparsity parameters $\theta_h$ and $\upsilon_h$, play a central role in the proposed formulation, as they determine which latent factors remain active and which are shrunk towards zero. To model these quantities, we employ the exchangeable shrinkage process (ESP) prior developed by Frühwirth-Schnatter (2023). The ESP prior takes the form of a sequence of unordered spike-and-slab distributions and enables selection of the columns of the factor loading matrix (equivalently, the rows of the factor score matrix) corresponding to active factors.

Let $H$ denote the maximum possible number of active latent common factors. Then the ESP prior is defined as follows. Assume $\boldsymbol{\tau} = \{\tau_h \in (0,1) : h = 1, \ldots, H\}$ to be a finite sequence of i.i.d. random probabilities. Let $\Theta = \{\theta_h\}$, $h = 1, \ldots, H$ be a finite sequence of column-specific shrinkage

parameters assumed to be independent conditional on $\boldsymbol{\tau}$ and independent of all $\tau_l$, $l \neq h$ for all $h$. If the prior $\theta_h|\tau_h$ takes the following spike-and-slab form:

$$\theta_h|\tau_h \sim (1-\tau_h)p_{spike}(\theta_h|\varphi_0) + \tau_h p_{slab}(\theta_h|\varphi_\theta), \tag{3.2}$$

where $\varphi_0$ and $\varphi_\theta$ are the hyperparameters of the spike and the slab distributions, respectively, then $\Theta$ follows an ESP prior. The prior is exchangeable in the sense that it is invariant to permutations of the column indices $h$.

It is often assumed in the literature that the slab probabilities $\tau_1, \ldots, \tau_H$ follow a beta distribution, where the first parameter depends on $H$. Following Frühwirth-Schnatter (2023), we assign the slab probabilities the prior

$$\tau_h|H, \alpha \sim \mathcal{B}\left(\frac{\alpha}{H}, 1\right), \qquad h = 1, \ldots, H. \tag{3.3}$$

This corresponds to the finite Beta-process construction of Ročková and George (2016) and converges to the Indian Buffet Process prior as $H \to \infty$ (Teh, 2006). An important property of the ESP prior is that it admits a finite generalised cumulative shrinkage process (CUSP) representation (Frühwirth-Schnatter, 2023). CUSP was introduced by Legramanti et al. (2020) in the context of sparse infinite factor models and provides a mechanism for automatic inference on the effective latent dimensionality. In the context of our formulation, it is obtained by permuting the column index $h$ of the parameters $\theta_1, \ldots, \theta_H$ according to the decreasing slab probabilities $\tau_{(1)} > \ldots > \tau_{(H)}$. Increasing spike probabilities $\pi_h$ for $h = 1, \ldots, H$, as for a CUSP prior, are obtained by defining $\pi_h = 1 - \tau_{(h)}$. Representation (3.2) allows one to choose an upper limit for the number of factors, $H$, and keep it fixed throughout the entire inference procedure. By performing the classification between spike and slab independently for each column, we end up with defining an effective number of active factors $H_+$, which is random a priori and a posteriori, and typically smaller than $H$.

The key properties of the ESP prior do not depend on the specific choice of the spike and slab distributions. For computational convenience and to preserve conjugacy, we assume a hierarchical structure with inverse Gamma distributions for both the spike and the slab, and Gamma hyperpriors on their rate parameters. Finally, since the concentration parameter $\alpha$ plays a central role in determining the allocation between spike and slab, we allow it to be learned from the data by assigning it a Gamma prior $\alpha \sim \mathcal{G}(a_\alpha, b_\alpha)$.

Combining the variance decomposition introduced in Section 3.1 with the ESP prior yields the following hierarchical prior for the common factor loadings:

$$\begin{aligned}
&\lambda_{ih} \,|\, \theta_h, \phi_{ih}, \kappa_\Lambda \sim \mathcal{N}(0, \theta_h \phi_{ih} \kappa_\Lambda), \quad \text{where } i = 1, \ldots, p, \;\; h = 1, \ldots, H \\
&\theta_h \,|\, \tau_h \sim \tau_h \mathcal{G}^{-1}(a_\Lambda, b_\Lambda) + (1-\tau_h)\mathcal{G}^{-1}(a_0^\Lambda, b_0^\Lambda), \qquad \tau_h \sim \mathcal{B}\left(\frac{\alpha_\Lambda}{H}, 1\right), \\
&\alpha_\Lambda \sim \mathcal{G}(a_\alpha^\Lambda, b_\alpha^\Lambda), \quad b_\Lambda \sim \mathcal{G}(a_1^\Lambda, b_1^\Lambda), \quad b_0^\Lambda \sim \mathcal{G}(a_2^\Lambda, b_2^\Lambda), \\
&\phi_{ih} \sim \mathcal{G}^{-1}(a_\phi, \beta_{i\phi}), \qquad \beta_{i\phi} \sim \mathcal{G}(c, d). \\
&\kappa_\Lambda \sim \mathcal{G}^{-1}(a^\kappa, b^\kappa),
\end{aligned} \tag{3.4}$$

An analogous prior is assigned to the common factor scores:

$$\begin{aligned}
&f_{hns} \,|\, \upsilon_h, \zeta_{hns}, \kappa_F \sim \mathcal{N}(0, \upsilon_h \zeta_{hns} \kappa_F), \quad \text{where } n = 1, \ldots, N_s, \;\; s = 1, \ldots, S, \;\; h = 1, \ldots, H \\
&\upsilon_h \,|\, \pi_h \sim \pi_h \mathcal{G}^{-1}(a_F, b_F) + (1-\pi_h)\mathcal{G}^{-1}(a_0^F, b_0^F), \qquad \pi_h \sim \mathcal{B}\left(\frac{\alpha_F}{H}, 1\right),
\end{aligned}$$

$$
\begin{aligned}
&\alpha_F \sim \mathcal{G}(a_\alpha^F, b_\alpha^F), \quad b_F \sim \mathcal{G}(a_1^F, b_1^F), \quad b_0^F \sim \mathcal{G}(a_2^F, b_2^F), \\
&\zeta_{hns} \sim \mathcal{G}^{-1}(a_\zeta, \beta_{\zeta ns}), \qquad \beta_{\zeta ns} \sim \mathcal{G}(e, w), \\
&\kappa_F \sim \mathcal{G}^{-1}(c^\kappa, d^\kappa),
\end{aligned} \tag{3.5}
$$

### 3.3 Prior specification for the study-specific signal

Since our primary interest lies in identifying biclusters in the common signal, we do not impose sparsity on the study-specific factor scores and instead assign them a standard Gaussian prior $\boldsymbol{\eta}_{ns} \sim N(\mathbf{0}, \boldsymbol{I}_{H_s})$, which is a standard choice in BFA. For the study-specific factor loadings, we again employ the ESP prior described in Section 3.2, but in a simplified form that omits the global sparsity component. This choice reflects a trade-off between sparsity and flexibility: although some degree of shrinkage is desirable, the study-specific loadings must remain sufficiently dense to capture technology-driven and batch-related variation. At the same time, automatic determination of the effective study-specific dimensionality remains important, motivating the use of factor-level sparsity through the ESP prior.

The prior on the study-specific factor loadings is given by:

$$
\begin{aligned}
&\lambda_{ihs} \,|\, \theta_{hs}, \phi_{ihs} \sim \mathcal{N}(0, \theta_{hs}\phi_{ihs}), \quad \text{where } i = 1, \ldots, p,\ h = 1, \ldots, H_s,\ s = 1, \ldots, S, \\
&\theta_{hs} \,|\, \tau_{hs} \sim \tau_{hs}\mathcal{G}^{-1}(a_\Lambda^S, b_\Lambda^S) + (1 - \tau_{hs})\mathcal{G}^{-1}(a_0^{S\Lambda}, b_0^{S\Lambda}), \qquad \tau_{hs} \sim \mathcal{B}\left(\frac{\alpha_\Lambda^S}{H_s}, 1\right), \\
&\alpha_\Lambda^S \sim \mathcal{G}(a_\alpha^{S\Lambda}, b_\alpha^{S\Lambda}), \quad b_0^{S\Lambda} \sim \mathcal{G}(a_2^{S\Lambda}, b_2^{S\Lambda}), \quad b_\Lambda^S \sim \mathcal{G}(a_1^{S\Lambda}, b_1^{S\Lambda}), \\
&\phi_{ihs} \sim \mathcal{G}^{-1}(a_\phi^S, \beta_{\phi is}^S), \qquad \beta_{\phi is}^S \sim \mathcal{G}(c^S, d^S).
\end{aligned} \tag{3.6}
$$

Note that the column-specific variance parameters $\theta_{hs}$, slab probabilities $\tau_{hs}$ and binary indicators $\gamma_{hs}$ are study-specific, whereas the corresponding hyperparameters are shared across studies. This allows information to be borrowed across studies when determining the dimensionality of the study-specific latent structure, while still permitting each study to have its own set of active factors. Similar to the common factors, the number of active study-specific factors (assigned to the slab) is denoted by $H_s^+$.

The performance of the proposed model depends critically on the specification of the spike-and-slab components. In particular, the spike distribution must exert substantially stronger shrinkage than the slab distribution in order to achieve effective separation between active and inactive coefficients. If the two components are insufficiently separated, factor selection becomes unstable and inactive loadings may not be adequately shrunk towards zero. General intuition and practical guidance on calibrating the spike and slab hyperparameters in the ESP framework are provided by Grushanina and Frühwirth-Schnatter (2025).

The relative amount of shrinkage imposed on the common and study-specific latent components also requires careful calibration. Since the primary objective of the model is to recover the shared latent structure, the common component is assigned less aggressive factor-level shrinkage. In contrast, stronger shrinkage is imposed on the study-specific factor-level sparsity parameters, resulting in more conservative activation of study-specific factors. This reduces the tendency of study-specific factors to absorb variation that is shared across studies and helps mitigate the information-switching phenomenon discussed in Section 4. The final hyperparameter values were selected empirically based on extensive simulation experiments and are reported in Section B.2 of the Supplementary Material.

### 3.4 Priors for other parameters

The regression coefficients are assigned independent Gaussian priors $\beta_{il}^X \sim N(0, q_{il})$, where $i = 1, \ldots, p$ and $l = 1, \ldots, p_X$ denote the variable and covariate indices, respectively.

Finally, we assume a diagonal covariance matrix for the idiosyncratic errors within each study:

$$\xi_{is}^2 \sim \mathcal{G}^{-1}(a_\xi, b_{\xi is}), \qquad b_{\xi is} \sim \mathcal{G}(a_g, b_{gis}).$$

The rate hyperparameters $b_{gis}$ are assigned the data-driven values $100/R_{is}^2$, where $R_{is}$ is the range of the data in dimension $i$ and study $s$ (see, for example, Frühwirth-Schnatter (2006) for the reasoning behind this choice).

## 4 Identification

In factor analysis, identifiability refers to the ability to uniquely recover the latent factors and their loadings from the observed data. Classical factor models are not identifiable because different rotations of the latent factor space yield observationally equivalent representations with identical covariance structure (Anderson and Rubin (1956), Lopes and West (2004), Frühwirth-Schnatter et al. (2024), among many others). In multi-study factor models, this issue is further complicated by information switching, whereby signal can be equivalently represented by either shared (common) or study-specific latent factors (De Vito et al., 2019; Chandra et al., 2025; Bortolato and Canale, 2024). In this section, we discuss these sources of non-identifiability and their implications for the MSFRB model.

### 4.1 Rotational indeterminacy

The rotational invariance problem inherent in factor models arises from the fact that the decomposition of the covariance structure is invariant under orthogonal transformations of the latent space. In particular, the decomposition $\mathbf{\Omega} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Lambda}_s\mathbf{\Lambda}_s^\top + \mathbf{\Xi}$ will also hold for any semi-orthogonal matrices $\mathbf{P}$ ($\mathbf{P}\mathbf{P}^\top = \mathbf{I}$), $\mathbf{P}_s$ ($\mathbf{P}_s\mathbf{P}_s^\top = \mathbf{I}$), and $\mathbf{\Theta} = \mathbf{\Lambda}\mathbf{P}$, $\mathbf{g}_{ns} = \mathbf{P}^\top\mathbf{f}_{ns}$, $\mathbf{\Theta}_s = \mathbf{\Lambda}_s\mathbf{P}_s$, $\mathbf{v}_{ns} = \mathbf{P}_s^\top\boldsymbol{\eta}_{ns}$, in which case the two models

$$\mathbf{y}_{ns} = \mathbf{\Lambda}\mathbf{f}_{ns} + \mathbf{\Lambda}_s\boldsymbol{\eta}_{ns} + \boldsymbol{\epsilon}_{ns} \quad \text{and} \quad \mathbf{y}_{ns} = \mathbf{\Theta}\mathbf{g}_{ns} + \mathbf{\Theta}_s\mathbf{v}_{ns} + \boldsymbol{\epsilon}_{ns}$$

are observationally equivalent. This problem can either be addressed *a priori* via imposing restrictions on the elements of $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_s$ (see, for example, Lopes and West (2004) and Carvalho et al. (2008)), or *a posteriori* by using some type of orthogonalisation procedure, such as Procrustes or Varimax rotation (Kaiser, 1958; Poworoznek et al., 2021; McParland et al., 2014).

Let us first consider the product of the common factor loadings and factor scores, $\mathbf{\Lambda}\mathbf{F}$, where biclustering occurs. Both matrices are endowed with sparse structure through the sparse ESP prior, which imposes sparsity via coordinate-specific column-selection and element-wise shrinkage parameters for both factor loadings and scores. As argued in Moran et al. (2021), generic rotations typically mix coordinates, spreading signal across multiple columns and entries, and thereby destroying the axis-aligned sparse structure favoured by the prior. Consequently, while the likelihood remains rotationally invariant, the posterior is not. Rotational indeterminacy is therefore substantially reduced, leaving only trivial ambiguities such as sign switches and permutations. A more detailed discussion of this phenomenon can be found in Moran et al. (2021) and Rohe and Zeng (2023). We therefore rely

on the sparsity-inducing priors to obtain effectively identifiable solutions. In Section 4.4, we further introduce a sufficient condition based on unique support that can be used to assess factor identification in practice.

It should be noted that this argument pertains only to the common factors and their loadings. Study-specific factors are expected to be less sparse, as they might capture effects inherent to an individual study, and are therefore more susceptible to rotational ambiguity. However, since the primary focus of this work is on the common factors, addressing the rotational indeterminacy of study-specific components lies beyond the scope of this paper.

## 4.2 Scaling

While sparsity-inducing priors substantially reduce rotational indeterminacy, another source of non-identifiability remains. When the product $\mathbf{\Lambda F}$ is estimated, the columns of the factor loadings and scores are estimated only up to scaling. This means that $\boldsymbol{\lambda}_h^\top \mathbf{f}_h$ is equivalent to $(\boldsymbol{\lambda}_h c_h^{-1})^\top c_h \mathbf{f}_h$ for any constant $c_h \in \mathbb{R}$. In biclustering, the primary objective is to identify the nonzero elements of the matrices $\mathbf{\Lambda}$ and $\mathbf{F}$, as these determine the subsets of features and samples that co-vary, rather than to estimate their exact magnitudes. Since the scale of these values is not of direct interest, we follow Moran et al. (2021) and rescale $\mathbf{\Lambda}$ and $\mathbf{F}$ at each iteration of the VI algorithm so that the corresponding columns have equal norms. More specifically,

$$c_h = \sqrt{\frac{\|\boldsymbol{\lambda}_h\|_2}{\|\boldsymbol{f}_h\|_2}}, \qquad \boldsymbol{\lambda}_h \leftarrow \frac{1}{c_h}\boldsymbol{\lambda}_h, \qquad \boldsymbol{f}_h \leftarrow c_h \boldsymbol{f}_h.$$

Rescaling also improves numerical stability and leads to faster convergence by eliminating drift along scale-equivalent representations and improving the conditioning of the variational updates.

## 4.3 Information switching

Information switching arises because variation that is shared across studies can often be represented either through common factors or through a combination of study-specific factors, leading to multiple observationally equivalent decompositions. In the existing literature on MSFA models, information switching has been addressed differently, depending on the nature and purpose of the model. In the frequentist MSFA framework, De Vito et al. (2019) address information switching by identifying common and study-specific factors with distinct linear subspaces. Thus, the common factors are defined as the intersection of the study-specific covariance subspaces, while the study-specific factors are defined as their complements. This requires a condition that the concatenated loading matrix $\mathbf{L} = [\mathbf{\Lambda}, \mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_S]$ has a full column rank $r(\mathbf{L}) = H_+ + \sum_{s=1}^S H_s^+$, with $H_+ + \sum_{s=1}^S H_s^+ \leq p$, which can be rather restrictive in some settings. In the Bayesian framework, De Vito et al. (2021) do not explicitly impose any constraints to prevent information switching between common and study-specific factors. Instead, they rely on sparsity-inducing shrinkage priors to encourage, though not guarantee, their practical separation. As they demonstrate in the discussion section, this can fail in the case of only partially shared common factors.

In subsequent years, several approaches have been proposed in the literature to address information switching. Roy et al. (2021) address this issue by replacing the additive MSFA decomposition with a perturbation-based model, in which a single common factor structure is assumed and group differences are captured through multiplicative perturbation matrices, thereby avoiding information

switching inherent in the additive structure. Unlike Roy et al. (2021), Grabski et al. (2023) do not depart from the additive MSFA formulation, but generalise the common-versus-study-specific decomposition by introducing a latent indicator matrix that allows factors to be shared across arbitrary subsets of studies. While their primary objective is not to address information switching, the resulting framework moves beyond the strict common-versus-study-specific dichotomy and treats the factor-sharing structure itself as an inferential target. Chandra et al. (2025) address information switching by constraining study-specific factors to lie within the subspace spanned by the common factors, effectively modelling them as structured perturbations of the shared component and thus ensuring a unique decomposition. While this prevents information switching by model design, it can be too restrictive in case of the data showing substantial heterogeneity across studies. Finally, Bortolato and Canale (2024) introduce a flexible approach in which the model learns factor activation patterns from the data at the observation level. This produces a matrix of $0/1$ activation indicators for individual observations, allowing factors to be shared across arbitrary subsets of observations and studies. Because these activation patterns are identifiable under mild conditions, they provide an anchoring mechanism for the latent factors and thereby prevent the factor-allocation ambiguities associated with information switching.

While Bortolato and Canale (2024) achieve identifiability through sparse observation-level activation patterns, our setting is fundamentally different because factors are represented by biclusters, with sparsity induced simultaneously in the factor loading and score matrices. Consequently, the relevant support structure is no longer a one-dimensional activation pattern across samples, but a two-dimensional feature-sample support. This motivates an identification condition based on bicluster support rather than sample-level activation patterns.

### 4.4 Identification based on unique support

Motivated by the idea in Bortolato and Canale (2024) of using factor activation patterns for identification, we propose a sufficient identification condition tailored to the biclustering structure of the MSFRB model. In particular, the combination of element-wise sparsity in both the common factor loadings and scores, together with sparse loadings in the study-specific components, implies that non-identifiability due to information switching is highly unlikely in sufficiently sparse settings, as overlapping representations of the same signal would require substantial coincidence of supports across multiple components.

Building on this idea, we propose a post hoc identification strategy that exploits this double sparsity by recovering factor supports from posterior summaries and assessing their uniqueness. Unlike APAFA, where identification is based on observation-level activation patterns, identification in the MSFRB model cannot generally be determined from the sample support alone. A common factor contributes to the observed data through the product of its loadings and scores, so both the features and the samples on which the factor is active determine its support. Consequently, the natural object for identification is the joint feature-sample support of the factor, rather than either support considered separately. More specifically, for each active factor, we construct its support over features and samples and define its unique support as the subset of entries not shared with any other factor. A factor is deemed identified if it has at least one element in its support that is unique. This is a sufficient condition to rule out information switching and provides a simple and interpretable diagnostic for assessing identification in practice. Below, we formulate this idea more formally.

**Proposition 1.** *Consider the common component of the MSFRB model:*

$$\mathbf{Y}^{(c)} = \sum_{h=1}^{H_+} \boldsymbol{\lambda}_h \boldsymbol{f}_h^\top,$$

*where $\boldsymbol{\lambda}_h \in \mathbb{R}^p$ is the $h$th column of the common loading matrix and $\boldsymbol{f}_h \in \mathbb{R}^N$ is the corresponding common factor score vector. Here $N = \sum_{s=1}^{S} N_s$ denotes the total number of observations across all studies. For each active common factor $h$, define its feature support and sample support as*

$$G_h = \{i \in \{1, \dots, p\} : \lambda_{ih} \neq 0\}$$
$$T_h = \{n \in \{1, \dots, N\} : f_{hn} \neq 0\},$$

*and let its total support be*

$$B_h = G_h \times T_h \subseteq \{1, \dots, p\} \times \{1, \dots, N\}.$$

*Similarly, for each active study-specific factor $r$ in study $s$, let*

$$G_r^s = \{i \in \{1, \dots, p\} : \lambda_{ir}^s \neq 0\}, \qquad T_r^s = \mathcal{I}_s$$

*where $\mathcal{I}_s$ is by the model design the set of all samples belonging to study $s$. Define the total study-specific factor support as*

$$B_r^s = G_r^s \times \mathcal{I}_s.$$

*For common factor $h$, compute its unique support as*

$$U_h = B_h \setminus \left( \bigcup_{l \neq h} B_l \cup \bigcup_{s=1}^{S} \bigcup_{r=1}^{H_s^+} B_r^s \right).$$

*Then if $U_h \neq \varnothing$, the $h$th common factor is identified with respect to information switching, in the sense that its contribution cannot be reproduced or absorbed by the remaining common factors or by any study-specific factor.*

*Proof.* Take $(i, n) \in U_h$. By construction, $(i, n) \in B_h$, therefore both $\lambda_{ih} \neq 0$ and $f_{hn} \neq 0$, so factor $h$ contributes nontrivially to the data entry $(i, n)$. Since $(i, n) \in U_h$, we have $(i, n) \notin B_l$ for all $l \neq h$ and $(i, n) \notin B_r^s$ for all study-specific factors $r$ and studies $s$. Therefore no other common or study-specific factor is active at $(i, n)$. It follows that the signal at entry $(i, n)$ is generated exclusively by factor $h$. Consequently, the contribution of factor $h$ at $(i, n)$ cannot be reproduced or absorbed by any other common or study-specific factor, since none of these components has support at $(i, n)$. Therefore no observationally equivalent decomposition can reallocate the contribution of factor $h$ to the remaining factors. Hence factor $h$ is identified with respect to information switching. □

**Remark 1.** *The uniqueness condition also explains why nontrivial rotations are incompatible with the recovered sparse support structure. Algebraically, the common component remains invariant under rotations of the latent factors. However, a nontrivial rotation mixes the original factor loadings and scores, thereby spreading signal across multiple factors. If $(i, n)$ belongs to the unique support of factor $h$, this entry is generated exclusively by factor $h$ in the sparse representation. A rotation*

*that mixes factor $h$ with other factors would generally make additional rotated factors active at this entry, destroying the uniqueness of the support. Thus, unique support acts as an anchor for a particular sparse representation of the common component, leaving only transformations that preserve the support structure, such as sign changes and column permutations.*

The details of the post hoc identification algorithm are described in the Supplementary Material, Section C.1.

# 5 Variational inference for multi-study factor regression biclustering

Let $\boldsymbol{\varphi}$ denote the collection of all unknown model parameters and latent variables. The goal of variational inference is to approximate the posterior distribution $p(\boldsymbol{\varphi}|\boldsymbol{Y})$ with some density $q^*(\boldsymbol{\varphi})$ in a class $\mathcal{Q}$. The optimal variational density $q^*(\boldsymbol{\varphi})$ is defined as the closest to the posterior $p(\boldsymbol{\varphi}|\boldsymbol{Y})$ in Kullback-Leibler (KL) divergence, i.e. which satisfies the following expression:

$$\underset{q(\boldsymbol{\varphi})\in\mathcal{Q}}{\arg\min}\left\{KL(q(\boldsymbol{\varphi}) \parallel p(\boldsymbol{\varphi}|\boldsymbol{Y}))\right\} = \underset{q(\boldsymbol{\varphi})\in\mathcal{Q}}{\arg\min}\left\{\int_{\boldsymbol{\Phi}} q(\boldsymbol{\varphi}) log\frac{q(\boldsymbol{\varphi})}{p(\boldsymbol{\varphi}|\boldsymbol{Y})}d\boldsymbol{\varphi}\right\}$$

This is equivalent to maximization of the evidence lower bound (ELBO):

$$\underset{q(\boldsymbol{\varphi})\in\mathcal{Q}}{\arg\max}\left\{ELBO(q(\boldsymbol{\varphi}))\right\} = \underset{q(\boldsymbol{\varphi})\in\mathcal{Q}}{\arg\max}\left\{\mathbb{E}_{q(\boldsymbol{\varphi})}\left[\log p(\boldsymbol{\varphi},\boldsymbol{Y})\right] - \mathbb{E}_{q(\boldsymbol{\varphi})}\left[\log q(\boldsymbol{\varphi})\right]\right\}, \tag{5.1}$$

where all the expectations are taken with respect to $q(\boldsymbol{\varphi})$ and $p(\boldsymbol{\varphi},\boldsymbol{Y})$ is the joint distribution of the data and model parameters, that factorises into the composition of the likelihood and prior(s) $p(\boldsymbol{\varphi},\boldsymbol{Y}) = p(\boldsymbol{Y}|\boldsymbol{\varphi})p(\boldsymbol{\varphi})$.

## 5.1 Mean-field variational inference

A common approach to select $\mathcal{Q}$ is to use a mean-field variational family that assumes a partition of parameters into independent blocks and approximates the posterior with a product of independent densities $\mathcal{Q}^{MF} = \{q : q(\boldsymbol{\varphi}) = \prod_{m=1}^{M} q(\boldsymbol{\varphi}_m)\}$. The independent densities $q(\boldsymbol{\varphi}_m)$ are referred to as variational factors. The optimal variational factors $q^*(\boldsymbol{\varphi}_m)$ can be obtained from:

$$q_m^*(\boldsymbol{\varphi}_m) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{\varphi}_{-m})}[\log p(\boldsymbol{\varphi}_m|\boldsymbol{\varphi}_{-m},\boldsymbol{Y})]\right), \tag{5.2}$$

where $\boldsymbol{\varphi}_{-m} = (\boldsymbol{\varphi}_1,\ldots,\boldsymbol{\varphi}_{m-1},\boldsymbol{\varphi}_{m+1},\ldots,\boldsymbol{\varphi}_M)$ is the vector of parameters excluding the $m$th one (Bishop (2006), chapter 10).

When the conditional distribution $p(\boldsymbol{\varphi}_m|\boldsymbol{\varphi}_{-m},\boldsymbol{Y})$ is of the exponential family, the corresponding optimal variational factors also belong to the same exponential family. Thus, the maximisation problem in equation (5.1) is reduced to learning the optimal variational factors $q_m^*(\boldsymbol{\varphi}_m)$. Choosing $\mathcal{Q}^{MF}$ and the prior in such a way that ensures conjugacy enables the development of a coordinate-ascent variational inference (CAVI) algorithm, which learns the optimal variational factors by iteratively updating them according to equation (5.2) until convergence is reached (Bishop (2006), chapter 10).

To implement CAVI for MSFRB as described above, we assume a mean field variational approximation where variational distributions of our parameters are independent:

$$q(\boldsymbol{\varphi}) = q(\boldsymbol{F},\boldsymbol{\Lambda},\boldsymbol{\eta},\boldsymbol{\Lambda}_s,\boldsymbol{\beta}^X,\boldsymbol{\Xi}_s,\boldsymbol{b}_g,\boldsymbol{\Theta},\boldsymbol{\tau},\boldsymbol{r},\boldsymbol{\Theta}_s,\boldsymbol{\tau}_s,\boldsymbol{r}_s,\boldsymbol{\Upsilon},\boldsymbol{\pi},\boldsymbol{z},\boldsymbol{\Phi},\boldsymbol{b}_\phi,\boldsymbol{\Phi}_s,\boldsymbol{b}_{\phi s},\boldsymbol{\zeta},\boldsymbol{b}_\zeta,$$

$$
\begin{aligned}
&\alpha_\Lambda, \alpha_\Lambda^S, \alpha_F, b_\Lambda, b_0^\Lambda, b_F, b_0^F, b_\Lambda^S, b_0^{S\Lambda}, \kappa_\Lambda, \kappa_F) = \\
&= q(\alpha_\Lambda) q(\alpha_\Lambda^S) q(\alpha_F) q(b_\Lambda) q(b_0^\Lambda) q(b_\Lambda^S) q(b_0^{S\Lambda}) q(b_F) q(b_0^F) q(\kappa_\Lambda) q(\kappa_F) \times \\
&\left[\prod_{i=1}^{p} q(\boldsymbol{\beta}_i^X) q_H(\boldsymbol{\lambda}_i) q(b_{i\phi}) \prod_{h=1}^{H} q(\phi_{ih})\right] \times \\
&\left[\prod_{i=1}^{p}\prod_{s=1}^{S} q(b_{gis}) q(\xi_{is}^2) q_{H_s}(\boldsymbol{\lambda}_{is}) q(b_{\phi is}) \prod_{h=1}^{H} q(\phi_{ihs})\right] \times \\
&\left[\prod_{h=1}^{H} q(r_h) q(\tau_h) q(\theta_h) q(z_h) q(\pi_h) q(\upsilon_h)\right] \times \\
&\prod_{s=1}^{S}\left[\left[\prod_{n=1}^{H_s} q(r_{hs}) q(\tau_{hs}) q(\theta_{hs})\right]\left[\prod_{n=1}^{N_s} q_{H_s}(\boldsymbol{\eta}_{ns}) q_H(\boldsymbol{f}_{ns}) q(\beta_{\zeta ns}) \prod_{h=1}^{H} q(\zeta_{hns})\right]\right]
\end{aligned}
$$

## 5.2 Variational inference for the ESP prior

While for many of the model's parameters mean field variational inference is rather standard, this is not the case for the parameters involved in the ESP prior. According to our knowledge, variational inference for the ESP prior has not been described before and involves certain particularities. In this section, we describe variational inference for the general ESP prior on latent factor loadings in a basic factor model as in equation (2.1). This can be easily extended to the sparse ESP priors in the MSFRB model as described in Section 3. The variational update steps for the sparse ESP priors on the common factor loadings and scores, as well as on the study-specific factor loadings, are presented in detail in Algorithm A.1 in the Supplementary Material.

Using auxiliary indicator $\gamma_h$, for a basic factor model as in equation (2.1), the ESP prior on the column variance parameter of factor loadings can be formulated as follows:

$$
\begin{aligned}
&\lambda_{ih}|\theta_h \sim \mathcal{N}(0, \theta_h), \quad i = 1, \ldots, p, \quad h = 1, \ldots, H, \\
&\theta_h|\gamma_h = 1 \sim \mathcal{G}^{-1}(a_\Lambda, b_\Lambda), \quad \theta_h|\gamma_h = 0 \sim \mathcal{G}^{-1}(a_0, b_0), \\
&\gamma_h \sim Ber(\tau_h), \quad \tau_h \sim \mathcal{B}(\frac{\alpha}{H}, 1).
\end{aligned} \tag{5.3}
$$

For simplicity, we assume that $a_\Lambda$, $b_\Lambda$, $a_0$, $b_0$, and $\alpha$ are fixed hyperparameters. Extending the algorithm to infer these quantities is straightforward through standard conjugate variational updates.

Then the mean-field variational family takes the form:

$$
q(-) = \prod_{h=1}^{H}\left[q(\gamma_h) q(\tau_h) q(\theta_h|\gamma_h) \prod_{i=1}^{p} q(\lambda_{ih})\right] \tag{5.4}
$$

We choose the following variational distribution for conjugacy: $q(\gamma_h) = Ber(r_h)$, $q(\tau_h) = \mathcal{B}(\hat{a}_{\tau_h}, \hat{b}_{\tau_h})$, $q(\theta_h|\gamma_h = 1) = \mathcal{G}^{-1}(\hat{a}_{\theta_h slab}, \hat{b}_{\theta_h slab})$, $q(\theta_h|\gamma_h = 0) = \mathcal{G}^{-1}(\hat{a}_{\theta_h spike}, \hat{b}_{\theta_h spike})$, $q(\boldsymbol{\lambda}_i) = \mathcal{N}_H(\boldsymbol{\mu}_i^\lambda, \boldsymbol{\Sigma}_i^\lambda)$. The algorithm below illustrates variational updates for the respective parameters.

1. For $h = 1, \ldots, H$ update $\theta_h|\gamma_h$ from $\mathcal{G}^{-1}(\hat{a}_{\theta_h}, \hat{b}_{\theta_h})$ using

   for $\gamma_h = 1$

$$\hat{a}_{\theta_h slab} = a_\Lambda + \frac{p}{2}$$
$$\hat{b}_{\theta_h slab} = b_\Lambda + \frac{1}{2}\sum_{i=1}^{p}\mathbb{E}(\lambda_{ih}^2)$$

for $\gamma_h = 0$

$$\hat{a}_{\theta_h slab} = a_0 + \frac{p}{2}$$
$$\hat{b}_{\theta_h slab} = b_0 + \frac{1}{2}\sum_{i=1}^{p}\mathbb{E}(\lambda_{ih}^2)$$

where $\mathbb{E}\lambda_{ih}^2 = \left[\mu_{ih}^\lambda\right]^2 + \mathbf{\Sigma}_{i[h,h]}^\lambda$ with $h$ and $[h,h]$ indicating the $h$th element of the mean vector $\boldsymbol{\mu}_i^\lambda$ and the $[h,h]$th element of the covariance matrix $\mathbf{\Sigma}_i^\Lambda$, correspondingly.

2. For $h = 1, \ldots, H$ update the responsibilities $r_h$ in the following way:

$$s_{h\,slab} = \mathbb{E}\log\tau_h + a_\Lambda \log b_\Lambda - \log\Gamma(a_\Lambda) - (a_\Lambda + \frac{p}{2} + 1)\mathbb{E}_{slab}\log\theta_h - \mathbb{E}_{slab}\left(\frac{1}{\theta_h}\right)\left(b_\Lambda + \frac{1}{2}\sum_{i=1}^{p}\mathbb{E}(\lambda_{ih}^2)\right)$$

$$s_{h\,spike} = \mathbb{E}\log(1-\tau_h) + a_0 \log b_0 - \log\Gamma(a_0) - (a_0 + \frac{p}{2} + 1)\mathbb{E}_{spike}\log\theta_h - \mathbb{E}_{spike}\left(\frac{1}{\theta_h}\right)\left(b_0 + \frac{1}{2}\sum_{i=1}^{p}\mathbb{E}(\lambda_{ih}^2)\right)$$

where

$$\mathbb{E}\lambda_{ih}^2 = \left[\mu_{ih}^\lambda\right]^2 + \mathbf{\Sigma}_{i[h,h]}^\lambda,$$
$$\mathbb{E}\log\tau_h = \psi(\hat{a}_{\tau_h}) - \psi(\hat{a}_{\tau_h} + \hat{b}_{\tau_h}),$$
$$\mathbb{E}\log(1-\tau_h) = \psi(\hat{b}_{\tau_h}) - \psi(\hat{a}_{\tau_h} + \hat{b}_{\tau_h}),$$
$$\mathbb{E}(\tfrac{1}{\theta_h}) = \frac{\hat{a}_{\theta_h}}{\hat{b}_{\theta_h}},$$
$$\mathbb{E}\log\theta_h = \ln(\hat{b}_{\theta_h}) - \psi(\hat{a}_{\theta_h}),$$

where $\psi(\cdot)$ is a digamma function.

Calculate responsibility (probability that factor $h$ is active) as

$$r_h = \frac{\exp(s_{h\,slab})}{\exp(s_{h\,slab}) + \exp(s_{h\,spike})}$$

As probabilities of a factor being active or inactive can take very small or very large values, which can lead to "Inf" or "NaN" values in the algorithm, for the sake of numerical stability we recommend to use the log-sum-exp trick, which leads to the following representation of $r_h$:

$$r_h = \frac{1}{1 + \exp(s_{h\,spike} - s_{h\,slab})}.$$

3. For $h = 1, \ldots, H$ update $\tau_h$ from $\mathcal{B}(\hat{a}_{\tau_h}, \hat{b}_{\tau_h})$ using

$$\hat{a}_{\tau_h} = \frac{\alpha}{H} + r_h$$
$$\hat{b}_{\tau_h} = 1 + (1 - r_h) = 2 - r_h$$

4. For $i = 1, \ldots, p$ update $\boldsymbol{\lambda}_i^\top$ from $\mathcal{N}_H(\boldsymbol{\mu}_i^\lambda, \boldsymbol{\Sigma}_i^\lambda)$ using

$$\boldsymbol{\Sigma}_i^\lambda = \left( \boldsymbol{\Psi}^{-1} + \mathbb{E}\left(\frac{1}{\xi_i^2}\right) \sum_{n=1}^{N} \mathbb{E}(\boldsymbol{f}_n \boldsymbol{f}_n^\top) \right)^{-1}$$

where

$$\boldsymbol{\Psi}^{-1} = \operatorname{diag}\left( \mathbb{E}\left(\frac{1}{\theta_1}\right), \ldots, \mathbb{E}\left(\frac{1}{\theta_H}\right) \right),$$
$$\mathbb{E}(\tfrac{1}{\theta_h}) = r_h \tfrac{\hat{a}_{\theta_h slab}}{\hat{b}_{\theta_h slab}} + (1 - r_h) \tfrac{\hat{a}_{\theta_h spike}}{\hat{b}_{\theta_h spike}}, \; h = 1, \ldots, H,$$
$$\mathbb{E}(\boldsymbol{f}_n \boldsymbol{f}_n^\top) = \boldsymbol{\mu}_n^f \left[\boldsymbol{\mu}_n^f\right]^\top + \boldsymbol{\Sigma}_n^f,$$
$$\mathbb{E}\left(\frac{1}{\xi_i^2}\right) = \frac{\hat{a}_\xi}{\hat{b}_{\xi_i}},$$

$$\boldsymbol{\mu}_i^\lambda = \boldsymbol{\Sigma}_i^\lambda \mathbb{E}\left(\frac{1}{\xi_i^2}\right) \left(\boldsymbol{\mu}_1^f, \ldots, \boldsymbol{\mu}_N^f\right) \boldsymbol{y}_i^\top.$$

## 5.3 CAVI algorithm for MSFRB model

The sparse ESP prior used in the biclustering algorithm is more elaborate than in Section 5.2 as in addition to the column shrinkage parameter, the variances of $\lambda_{ih}$, $f_{hn}$ and $\lambda_{ihs}$ also include additional sparsity parameters. Moreover, the rate parameters of the spike and slab inverse gamma distributions, namely $b_0^F$ and $b_F$ in case of common factors, $b_0^\Lambda$ and $b_\Lambda$ in case of common factor loadings, and $b_0^{S\Lambda}$ and $b_\Lambda^S$ in case of study-specific factor loadings are assigned hyperpriors. Finally, the concentration parameters of the beta distribution of slab probabilities $\alpha_\lambda$, $\alpha_F$ and $\alpha_\lambda^S$ are also learned from data (see equations 3.4, 3.5) and 3.6 for the full prior formulations). However, for prior choices that ensure conjugacy, these changes are easy to implement with standard enhancements of the algorithm presented in the previous section. For the inference on the spike-and-slab part of the ESP prior on all three variables $\lambda_{ih}$, $f_{hn}$ and $\lambda_{ihs}$, we utilize the same data augmentation scheme as in equation 5.3: $\gamma_h = 1$ if the $h$th column of the common factor loading matrix $\boldsymbol{\Lambda}$ is active and $\gamma_h = 0$ otherwise; $\nu_h = 1$ if the $h$th row of the common factor score matrix $\boldsymbol{F}$ is active and $\nu_h = 0$ otherwise; and $\gamma_{hs} = 1$ if the $h$th column of the study $s$ factor loading matrix $\boldsymbol{\Lambda}_s$ is active and $\gamma_{hs} = 0$ otherwise. The full details of the CAVI algorithm for the MSFRB model are presented in Algorithm A.1 in the Supplementary Material.

## 5.4 Stochastic variational inference

Molecular profiling data can occasionally be high dimensional not only in $p$ but also in $N_s$, in which case the CAVI algorithm might be relatively slow. To overcome this bottleneck, we employ stochastic variational inference (SVI) method (Hoffman et al., 2013), which deals with high dimensionality by updating local parameters (those parameters which updates scale with $N_s$ e.g. subject-specific or common factor scores for $n = 1, \ldots, N_s$) for only a subset of the $N_s$ observations and then uses this subset to approximate the remaining parameters common for all individuals $n = 1, \ldots, N_s$ and $s = 1, \ldots, S$. This is done in the following way.

First, a batch proportion parameter $b \in (0, 1)$ needs to be selected, then a random sample of data in the size of $\tilde{N}_s = b \times N_s$ is drawn at each iteration of the algorithm. So, at each iteration a $p \times \tilde{N}_s$ subset of $\boldsymbol{y}_{ns}$ will be taken, with randomly chosen $\tilde{N}_s$ out of $N_s$ columns. SVI will proceed updating the local parameters $\boldsymbol{\eta}_{ns}$, $\boldsymbol{f}_{ns}$, $\zeta_{hns}$ and $\beta_{ns\zeta}$ from this subset.

At the next step, the local parameters updated from this batch are used for updating the global parameters ($\boldsymbol{\beta}_i^X$, $\xi_{is}^2$, $\boldsymbol{\lambda}_i$, $\boldsymbol{\lambda}_{is}$, $\upsilon_h$, $z_h$ and $\kappa_F$). For this purpose, to ensure the stable performance and convergence, the algorithm is reformulated as a gradient-based optimization algorithm for the global parameters. Thus, the derivative of the ELBO with respect to natural parameters takes the following form:

$$\nabla ELBO(q(\boldsymbol{\varphi}_g)) = \mathbb{E}_{q(\boldsymbol{\varphi}_{-g})}\left[\eta(\boldsymbol{Y}, \boldsymbol{\varphi})\right] - \boldsymbol{\vartheta}_g, \tag{5.5}$$

where $\boldsymbol{\varphi}$ denotes the collection of model parameters and latent variables, and $\boldsymbol{\vartheta}_g$ denotes the variational parameters associated with the $g$th variational factor. Setting the gradient in equation (5.5) to 0, we get $\boldsymbol{\vartheta}_g = \mathbb{E}_{q(\boldsymbol{\varphi}_{-g})}\left[\eta(\boldsymbol{Y}, \boldsymbol{\varphi})\right]$.

The expectation of the likelihood is estimated with the weight of $N_s/\tilde{N}_s$ assigned to each individual observation. At each iteration of the algorithm the global parameters are estimated as a weighted average of the global parameters updated at the current and previous iterations:

$$\boldsymbol{\vartheta}_g(t) = (1 - \rho_t)\boldsymbol{\vartheta}_g(t-1) + \rho_t \hat{\boldsymbol{\vartheta}}_g(t),$$

where $\rho_t$ is a step size parameter such that $\sum_t \rho_t \to \infty$ and $\sum_t \rho_t^2 < \infty$ and $\boldsymbol{\vartheta}_g$ denotes the natural variational parameters associated with the $g$th global variational factor, with $g = 1, \ldots, G$ (in our case $G = 7$ and the set of global variational factors corresponds to the model parameters $\{\boldsymbol{\beta}_i^X$, $\xi_{is}^2$, $\boldsymbol{\lambda}_i$, $\boldsymbol{\lambda}_{is}$, $\upsilon_h$, $z_h$,$\kappa_F\}$). The choice of $\rho_t$ is rather influential as it can impact the convergence of the algorithm. A typical choice, suggested in Hoffman et al. (2013) is to set $\rho_t = (t + \tau^\rho)^{-\kappa^\rho}$, where $\kappa^\rho \in (0.5, 1]$ is the forgetting rate that controls how quickly the variational parameters change across iterations, and $\tau^\rho > 0$ is the delay, which downweights early iterations. The details of the SVI algorithm for the MSFRB model are presented in Section A.2 in the Supplementary Material.

# 6 Simulation Studies

To assess the performance of the proposed MSFRB model, we conducted two simulation studies. The first study is designed to provide an illustrative example of the model behaviour. The common factor loadings and factor scores are generated to produce visible block patterns both in the latent structure and in the observed data matrix. We consider both sparse and dense common factors and assess the ability of the model to recover the underlying parameters and correctly distinguish between these two different types of factors. The second simulation study investigates the performance of the MSFRB model across a range of various dimensional settings. In particular, we consider the following combinations of the number of features and sample sizes: $p \in \{100, 500, 1000\}$ and $N_s \in \{100, 300, 500\}$, while fixing the number of studies at $S = 4$. To assess the effect of the number of studies separately, we additionally perform a sensitivity analysis with $S \in \{4, 8\}$ for a fixed moderate-dimensional setting.

We compare the performance of the proposed MSFRB model with four competing methods commonly used for the analysis of molecular profiling data. The first comparator is the Bayesian multi-study factor analysis (MSFA) model of De Vito et al. (2021); Hansen et al. (2025). Like MSFRB,

this model explicitly accounts for the multi-study structure of the data by decomposing variation into common and study-specific latent factors. However, it does not perform biclustering, as sparsity is imposed only on the factor loadings, whereas the latent factors are assigned standard Gaussian priors with unit variance. Consequently, the model is able to identify sparse groups of variables associated with a factor, but not sparse groups of observations. We fit the model using the variational inference implementation provided by Hansen et al. (2025)[1].

The remaining three competing methods provide biclustering of factor loadings and scores, but do not explicitly account for the multi-study structure of the data. The first is empirical Bayes matrix factorisation (EBMF) (Wang and Stephens, 2021), which is fitted using the R package `flashier` and employs an efficient variational inference algorithm. The second is the Bayesian biclustering model of Gao et al. (2016), fitted using the R package `BicMix`, which combines variational inference with expectation-maximisation updates. Finally, we consider spike-and-slab lasso biclustering (SSLB) (Moran et al., 2021), fitted using the `SSLB` R package, which implements a fast expectation-maximisation algorithm with variational updates.

To compare the recovered matrices with their true (simulated) values, we use two complementary metrics. The first is the relative Frobenius distance, which quantifies the reconstruction error relative to the overall signal strength. It measures how far the estimated matrix is from the true matrix after accounting for the overall magnitude of the signal. The relative Frobenius distance between two matrices A and B is calculated as

$$\frac{\|A - B\|_F}{\|A\|_F},$$

where

$$\|A - B\|_F = \sqrt{\left(\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{ij} - B_{ij})^2\right)}.$$

Smaller values indicate better recovery, with a value of zero corresponding to perfect reconstruction.

The second metric is the RV coefficient (Abdi, 2007), which measures the similarity between two matrices in terms of their underlying structure. For two matrices $A$ and $B$, the RV coefficient is defined as

$$\mathrm{RV}(A, B) = \frac{\mathrm{tr}\left(AA^\top BB^\top\right)}{\sqrt{\mathrm{tr}\left(AA^\top AA^\top\right)\ \mathrm{tr}\left(BB^\top BB^\top\right)}}.$$

The RV coefficient takes values in $(0, 1)$, with values close to $0$ indicating low structural similarity and values close to $1$ indicating highly similar matrices.

The variational inference implementation of MSFA provided by Hansen et al. (2025) performs best when the data are centred and standardised prior to model fitting. Therefore, each study-specific data matrix is normalised before applying this method. To ensure comparability with the other approaches, the fitted values are transformed back to the original measurement scale using the corresponding study-specific means and standard deviations before computing reconstruction errors. All reported performance metrics are therefore evaluated on the original data scale. Unlike the biclustering methods considered here, MSFA imposes sparsity only on the factor loadings and therefore

[1]Other Bayesian multi-study factor analysis models suitable for molecular profiling data, such as Chandra et al. (2025) and Bortolato and Canale (2024), have also been proposed. However, these approaches are primarily aimed at covariance estimation and rely on MCMC-based inference, which can be prohibitively slow in the high-dimensional settings considered in this work.

remains susceptible to rotational indeterminacy. Consequently, direct comparison of the estimated factor loadings and factor scores with their simulated values is not meaningful. To avoid identifiability issues and ensure a fair comparison across methods, we evaluate recovery of the common latent structure through the product $\boldsymbol{\Lambda F}$, which is invariant to orthogonal rotations of the latent factors.

The remaining biclustering methods (EBMF, BicMix and SSLB) are applied directly to the pooled data matrix without additional preprocessing. Preliminary experiments based on residualising covariate and study effects prior to model fitting generally led to poorer recovery of the latent structure, suggesting that such preprocessing may remove part of the signal of interest. We therefore report results obtained from the original simulated data.

For the BicMix model of Gao et al. (2016), implemented in the `BicMix` R package, the estimated biclustering component $\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}}$ is reconstructed using the subset of latent factors that exhibit sparsity in both the loading and factor-score directions. BicMix provides posterior inclusion probabilities for the loading and score components of each factor. We classify a component as sparse if both the loading and score inclusion probabilities exceed a predefined threshold (we set it to $0.5$), and construct the estimated common component as the product of the corresponding loading and factor matrices. This ensures that only factors exhibiting simultaneous sparsity across features and samples contribute to the reconstructed biclustering structure.

Data reconstruction accuracy was evaluated on the original data scale. All reported reconstruction errors were computed by comparing the estimated and true data matrices expressed on this scale.

## 6.1 Simulation study 1: Block-structured common factors and loadings

### 6.1.1 Reconstruction of data Y

The goal of this simulation study is to demonstrate that the proposed model is able to recover complex sparse structure present in the data. For this, we create a data matrix with visible block structure in the following way. We generate data coming from $S = 3$ studies, with $p = 60$ features and $N_s = 100$ samples per study, thus making $N = 300$ total samples in the dataset. We consider three active common factors and two active study-specific factors in each study. Study-specific factors are generated as $\boldsymbol{\eta}_{ns}^{\top} \sim \mathcal{N}_2(\mathbf{0}, \mathbf{I}_2)$ for $n = 1, \ldots, N_s$ and $s = 1, \ldots, S$, while study-specific factors loadings are generated as $\lambda_{ihs} \sim \mathcal{N}(0, 0.35^2)$ for all $i$, $s$, and $h = 1, 2$ denoting the number of active factors.

To create clearly visible sparse block structure in the data matrix, the nonzero values of common factor loadings and factors are generated in the following way:

$$
\begin{aligned}
&\lambda_{i1} \sim \mathcal{N}(0, 4^2), \quad i = 1, \ldots, 24, \\
&\lambda_{i2} \sim \mathcal{N}(0, 4^2), \quad i = 21, \ldots, 42, \\
&\lambda_{i3} \sim \mathcal{N}(0, 4^2), \quad i = 39, \ldots, 60, \\
&f_{1n} \sim \mathcal{N}(0, 4^2), \quad n \in \{6, \ldots, 28\} \cup \{114, \ldots, 132\} \cup \{201, \ldots, 225\}, \\
&f_{2n} \sim \mathcal{N}(0, 4^2), \quad n \in \{37, \ldots, 63\} \cup \{131, \ldots, 161\} \cup \{245, \ldots, 263\}, \\
&f_{3n} \sim \mathcal{N}(0, 4^2), \quad n \in \{72, \ldots, 94\} \cup \{176, \ldots, 196\} \cup \{265, \ldots, 289\},
\end{aligned}
$$

while other values are set to $0$. In addition, small independent Gaussian perturbations are added to all entries, with

$$
\lambda_{ih} \leftarrow \lambda_{ih} + \varepsilon_{ih}, \quad \varepsilon_{ih} \sim \mathcal{N}(0, 0.15^2)
$$

| Model | Relative Frobenius distance | | | RV coefficient | | |
|---|---|---|---|---|---|---|
| | $d(\boldsymbol{Y}_1, \widehat{\boldsymbol{Y}}_1)$ | $d(\boldsymbol{Y}_2, \widehat{\boldsymbol{Y}}_2)$ | $d(\boldsymbol{Y}_3, \widehat{\boldsymbol{Y}}_3)$ | $RV_{\boldsymbol{Y}_1, \widehat{\boldsymbol{Y}}_1}$ | $RV_{\boldsymbol{Y}_2, \widehat{\boldsymbol{Y}}_2}$ | $RV_{\boldsymbol{Y}_3, \widehat{\boldsymbol{Y}}_3}$ |
| MSFRB | **0.059** (0.005) | **0.066** (0.006) | **0.064** (0.007) | **0.9999** | **0.9999** | **0.9999** |
| MSFA | 0.064 (0.005) | 0.074 (0.010) | 0.068 (0.007) | 0.9995 | 0.9993 | 0.9996 |
| EBMF | 0.062 (0.007) | 0.071 (0.008) | 0.068 (0.010) | **0.9999** | **0.9999** | **0.9999** |
| BicMix | 0.083 (0.009) | 0.096 (0.012) | 0.091 (0.013) | 0.9994 | 0.9993 | 0.9994 |
| SSLB | 0.278 (0.048) | 0.321 (0.054) | 0.293 (0.054) | 0.9975 | 0.9975 | 0.9962 |

Table 1: *Relative Frobenius distances and RV coefficients between the observed data $\boldsymbol{Y}_s$ and fitted values $\widehat{\boldsymbol{Y}}_s$ for each study. Relative Frobenius distances are reported as mean (standard deviation) across ten replicates, and smaller values indicate better reconstruction. RV coefficients are reported as mean values, with values closer to 1 indicating better agreement. The best values in each column are highlighted in bold.*

$$f_{hn} \leftarrow f_{hn} + \varepsilon_{hn}, \quad \varepsilon_{hn} \sim \mathcal{N}(0, 0.1^2)$$

We generate a single binary covariate representing biological sex of an individual from which the sample is taken, and it is randomly assigned to each sample in each study as a binomial distributed variable with study-specific probabilities of being male equal to $P(\text{male}) = (0.55, 0.45, 0.60)$. The regression coefficients $\beta_i^X$ are generated from $\mathcal{N}(0, 0.6^2)$ for all $i$. Finally, the idyiosyncratic variances are simulated as $\xi_{is}^2 \sim \mathcal{G}^{-1}(100, 30)$ for all $i$ and $s$. This ensures that the noise is small enough not to mask the block structure of the data matrix induced by the common factors and their loadings.

Then, for each study $s$ and sample $n$, the data is drawn from:

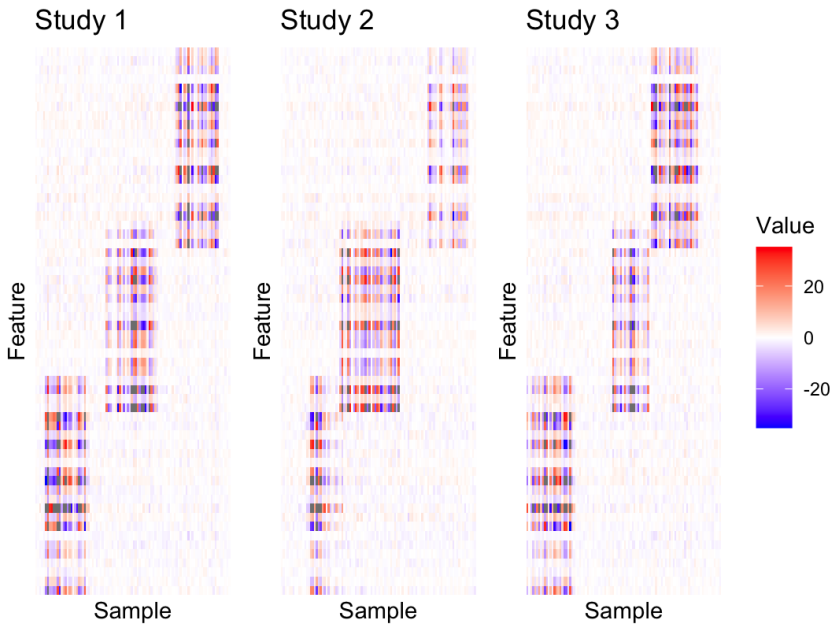
$$\mathbf{y}_{ns} \sim \mathcal{N}_p\big(\boldsymbol{\beta}^X \mathbf{x}_{ns} + \boldsymbol{\Lambda}\mathbf{f}_{ns} + \boldsymbol{\Lambda}_s\boldsymbol{\eta}_{ns},\ \boldsymbol{\Xi}_s\big)\,.$$

with $\boldsymbol{\Xi}_s = \text{diag}(\xi_{1s}^2, \ldots, \xi_{ps}^2)$.

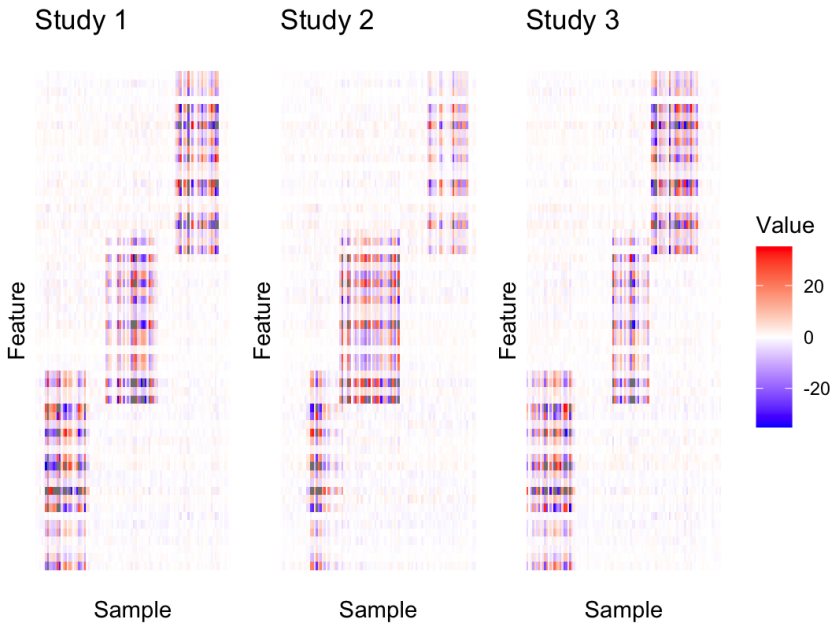
We set the maximum possible number of active common factors as $H = 20$ and study-specific factors as $H_s = 10$. Hyperparameters and initial values are set as described in Section B.1 of the Supplementary Material. The estimated data matrix is calculated as follows:

$$\widehat{\boldsymbol{Y}}_s = \mathbb{E}[\widehat{\boldsymbol{\beta}}^X]\boldsymbol{X}_s + \mathbb{E}[\widehat{\boldsymbol{\Lambda}}]\mathbb{E}[\widehat{\boldsymbol{F}}_s] + \mathbb{E}[\widehat{\boldsymbol{\Lambda}}_s]\mathbb{E}[\widehat{\boldsymbol{\eta}}_s].$$

We simulate the data matrix ten times with different seeds. Across all ten simulation replicates, the model correctly recovered the three active common factors. Moreover, all recovered common factors satisfied the unique-support identification criterion described in Section 4.4. Figure 2 presents the heatmap of the mean of the simulated true (a) and the estimated (b) data matrices for each of the three studies.




(a) *Heatmap of the mean across ten simulated data matrices* $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$




(b) *Heatmap of the mean across ten estimated by MSFRB data matrices* $\widehat{\mathbf{Y}}_1, \widehat{\mathbf{Y}}_2, \widehat{\mathbf{Y}}_3$

Figure 2: *Comparison of the heatmaps of the simulated and estimated by the MDFRB model data matrices. Color scales are truncated at the* 99*th percentile to improve visual contrast.*

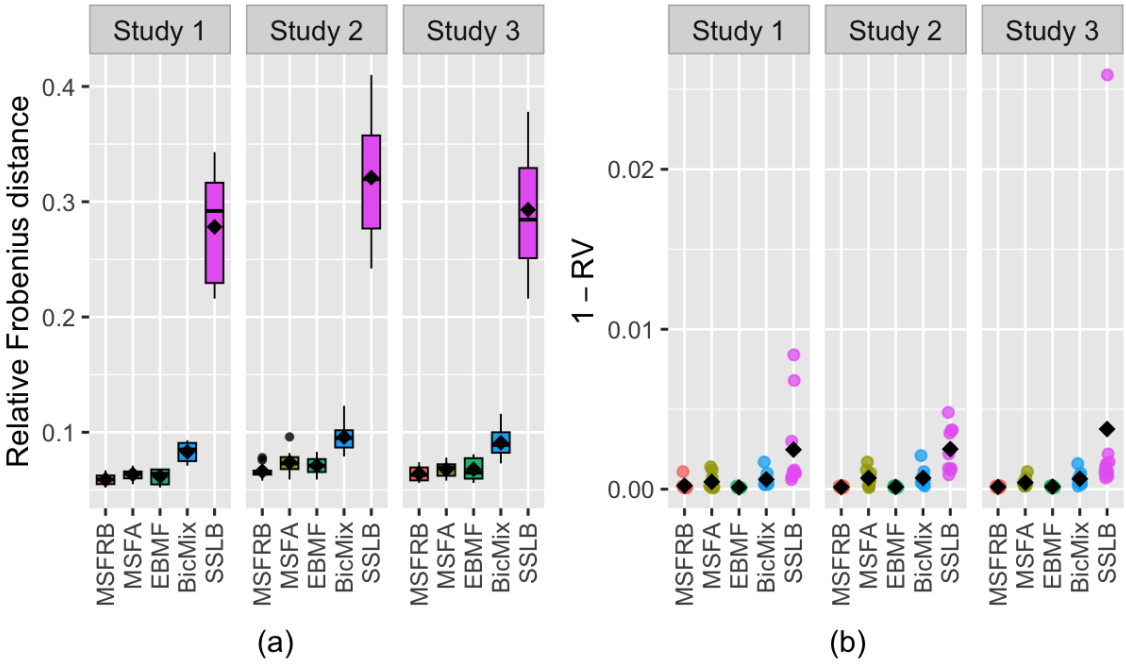



Figure 3: *Comparison of the data reconstruction accuracy across methods and studies based on ten simulation replicates. Panel (a) displays boxplots of the relative Frobenius distance between the observed data* $\boldsymbol{Y}_s$ *and the reconstructed matrices* $\widehat{\boldsymbol{Y}}_s$*, with points indicating the mean values. Smaller values correspond to better reconstruction accuracy. Panel (b) shows the quantity* $1 - \mathrm{RV}$*, where* RV *denotes the RV coefficient between* $\boldsymbol{Y}_s$ *and* $\widehat{\boldsymbol{Y}}_s$*. Values closer to zero indicate better agreement between the observed and reconstructed data. Results are presented separately for each study.*

As illustrated in Table 1 and Figure 3, MSFRB achieves the lowest reconstruction errors overall, although the differences relative to MSFA, EBMF and BicMix are modest. All four methods provide highly accurate reconstructions of the observed data, with relative Frobenius distances below 0.1 and RV coefficients close to one across all studies. In contrast, SSLB exhibits substantially larger reconstruction errors. This is likely due to its strong sparsity-inducing prior, which encourages both loadings and factor scores to be exactly or nearly zero outside a small number of biclusters. The simulated data, however, exhibit only approximate sparsity, with overlapping supports and small but nonzero values outside the main signal regions. Consequently, SSLB tends to shrink these weaker signals towards zero, leading to an underestimation of the underlying structure. The still high RV coefficients indicate that, despite inaccuracies at the level of individual entries, the overall structure of the data is largely preserved.

### 6.1.2 Recovery of model parameters $\boldsymbol{\Lambda}$, $\mathbf{F}_s$, $\boldsymbol{\Lambda}_s$, $\eta_s$, and $\beta_X$.

This simulation study is designed to assess the ability of the MSFRB model to recover the underlying model parameters, including common and study-specific factors, their corresponding loadings, and the covariate effects. We retain the same data-generating mechanism and dimensionality as in Section 6.1.1, but modify the construction of the common component to induce a clearly visible structure in both the common factor loadings and factor scores. In addition, we include one dense

common factor alongside three sparse common factors to evaluate the ability of the model to recover both sparse and dense latent structure. The first three sparse common factors and loadings are generated to exhibit sparse support pattern:

$$
\begin{aligned}
&\lambda_{i1} \sim \mathcal{N}(0, 2^2), \quad i = 1, \ldots, 20, \\
&\lambda_{i2} \sim \mathcal{N}(0, 2^2), \quad i = 21, \ldots, 40, \\
&\lambda_{i3} \sim \mathcal{N}(0, 2^2), \quad i = 41, \ldots, 60, \\
&f_{1n} \sim \mathcal{N}(0, 2^2), \quad n \in \{1, \ldots, 50\} \cup \{101, \ldots, 150\} \cup \{201, \ldots, 250\}, \\
&f_{2n} \sim \mathcal{N}(0, 2^2), \quad n \in \{51, \ldots, 100\} \cup \{151, \ldots, 200\} \cup \{251, \ldots, 300\}, \\
&f_{3n} \sim \mathcal{N}(0, 2^2), \quad n \in \{1, \ldots, 100\} \cup \{121, \ldots, 170\} \cup \{201, \ldots, 300\},
\end{aligned}
$$

while all remaining entries are set to $0$. The dense common factor is generated by sampling its factor scores from $\mathcal{N}(0, 1.5^2)$, and its loadings from $\mathcal{N}(0, 2^2)$.

Study-specific factors are generated as $\boldsymbol{\eta}_{ns}^{\top} \sim \mathcal{N}_2(\mathbf{0}, \mathbf{I}_2)$ for $n = 1, \ldots, N_s$ and $s = 1, \ldots, S$, with their corresponding loadings generated as $\lambda_{ihs} \sim \mathcal{N}(0, 1)$ for all $i$, $s$, and $h = 1, 2$. The idyiosyncratic variances are simulated as $\xi_{is}^2 \sim \mathcal{G}^{-1}(100, 30)$ for all $i$ and $s$.

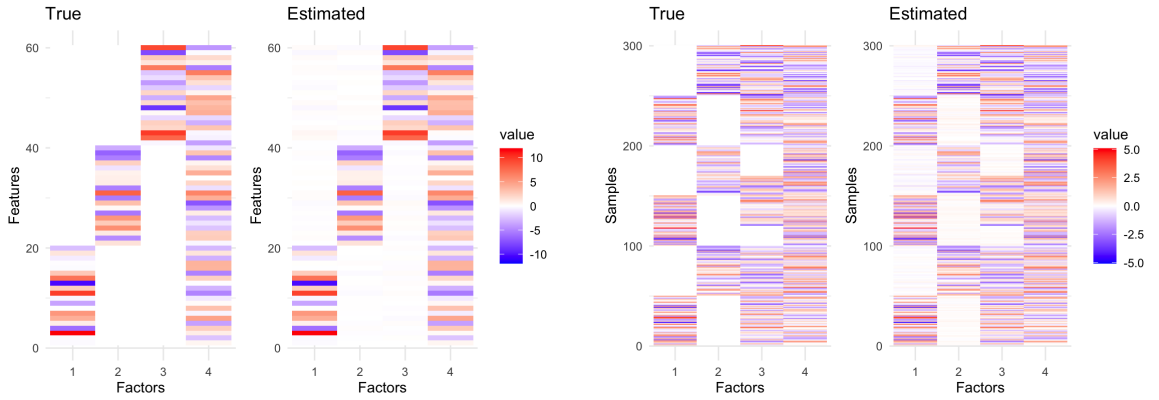
We retain the sex covariate generated as described in Section 6.1.1 and additionally introduce a continuous age covariate. Age is simulated from a normal distribution with standard deviation of $8$ and study-specific means of $55$, $58$, and $52$ years in the first, second, and third studies, respectively. To ensure realistic values, age values are truncated to lie between $18$ and $90$ years.

We set the maximum possible number of active common and study-specific factors to $H = 20$ and $H_s = 10$, respectively. Hyperparameters and initial values are chosen as described in Section B.1 of the Supplementary Material. We again simulate $10$ replicate datasets. In all replicates, the model correctly recovered the four active common factors, all of which satisfied the unique-support identification criterion.

Below, we present heatmaps of the true and estimated common factor loadings, common factor scores, study-specific factor loadings, and covariates for one randomly selected replicate. As factor loadings and factor scores are identified up to sign switching and column permutations, we re-arranged the estimated matrices to have the same sign an order as the true factor loadings/scores. For better visual clarity, we only exhibited the columns corresponding to the active factors in the estimated factor loadings and scores matrices. We take posterior means as estimated values for all the parameters in question.

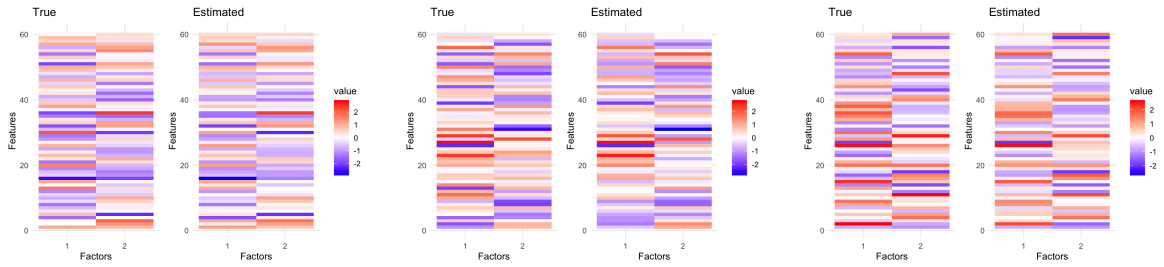
Figure 4 demonstrates that the model successfully uncovered both sparse and dense common factors and factor loadings. As shown in Figure 5, the study-specific loadings are also recovered accurately. Nevertheless, the primary objective of the model is to identify the latent structure shared across studies, which is encoded in the common factors and loadings. In contrast, the study-specific factors capture variation unique to individual studies, which in transcriptomic applications may arise from differences in experimental platforms, laboratory protocols, batch effects, and other study-specific sources of variation.

Finally, Figure 6 compares the simulated and estimated covariate coefficients. The estimated coefficient matrix includes an intercept term, which is inferred automatically by the model. The intercept is necessary as, contrary to the usual practice in factor models, he MSFRB model is fitted directly to the original data without prior centering or scaling. This avoids removing variation that may be associated with the observed covariates and allows covariate effects to be estimated jointly




(a) *Common factor loadings* (b) *Common factors*

Figure 4: *The heatmap of the mean across ten true and estimated common factor loadings and scores.*




(a) *Factor loadings: study 1* (b) *Factor loadings: study 2* (c) *Factor loadings: study 3*

Figure 5: *The heatmap of the mean across ten true and estimated subject-specific factor loadings.*

with the latent structure. As evident from Figure 6, the MSFRB model recovered the true covariate effects with a high degree of accuracy.

The primary objective of this simulation study is to evaluate the extent to which the competing methods recover the underlying latent structure. As most of the methods considered can be viewed as biclustering approaches, or more broadly as methods for identifying shared structure across features and samples (e.g. common pathways), we focus on recovery of the common component. Specifically, we evaluate the estimation accuracy of the matrix of common factor loadings, $\mathbf{\Lambda}$, the matrix of common factor scores, $\mathbf{F}$, and their product $\mathbf{\Lambda F}$, which represents the shared signal. This allows us to distinguish between accurate reconstruction of the overall signal and accurate recovery of its underlying factorised representation. To assess performance, we report relative Frobenius distances and RV coefficients, which quantify discrepancies in magnitude and agreement in overall structure, respectively, between the true and estimated matrices.

For all methods, estimated factor loadings and factor scores were first aligned to the true factors to account for sign and column permutation indeterminacies inherent to factor models. Alignment was performed by matching estimated and true factors based on the combined similarity of their loading and score vectors, followed by sign correction. The common signal was then reconstructed as the product of the aligned estimated loading and factor matrices, $\widehat{\mathbf{\Lambda}}\widehat{\boldsymbol{F}}$. Since BicMix allows an over-

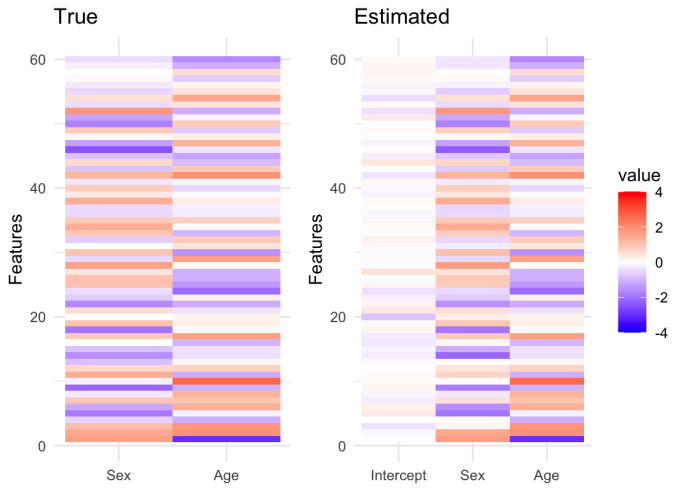



Figure 6: *The heatmap of the mean across ten true and estimated covariate coefficients*

| Model | Relative Frobenius distance | | | RV coefficient | | |
|---|---|---|---|---|---|---|
| | $d(\boldsymbol{\Lambda}, \widehat{\boldsymbol{\Lambda}})$ | $d(\boldsymbol{F}, \widehat{\boldsymbol{F}})$ | $d(\boldsymbol{\Lambda F}, \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}})$ | $RV_{\boldsymbol{\Lambda}, \widehat{\boldsymbol{\Lambda}}}$ | $RV_{\boldsymbol{F}, \widehat{\boldsymbol{F}}}$ | $RV_{\boldsymbol{\Lambda F}, \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}}}$ |
| MSFRB | 0.119 [0.059] | 0.148 [0.054] | **0.085** [0.023] | **0.9966** | 0.9965 | **0.9996** |
| MSFA | 0.780 [0.015] | 0.762 [0.019] | 0.942 [0.006] | 0.7886 | 0.9919 | 0.7591 |
| EBMF | **0.109** [0.103] | **0.088** [0.094] | 0.108 [0.055] | 0.9881 | **0.9982** | 0.9864 |
| BicMix | 1.701 [0.251] | 1.336 [0.344] | 2.991 [1.253] | 0.1637 | 0.8561 | 0.0845 |
| SSLB | 0.901 [0.460] | 0.873 [0.272] | 1.919 [1.037] | 0.4676 | 0.9065 | 0.4008 |

Table 2: *Relative Frobenius distances and RV coefficients between the true and estimated common loading matrix $\boldsymbol{\Lambda}$, the common factor score matrix $\boldsymbol{F}$, and their product $\boldsymbol{\Lambda F}$. Values are reported as mean [standard deviation] across simulation replicates. Smaller relative Frobenius distances indicate better recovery, while RV coefficients closer to one indicate stronger agreement. The best values in each column are highlighted in bold.*

complete representation and the simulated common signal contains both sparse and dense factors, all recovered BicMix components were considered during the matching procedure rather than restricting attention to factors classified as sparse in both dimensions. This ensures that recovery is assessed with respect to the complete common signal generated under the simulation design.

The results presented in Table 2 indicate that both MSFRB and EBMF are able to recover the underlying common structure accurately, although their strengths differ. EBMF achieves the lowest relative Frobenius distances and the highest RV coefficients for the individual loading and factor score matrices, $\boldsymbol{\Lambda}$ and $\boldsymbol{F}$, suggesting particularly accurate recovery of the factorised representation. In contrast, MSFRB achieves the best performance in recovering the common signal $\boldsymbol{\Lambda F}$, attaining the smallest reconstruction error and the highest RV coefficient among all competing methods. This indicates that MSFRB is especially effective at capturing the shared latent structure represented by the common component. MSFA performs substantially worse, particularly with respect to recovery of the common signal, despite achieving reasonably high RV coefficients for the factor score matrix. BicMix and SSLB exhibit considerably poorer performance across all metrics, with large reconstruc-

tion errors and substantially lower RV coefficients. In particular, BicMix appears unable to recover the sparse and dense common structure simultaneously, while SSLB shows high variability across simulation replicates. These findings are consistent with the boxplots in Figure 7, where MSFRB displays the most stable performance and the lowest errors in recovering the common signal across repeated simulations.

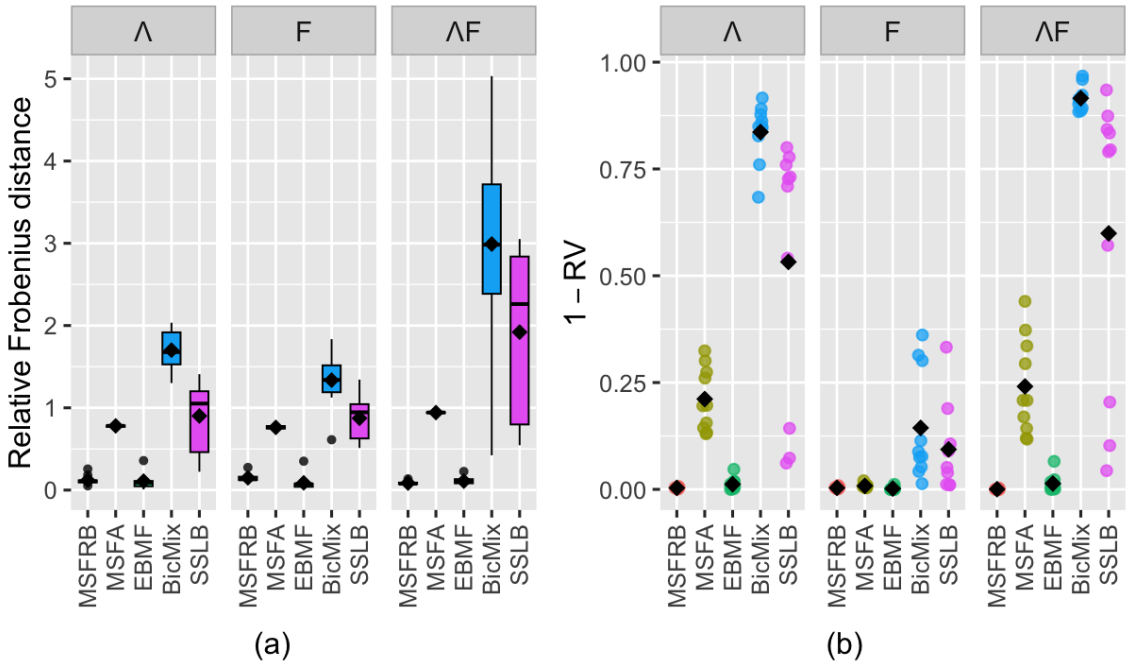



Figure 7: *Comparison of the recovery of the common signal across methods based on ten simulation replicates. Panel (a) displays boxplots of the relative Frobenius distance between the true common component* $\boldsymbol{\Lambda F}$ *and its estimate* $\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}}$*, with points indicating the mean values. Smaller values correspond to more accurate recovery of the common signal. Panel (b) shows the quantity* $1-\mathrm{RV}$*, where* RV *denotes the RV coefficient between* $\boldsymbol{\Lambda F}$ *and* $\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}}$*. Values closer to zero indicate stronger agreement between the true and estimated common components.*

The poor performance of BicMix and SSLB can likely be attributed to the fact that the simulated common structure combines sparse, dense, and overlapping factors while also including study-specific and covariate-driven variation. This creates a more complex decomposition problem than a purely sparse biclustering setting, and could make it difficult for these methods to separate the shared common component from other sources of structure.

## 6.2 Simulation study 2: Model performance on various dimensions

This simulation study investigates the comparative performance of the MSFRB model as the dimensionality of the data increases. In particular, we examine how recovery of the shared latent structure varies with the number of features $p$ and the number of samples per study $N_s$, while fixing the number of studies at $S = 4$. We consider the combinations $p \in \{100, 500, 1000\}$ and $N_s \in \{100, 500\}$. For each setting, ten replicate datasets are generated.

Performance is assessed using the relative Frobenius distance between the observed and recon-

structed data matrices $d(\boldsymbol{Y}_s, \widehat{\boldsymbol{Y}}_s)$ for each study $s$, as well as the distance between the true and estimated common signal $d(\boldsymbol{\Lambda F}, \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}})$. The latter serves as the primary measure of performance, as it captures the recovery of the structure shared across studies. By focusing on the product $\boldsymbol{\Lambda F}$, we avoid issues of scale and rotational indeterminacy inherent in factor models.

The data are generated as follows. We fix the number of common factors to $H = 5$, of which two are dense and three are sparse, and the number of study-specific factors to $H_s = 2$ for each study. For each study $s = 1, \ldots, S$ and sample $n = 1, \ldots, N_s$, the study-specific latent factors are drawn as $\boldsymbol{\eta}_{ns} \sim \mathcal{N}_3(\mathbf{0}, \mathbf{I}_3)$, with corresponding loadings generated independently as

$$\lambda_{ihs} \sim \mathcal{N}(0, 0.3^2), \qquad i = 1, \ldots, p \text{ and } h = 1, \ldots, 2.$$

We consider an intercept-only design with $p_x = 1$ and $\mathbf{x}_{ns} = 1$ for all $n$ and $s$, and generate regression coefficients as

$$\beta_i^X \sim \mathcal{N}(0, 2^2), \qquad i = 1, \ldots, p.$$

The common loading matrix $\boldsymbol{\Lambda}$ is constructed as a mixture of dense and sparse columns. For the two dense factors,

$$\lambda_{ih} \sim \mathcal{N}(0, 2^2), \qquad i = 1, \ldots, p \text{ and } h = 1, 2.$$

For the remaining three factors, sparsity is induced by randomly selecting between $30\%$ and $60\%$ of entries to be nonzero, drawn from $\mathcal{N}(0, 2^2)$, with the remaining entries set to zero.

The common factor matrix $\boldsymbol{F}$ is generated as

$$f_{hn} \sim \mathcal{N}(0, 2^2), \qquad h = 1, \ldots, 5 \text{ and } n = 1, \ldots, N,$$

where $N = \sum_{s=1}^{S} N_s$. The first two factors are dense, while the remaining factors are sparsified by randomly setting between $40\%$ and $70\%$ of their entries to zero.

We assume diagonal noise covariance matrices $\boldsymbol{\Xi}_s = \mathrm{diag}(\xi_{1s}^2, \ldots, \xi_{ps}^2)$, where the idiosyncratic errors are generated independently as

$$\xi_{is}^2 \sim \mathcal{G}^{-1}(7, 1), \qquad i = 1, \ldots, p \text{ and } s = 1, \ldots, S.$$

Finally, for each study $s$ and sample $n$, the data are generated according to

$$\mathbf{y}_{ns} \sim \mathcal{N}_p\big(\boldsymbol{\beta}^X \mathbf{x}_{ns} + \boldsymbol{\Lambda}\mathbf{f}_{ns} + \boldsymbol{\Lambda}_s \boldsymbol{\eta}_{ns},\ \boldsymbol{\Xi}_s\big)\,.$$

This construction yields $S = 4$ data matrices $\mathbf{Y}_s$ that share a common latent structure while retaining study-specific variation and noise, allowing us to assess how well each method recovers the shared signal as the dimensionality increases.

As in the first simulation study, we compare the performance of the MSFRB model with four competing approaches, namely the MSFA model of Hansen et al. (2025), EBMF of Wang and Stephens (2021), BicMix of Gao et al. (2016), and SSLB of Moran et al. (2021). This simulation setting is particularly challenging because the latent structure is characterised by complex sparsity patterns and weak signals, including overlapping supports, rather than by extreme dimensionality alone. These features make recovery of the common component difficult even when the sample size is relatively large.

| Model | $N_s$ | Relative Frobenius distance | | | | | RV coefficient |
|---|---|---|---|---|---|---|---|
| | | $d(\boldsymbol{Y}_1, \widehat{\boldsymbol{Y}}_1)$ | $d(\boldsymbol{Y}_2, \widehat{\boldsymbol{Y}}_2)$ | $d(\boldsymbol{Y}_3, \widehat{\boldsymbol{Y}}_3)$ | $d(\boldsymbol{Y}_4, \widehat{\boldsymbol{Y}}_4)$ | $d(\boldsymbol{\Lambda F}, \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}})$ | $RV_{\boldsymbol{\Lambda F}, \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}}}$ |
| MSFRB | 100 | **0.057** [0.003] | **0.057** [0.002] | **0.057** [0.002] | **0.057** [0.004] | **0.050** [0.025] | 0.9992 |
| MSFA | 100 | 0.060 [0.003] | 0.061 [0.003] | 0.061 [0.003] | 0.061 [0.004] | 0.864 [0.005] | 0.8483 |
| EBMF | 100 | 0.062 [0.007] | 0.063 [0.004] | 0.061 [0.005] | 0.060 [0.005] | 0.105 [0.023] | 0.**9998** |
| BicMix | 100 | 0.063 [0.005] | 0.062 [0.004] | 0.062 [0.005] | 0.061 [0.005] | 1.236 [0.238] | 0.1593 |
| SSLB | 100 | 0.689 [0.014] | 0.678 [0.063] | 0.734 [0.039] | 0.682 [0.080] | 0.694 [0.041] | 0.9769 |
| MSFRB | 500 | **0.056** [0.002] | 0.057 [0.002] | 0.056 [0.002] | 0.057 [0.002] | **0.026** [0.004] | **0.9999** |
| MSFA | 500 | **0.056** [0.002] | **0.055** [0.002] | **0.055** [0.002] | **0.055** [0.002] | 0.863 [0.005] | 0.8479 |
| EBMF | 500 | 0.059 [0.002] | 0.060 [0.002] | 0.060 [0.002] | 0.059 [0.003] | 0.046 [0.004] | **0.9999** |
| BicMix | 500 | 0.060 [0.003] | 0.061 [0.004] | 0.060 [0.002] | 0.060 [0.003] | 1.008 [0.008] | 0.1145 |
| SSLB | 500 | 1 | 1 | 1 | 1 | – | – |

Table 3: *Relative Frobenius distances between the simulated data $\boldsymbol{Y}_s$ and fitted values $\widehat{\boldsymbol{Y}}_s$ for each study, together with the relative Frobenius distance and RV coefficient for the recovered common signal $\boldsymbol{\Lambda F}$, for fixed $p = 100$ and varying $N_s$. Values are reported as mean [standard deviation] across ten replicate datasets, while RV coefficients are reported as mean values.For BicMix, common component recovery metrics are computed only for four replicates in which sparse-sparse components were selected. The best values in each column are highlighted in bold.*

Across Tables 3 and 4, the reconstruction errors for the observed data $\boldsymbol{Y}_s$ are very similar for MSFRB, MSFA, and EBMF, indicating that all three methods capture the overall variation in the data to a comparable extent. However, differences become apparent when considering the recovery of the common signal. MSFRB and EBMF achieve substantially lower errors and higher RV coefficients for the common component, whereas MSFA exhibits much larger errors despite comparable reconstruction of the observed data, which is likely due to information switching between shared and study-specific components.

The behaviour of SSLB is unstable. In particular, for $p = 100$ and $N_s = 500$, SSLB fails to recover any common signal, resulting in a degenerate solution. More generally, its performance deteriorates as the sample size increases. This may be explained by the strong sparsity-inducing prior in SSLB, which can over-shrink weak and overlapping signals, making it difficult to recover the common structure in this setting. In some $p, N_s$ settings, also the BicMix model did not return any components classified as sparse in both loadings and factor scores, and the corresponding biclustering component was therefore not available for comparison.

Overall, these results highlight that the main difficulty lies in recovering the shared latent structure rather than reconstructing the observed data, and that only MSFRB and EBMF perform reliably in this respect.

# 7 Real data application

We applied the proposed MSFRB model to a multi-study colorectal cancer gene expression dataset constructed from the R-package `curatedCRCData_mRNA`. The dataset was assembled using a custom preprocessing pipeline designed to harmonise microarray-based mRNA expression studies while retaining only molecular features observed across all selected studies. The pipeline allows the user to specify the expression platform, the minimum number of observations per study, the max-

| Model | $N_s$ | Relative Frobenius distance | | | | | RV coefficient |
|---|---|---|---|---|---|---|---|
| | | $d(\boldsymbol{Y}_1, \widehat{\boldsymbol{Y}}_1)$ | $d(\boldsymbol{Y}_2, \widehat{\boldsymbol{Y}}_2)$ | $d(\boldsymbol{Y}_3, \widehat{\boldsymbol{Y}}_3)$ | $d(\boldsymbol{Y}_4, \widehat{\boldsymbol{Y}}_4)$ | $d(\boldsymbol{\Lambda F}, \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}})$ | $RV_{\boldsymbol{\Lambda F}, \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}}}$ |
| MSFRB | 100 | **0.058** [0.002] | **0.058** [0.004] | **0.058** [0.003] | **0.057** [0.003] | 0.439 [0.101] | 0.9403 |
| MSFA | 100 | **0.058** [0.002] | **0.058** [0.002] | **0.058** [0.003] | 0.058 [0.003] | 0.865 [0.007] | 0.8476 |
| EBMF | 100 | 0.061 [0.003] | 0.061 [0.002] | 0.062 [0.002] | 0.062 [0.003] | **0.105** [0.018] | **0.9998** |
| BicMix | 100 | 0.062 [0.003] | 0.061 [0.004] | 0.061 [0.004] | 0.061 [0.003] | 1.006 [–] | 0.0028 |
| SSLB | 100 | 0.173 [0.011] | 0.168 [0.010] | 0.173 [0.016] | 0.177 [0.015] | 0.323 [0.119] | 0.9690 |
| MSFRB | 500 | **0.058** [0.002] | 0.059 [0.002] | **0.059** [0.002] | 0.060 [0.002] | 0.235 [0.094] | 0.9808 |
| MSFA | 500 | **0.058** [0.002] | **0.058** [0.002] | **0.059** [0.002] | **0.058** [0.003] | 0.862 [0.004] | 0.8459 |
| EBMF | 500 | 0.062 [0.003] | 0.062 [0.003] | 0.062 [0.003] | 0.062 [0.004] | **0.045** [0.009] | 0.**9999** |
| BicMix | 500 | 0.061 [0.003] | 0.061 [0.004] | 0.061 [0.004] | 0.061 [0.003] | 1.006 [–] | 0.0028 |
| SSLB | 500 | 0.824 [0.052] | 0.828 [0.043] | 0.819 [0.043] | 0.825 [0.040] | 0.810 [0.022] | 0.8849 |

Table 4: *Relative Frobenius distances between the simulated data $\boldsymbol{Y}_s$ and fitted values $\widehat{\boldsymbol{Y}}_s$ for each study, together with the relative Frobenius distance and RV coefficient for the recovered common signal $\boldsymbol{\Lambda F}$, for fixed $p = 500$ and varying $N_s$. Values are reported as mean [standard deviation] across ten replicate datasets, while RV coefficients are reported as mean values. For BicMix, common component recovery metrics are reported only for two replicates in which sparse-sparse components were identified. The best values in each column are highlighted in bold.*

imum acceptable fraction of missing values before gene exclusion, and the demographic or clinical variables to retain. After filtering, studies are merged by restricting the analysis to a common set of genes with complete expression values for all observations across all included studies.

Using the pipeline, we selected the $2,000$ most variable common genes. The final dataset comprised eight colorectal cancer studies: `GSE14333_eset` (290 samples), `GSE17536_eset` (177 samples), `GSE17538.GPL570_eset` (232 samples), `GSE2109_eset` (426 samples), `GSE21815_eset` (137 samples), `GSE26682.GPL570_eset` (156 samples), `GSE26682.GPL96_eset` (155 samples), and `GSE39582_eset` (565 samples), yielding a total of $2,138$ samples. Each study contributed measurements for the same set of $2,000$ genes. Study membership was encoded by the variable `study_id`. Age and gender were retained as observed covariates and were included in the regression component of the model where available.

This dataset provides a natural application for MSFRB because it combines gene expressions collected across multiple independent studies conducted over different time periods and using different experimental protocols and microarray platforms. Such heterogeneity is expected to induce both biological signal shared across studies and study-specific variation due to technical, platform-related, and cohort-specific effects. The MSFRB decomposition is therefore well suited to this setting: common sparse factors are used to identify biclusters of genes and samples that represent reproducible structure across studies, while study-specific factors account for residual heterogeneity and potential batch effects. The inclusion of age and gender allows the shared latent structure to be estimated after adjustment for observed demographic variation.

For this real data analysis we used the CAVI implementation of MSFRB. The maximum number of common factors was set to be $H = 30$ and study-specific factors was set to be $H_s = 20$, with the sparse ESP prior allowing the effective number of active factors to be inferred from the data. The algorithm converged after 198 iterations. A total of six common factors were identified, all satisfying the identifiability condition based on unique support. The numbers of study-specific factors across

| Model | $N_s$ | Relative Frobenius distance | | | | | RV coefficient |
|---|---|---|---|---|---|---|---|
| | | $d(\boldsymbol{Y}_1, \widehat{\boldsymbol{Y}}_1)$ | $d(\boldsymbol{Y}_2, \widehat{\boldsymbol{Y}}_2)$ | $d(\boldsymbol{Y}_3, \widehat{\boldsymbol{Y}}_3)$ | $d(\boldsymbol{Y}_4, \widehat{\boldsymbol{Y}}_4)$ | $d(\boldsymbol{\Lambda F}, \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}})$ | $RV_{\boldsymbol{\Lambda F}, \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}}}$ |
| MSFRB | 100 | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.0000 |
| MSFA | 100 | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.0000 |
| EBMF | 100 | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.0000 |
| BicMix | 100 | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.0000 |
| SSLB | 100 | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.0000 |
| MSFRB | 500 | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.0000 |
| MSFA | 500 | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.0000 |
| EBMF | 500 | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.0000 |
| BicMix | 500 | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.0000 |
| SSLB | 500 | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.000 [0.000] | 0.0000 |

Table 5: *Relative Frobenius distances between the simulated data $\boldsymbol{Y}_s$ and fitted values $\widehat{\boldsymbol{Y}}_s$ for each study, together with the relative Frobenius distance and RV coefficient for the recovered common signal $\boldsymbol{\Lambda F}$, for fixed $p = 1000$ and varying $N_s$. Values are reported as mean [standard deviation] across ten replicate datasets, while RV coefficients are reported as mean values. The best values in each column are highlighted in bold.*

the eight studies were $5, 8, 11, 8, 9, 3, 14$, and $9$, respectively. Biclusters of genes and samples were extracted from the estimated common factors by thresholding the posterior second moments of the factor loadings and factor scores. Active factors were selected using a relative energy criterion, and gene–sample memberships were defined via factor-specific supports.

The extracted biclusters exhibit clear block-like patterns, indicating subsets of genes that are coherently up- or down-regulated within specific groups of patients. These patterns are consistent across studies and suggest that the inferred common factors capture shared sources of biological variation, while allowing for heterogeneity in the remaining samples.

We compared the performance of MSFRB with the same competing methods considered in the simulation studies, namely MSFA, EBMF, SSLB, and BixMix, using the relative Frobenius distance between the observed data and its reconstruction as a common evaluation metric. Although reconstruction accuracy is not the primary objective of the model, this metric provides a convenient way to assess how well each method captures the main structure of the data.

The results are reported in Table 6. MSFRB achieved the lowest overall reconstruction error, with a relative Frobenius distance of $0.099$. EBMF performed almost identically, attaining the same overall error and comparable reconstruction accuracy across all studies. In contrast, SSLB and MSFA exhibited substantially larger reconstruction errors, while BicMix performed worse than all other methods. These results suggest that MSFRB and EBMF are better able to capture the dominant sources of variation present in the data than the competing approaches.

# 8 Conclusion

The paper proposes a multi-study factor regression biclustering (MSFRB) model for the comprehensive analysis of high-dimensional molecular data collected across multiple studies. The model integrates multi-study factor analysis, factor regression, and biclustering within a unified Bayesian framework, enabling simultaneous modelling of shared latent structure as well as accounting for study-specific heterogeneity and observed covariate effects. Such a combination has not been con-

| Model | Study 1 | Study 2 | Study 3 | Study 4 | Study 5 | Study 6 | Study 7 | Study 8 | Overall |
|---|---|---|---|---|---|---|---|---|---|
| MSFRB | 0.131 | 0.056 | 0.059 | 0.116 | 0.143 | 0.066 | 0.062 | 0.099 | 0.099 |
| MSFA | 0.528 | 0.297 | 0.290 | 0.473 | 0.557 | 0.421 | 0.350 | 0.416 | 0.444 |
| EBMF | 0.133 | 0.056 | 0.060 | 0.115 | 0.150 | 0.069 | 0.062 | 0.098 | 0.099 |
| SSLB | 0.243 | 0.125 | 0.139 | 0.209 | 0.225 | 0.191 | 0.103 | 0.186 | 0.188 |
| BixMix | 0.794 | 0.734 | 0.697 | 0.774 | 0.800 | 0.781 | 0.796 | 0.711 | 0.764 |

Table 6: *Relative Frobenius distance between observed and reconstructed data across studies.*

sidered within a single coherent modelling framework, providing a flexible approach to integrative analysis in complex multi-study settings.

A central component of the proposed framework is the use of a sparse exchangeable shrinkage process (ESP) prior on both common factor loadings and factor scores. This prior induces sparsity at multiple levels and supports automatic factor selection, allowing the model to adapt to both sparse and moderately dense latent structures. While variational inference yields continuous approximations, the shrinkage mechanism leads to a clear practical separation between active and negligible components, which can be formalised through simple post-processing.

An additional contribution of this work is a principled approach to factor identification. Building on the structure induced by the model, we characterise active factors through their posterior energy and assess their identifiability via uniqueness of support across features and samples. This provides an interpretable criterion for distinguishing shared and study-specific components and complements the biclustering interpretation of the model. To enable scalable inference under the ESP prior, we developed a variational inference framework, including both coordinate ascent and stochastic variational algorithms. In particular, we derive mean-field variational updates for the ESP prior, enabling tractable and efficient approximation of its multi-level sparsity structure. The resulting procedures exploit the conditional structure of the model and scale well to high-dimensional settings.

Simulation studies and real data applications demonstrate that the MSFRB model accurately recovers latent structure in most settings, effectively separates shared and study-specific sources of variation, and reliably estimates covariate effects across a range of problem dimensions. These results highlight the advantages of the proposed framework over existing multi-study factor analysis and biclustering approaches.

The MSFRB model provides a flexible and principled tool for integrative analysis in multi-study settings. Several extensions can be developed within this framework. In particular, the Gaussian observation model may be relaxed to accommodate non-Gaussian response distributions, and additional structure—such as dependence or grouping among latent factors—can be incorporated through structured shrinkage priors (in the spirit of Schiavon et al. (2022)). These extensions can be pursued while retaining the factor regression structure and multi-study decomposition underlying the model.

# References

Abdi, H. (2007). Rv coefficient and congruence coefficient. *Encyclopedia of measurement and statistics*, 849(853):92.

Anderson, T. and Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume V:111–150.

Avalos-Pacheco, A., Rossell, D., and Savage, R. S. (2022). Heterogeneous Large Datasets Integration Using Bayesian Factor Regression. *Bayesian Analysis*, 17(1):33–66.

Bhattacharya, A. and Dunson, D. (2011). Sparse bayesian infinite factor models. *Biometrika*, 98(2):291–306.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Bortolato, E. and Canale, A. (2024). Adaptive partition factor analysis. *Journal of the American Statistical Association*.

Chandra, N. K., Dunson, D. B., and Xu, J. (2025). Inferring covariance structure from multiple data sources via subspace factor analysis. *Journal of the American Statistical Association*, 120(550):1239–1253.

Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, 8:93–103.

De Vito, R. and Avalos-Pacheco, A. (2025). Multi-study factor regression model: An application in nutritional epidemiology. *Statistics in Medicine*, 44(10–12):e70108.

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2019). Multi-study factor analysis. *Biometrics*, 75(1):337–346.

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2021). Bayesian multistudy factor analysis for high-throughput biological data. *The Annals of Applied Statistics*, 15(4):1723 – 1741.

Denitto, M., Bicego, M., Farinelli, A., and Figueiredo, M. (2017). Spike and slab biclustering. *Pattern Recognition*, 72:186–195.

Eren, K., Deveci, M., Küçüktunç, O., and Çatalyürek, V. (2012). A comparative analysis of biclustering algorithms for gene expression data. *Briefings in Bioinformatics*, 14(3):279–292.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer-Verlag, New York.

Frühwirth-Schnatter, S. (2023). Generalized cumulative shrinkage process priors with applications to sparse Bayesian factor analysis. *Philosophical Transactions of the Royal Society A*, (381):381:20220148.

Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2024). Sparse finite bayesian factor analysis when the number of factors is unknown. *Bayesian Analysis*, page accepted for publication.

Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2025). Sparse bayesian factor analysis when the number of factors is unknown (with discussion). *Bayesian Analysis*, 20(1).

Gao, C., Brown, C. D., and Engelhardt, B. E. (2013). A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects. *ArXiv: 13104792*.

Gao, C., McDowell, I. C., Zhao, S., Brown, C. D., and Engelhardt, B. E. (2016). Context specific and differential gene co-expression networks via bayesian biclustering. *PLoS Comput Biol*, 12(7).

Grabski, I. N., Vito, R. D., Trippa, L., and Parmigiani, G. (2023). Bayesian combinatorial MultiStudy factor analysis. *The Annals of Applied Statistics*, 17(3):2212 – 2235.

Grushanina, M. and Frühwirth-Schnatter, S. (2025). Dynamic Mixture of Finite Mixtures of Factor Analysers. *Bayesian Analysis*, pages 1 – 30.

Hansen, B., Avalos-Pacheco, A., Russo, M., and Vito, R. D. (2025). Fast variational inference for bayesian factor analysis in single and multi-study settings. *Journal of Computational and Graphical Statistics*, 34(1):96–108.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67:123–129.

Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., Khamiakova, T., Van Sanden, S., Lin, D., Talloen, W., Bijnens, L., Göhlmann, H. W. H., Shkedy, Z., and Clevert, D.-A. (2010). Fabia: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347.

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M. B. (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *Genome research*, 13 4:703–16.

Knowles, D. and Ghahramani, Z. (2011). Nonparametric bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, 5(2B):1534–1552.

Kowal, D. R. and Canale, A. (2023). Semiparametric functional factor models with Bayesian rank selection. *Bayesian Analysis*, 18:1161–1189.

Lazzeroni, L. and Owen, A. (2002). Plaid models for gene expression data. *Statistica Sinica*, 12(1):61–86.

Legramanti, S., Durante, D., and Dunson, D. (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika*, 107(3):745–752.

Li, G., Ma, Q., Tang, H., Paterson, A. H., and Xu, Y. (2009). Qubic: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Research*, 37:e101 – e101.

Liang, M., Hansen, B., Avalos-Pacheco, A., and De Vito, R. (2026). A tutorial on bayesian multistudy factor analysis with applications in nutrition and genomics. *Statistics in Medicine*, 45(10-12):e70531.

Lopes, H. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–67.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2006). *Sparse Statistical Modelling in Gene Expression Genomics*, pages 155–176.

Madeira, S. and Oliveira, A. (2004). Biclustering algorithms for biological data analysis: a survey. ieee/acm trans comput biol bioinform 1:24-45. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 1:24–45.

McParland, D., Gormley, I. C., McCormick, T. H., Clark, S. J., Kabudula, C. W., and Collinson, M. A. (2014). Clustering South African households based on their asset status using latent variable models. *The Annals of Applied Statistics*, 8(2):747 – 776.

Moran, G. E., Ročková, V., and George, E. I. (2021). Spike-and-slab Lasso biclustering. *The Annals of Applied Statistics*, 15(1):148 – 173.

Padilha, V. and Campello, R. (2017). A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics*, 18.

Pontes, B., Giráldez, R., and Aguilar-Ruiz, J. S. (2015). Biclustering on expression data: A review. *Journal of Biomedical Informatics*, 57:163–180.

Poworoznek, E., Ferrari, F., and Dunson, D. (2021). Efficiently resolving rotational ambiguity in bayesian matrix sampling with matching. *ArXiv: 2107.13783*.

Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129.

Rohe, K. and Zeng, M. (2023). Vintage factor analysis with varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1037–1060.

Ročková, V. and George, E. (2016). Fast bayesian factor analysis via automatic rotation to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622.

Roy, A., Lavine, I., Herring, A. H., and Dunson, D. B. (2021). Perturbed factor analysis: Accounting for group differences in exposure profiles. *The Annals of Applied Statistics*, 15(3):1386 – 1404.

Schiavon, L., Canale, A., and Dunson, D. (2022). Generalized infinite factorization models. *Biometrika*, 109(3):817–835.

Tanay, A., Sharan, R., and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics (Oxford, England)*, 18 Suppl 1:S136–44.

Teh, Y. (2006). A hierarchical bayesian language model based on pitman-yor processes. pages 985–992.

Wang, W. and Stephens, M. (2021). Empirical bayes matrix factorization. *Journal of Machine Learning Research*, 22(120):1–40.

## Acknowledgements

The authors thank Sylvia Frühwirth-Schnatter and Bettina Grün for their helpful discussion and comments to an earlier version of this paper.

# Supplementary Material for: "Multi-subject factor regression biclustering"

## A Variational inference algorithms for multi-study factor regression biclustering

In this section, we provide the detailes of the CAVI and SVI algorithms outlined in the main paper.

### A.1 CAVI algorithm for MSFRB

**Algorithm A.1:** Details of the CAVI algorithm steps for the MSFRB model

1: Set initial variational parameters

**Repeat until converged**

2: Update $\xi_{is}^2$ for $i = 1, \ldots, p$ and $s = 1, \ldots, S$ from $q^*(\xi_{is}^2) = \mathcal{G}^{-1}(\hat{a}_s^\xi, \hat{b}_{is}^\xi)$, where

$$\hat{a}_s^\xi = a_\xi + \frac{N_s}{2},$$
$$\hat{b}_{is}^\xi = \frac{\hat{a}^{b_\xi}}{\hat{b}_{is}^{b_\xi}} + \frac{1}{2}\sum_{n=1}^{N_s} \mathbb{E}\left(y_{ins} - \boldsymbol{\lambda}_i \boldsymbol{f}_{ns}^\top - \boldsymbol{\lambda}_{is}\boldsymbol{\eta}_{ns}^\top - \boldsymbol{\beta}_i^X \boldsymbol{x}_{ns}\right)^2,$$

where

$$\mathbb{E}\left(y_{ins} - \boldsymbol{\lambda}_i \boldsymbol{f}_{ns}^\top - \boldsymbol{\lambda}_{is}\boldsymbol{\eta}_{ns}^\top - \boldsymbol{\beta}_i^X \boldsymbol{x}_{ns}\right)^2 = y_{ins}^2 - 2y_{ins}\left(\boldsymbol{\mu}_i^\lambda \left[\boldsymbol{\mu}_{ns}^f\right]^\top + \boldsymbol{\mu}_{is}^{\lambda^s}\left[\boldsymbol{\mu}_{ns}^\eta\right]^\top + \boldsymbol{\mu}_i^{\beta^X}\boldsymbol{x}_{ns}\right) +$$
$$2\left(\boldsymbol{\mu}_i^\lambda \left[\boldsymbol{\mu}_{ns}^f\right]^\top \boldsymbol{\mu}_{is}^{\lambda^s}\left[\boldsymbol{\mu}_{ns}^\eta\right]^\top + \boldsymbol{\mu}_i^\lambda \left[\boldsymbol{\mu}_{ns}^f\right]^\top \boldsymbol{\mu}_i^{\beta^X}\boldsymbol{x}_{ns} + \boldsymbol{\mu}_{is}^{\lambda^s}\left[\boldsymbol{\mu}_{ns}^\eta\right]^\top \boldsymbol{\mu}_i^{\beta^X}\boldsymbol{x}_{ns}\right) +$$
$$\boldsymbol{\mu}_i^{\beta^X}\boldsymbol{x}_{ns}\left[\boldsymbol{x}_{ns}\right]^\top\left[\boldsymbol{\mu}_i^{\beta^X}\right]^\top + \boldsymbol{\mu}_i^\lambda\left[\boldsymbol{\mu}_{ns}^f\right]^\top\boldsymbol{\mu}_{ns}^f\left[\boldsymbol{\mu}_i^\lambda\right]^\top + \boldsymbol{\mu}_{is}^{\lambda^s}\left[\boldsymbol{\mu}_{ns}^\eta\right]^\top\boldsymbol{\mu}_{ns}^\eta\left[\boldsymbol{\mu}_{is}^{\lambda^s}\right]^\top +$$
$$\boldsymbol{\mu}_i^\lambda\boldsymbol{\Sigma}_{ns}^f\left[\boldsymbol{\mu}_i^\lambda\right]^\top + \left[\boldsymbol{\mu}_{ns}^f\right]^\top\boldsymbol{\Sigma}_i^\lambda\boldsymbol{\mu}_{ns}^f + \boldsymbol{\mu}_{is}^{\lambda^s}\boldsymbol{\Sigma}_{ns}^\eta\left[\boldsymbol{\mu}_{is}^{\lambda^s}\right]^\top + \left[\boldsymbol{\mu}_{ns}^\eta\right]^\top\boldsymbol{\Sigma}_{is}^{\lambda^s}\boldsymbol{\mu}_{ns}^\eta + Tr\left(\boldsymbol{\Sigma}_i^\lambda\boldsymbol{\Sigma}_{ns}^f\right) + Tr\left(\boldsymbol{\Sigma}_{is}^{\lambda^s}\boldsymbol{\Sigma}_{ns}^\eta\right)$$

and the mean of $\mathbb{E}(\xi_{is}^{-2}) = \mathbb{E}(\frac{1}{\xi_{is}^2}) = \hat{a}_{is}^\xi/\hat{b}_{is}^\xi$.

3: Update $b_{\xi is}$ for $i = 1, \ldots, p$ and $s = 1, \ldots, S$ from $q^*(b_{\xi is}) = \mathcal{G}(\hat{a}^{b_\xi}, \hat{b}_{is}^{b_\xi})$, where

$$\hat{a}^{b_\xi} = a_g + a_\xi,$$
$$\hat{b}_{is}^{b_\xi} = b_{gis} + \mathbb{E}(\frac{1}{\xi_{is}^2}) = b_{gis} + \frac{\hat{a}_s^\xi}{\hat{b}_{is}^\xi},$$

where $\mathbb{E}(b_{\xi is}) = \mu_{is}^{b_\xi} = \hat{a}^{b_\xi}/\hat{b}_{is}^{b_\xi}$.

4: Update $\theta_h|\gamma_h$ for $h = 1, \ldots, H$ from

- $q^*(\theta_h|\gamma_h = 1) = \mathcal{G}^{-1}(\hat{a}_{\theta_h slab}, \hat{a}_{\theta_h slab})$, with

$$\hat{a}_{\theta_h slab} = a_\Lambda + \frac{p}{2},$$
$$\hat{b}_{\theta_h slab} = \frac{\hat{a}_{b_\theta}}{\hat{b}_{b_\theta}} + \frac{1}{2}\frac{\hat{a}_{\kappa_\Lambda}}{\hat{b}_{\kappa_\Lambda}}\tilde{S}_h^\Lambda,$$

- $q^*(\theta_h|\gamma_h = 0) = \mathcal{G}^{-1}(\hat{a}_{\theta_h spike}, \hat{a}_{\theta_h spike})$ , with

$$\hat{a}_{\theta_h spike} = a_0^\Lambda + \frac{p}{2},$$
$$\hat{b}_{\theta_h spike} = \frac{\hat{a}_{b_0^\Lambda}}{\hat{b}_{b_0^\Lambda}} + \frac{1}{2}\frac{\hat{a}_{\kappa_\Lambda}}{\hat{b}_{\kappa_\Lambda}}\tilde{S}_h^\Lambda,$$

where $\tilde{S}_h^\Lambda = \sum_{i=1}^p \left[\mathbb{E}\lambda_{ih}^2 \mathbb{E}\left(\frac{1}{\phi_{ih}}\right)\right] = \sum_{i=1}^p \left[\left(\left[\boldsymbol{\mu}_{ih}^\lambda\right]^2 + \boldsymbol{\Sigma}_{i[h,h]}^\lambda\right)\frac{\hat{a}_\phi}{\hat{b}_{\phi_{ih}}}\right]$, with $h$ and $[h,h]$ indicating the $h$th element of the mean vector and the $[h,h]$th element of the covariance matrix $\boldsymbol{\Sigma}_i^\Lambda$, correspondingly. Calculate weighted mixture of expectations for $h = 1, \ldots, H$

$$\mathbb{E}_{mix}\left(\frac{1}{\theta_h}\right) = r_h \frac{\hat{a}_{\theta_h slab}}{\hat{b}_{\theta_h slab}} + (1 - r_h)\frac{\hat{a}_{\theta_h spike}}{\hat{b}_{\theta_h spike}}.$$

5: Update responsibilities for $h = 1, \ldots, H$ from $q^*(\gamma_h) = Ber(r_h)$. First calculate

$$l_{h\,slab} = \mathbb{E}\log\tau_h + a_\Lambda \mathbb{E}\log b_\Lambda - \log\Gamma(a_\Lambda) - (a_\Lambda + \frac{p}{2} + 1)\mathbb{E}_{slab}\log\theta_h - \mathbb{E}_{slab}\left(\frac{1}{\theta_h}\right)\left(\mathbb{E}b_\Lambda + \frac{1}{2}\mathbb{E}\left(\frac{1}{\kappa_\Lambda}\right)\tilde{S}_h^\Lambda\right),$$

$$l_{h\,spike} = \mathbb{E}\log(1-\tau_h) + a_0^\Lambda \mathbb{E}\log b_0^\Lambda - \log\Gamma(a_0^\Lambda) - (a_0^\Lambda + \frac{p}{2} + 1)\mathbb{E}_{spike}\log\theta_h - \mathbb{E}_{spike}\left(\frac{1}{\theta_h}\right)\left(\mathbb{E}b_0^\Lambda + \frac{1}{2}\mathbb{E}\left(\frac{1}{\kappa_\Lambda}\right)\tilde{S}_h^\Lambda\right),$$

where

$$\mathbb{E}\log\tau_h = \psi(\hat{a}_{\tau_h}) - \psi(\hat{a}_{\tau_h} + \hat{b}_{\tau_h}),$$
$$\mathbb{E}\log(1-\tau_h) = \psi(\hat{b}_{\tau_h}) - \psi(\hat{a}_{\tau_h} + \hat{b}_{\tau_h}),$$
$$\mathbb{E}\log b_\Lambda = \psi(\hat{a}_{b_\theta}) - \ln\hat{b}_{b_\theta},$$
$$\mathbb{E}\log b_0^\Lambda = \psi(\hat{a}_{b_0^\Lambda}) - \ln\hat{b}_{b_0^\Lambda},$$
$$\mathbb{E}_{slab}\log\theta_h = \ln\hat{b}_{\theta_h slab} - \psi(\hat{a}_{\theta_h slab}),$$
$$\mathbb{E}_{spike}\log\theta_h = \ln\hat{b}_{\theta_h spike} - \psi(\hat{a}_{\theta_h spike}),$$
$$\mathbb{E}_{slab}\left(\frac{1}{\theta_h}\right) = \frac{\hat{a}_{\theta_h slab}}{\hat{b}_{\theta_h slab}},$$
$$\mathbb{E}_{spike}\left(\frac{1}{\theta_h}\right) = \frac{\hat{a}_{\theta_h spike}}{\hat{b}_{\theta_h spike}},$$
$$\mathbb{E}\left(\frac{1}{\kappa_\Lambda}\right) = \frac{\hat{a}_{\kappa_\Lambda}}{\hat{b}_{\kappa_\Lambda}},$$
$$\tilde{S}_h^\Lambda = \sum_{i=1}^p \left[\mathbb{E}\lambda_{ih}^2 \mathbb{E}\left(\frac{1}{\phi_{ih}}\right)\right] = \sum_{i=1}^p \left[\left(\left[\boldsymbol{\mu}_{ih}^\lambda\right]^2 + \boldsymbol{\Sigma}_{i[h,h]}^\lambda\right)\frac{\hat{a}_\phi}{\hat{b}_{\phi_{ih}}}\right].$$

Calculate the responsibility of slab $r_h$ as

$$r_h = \frac{1}{1 + \exp(l_{h\,spike} - l_{h\,slab})}.$$

6: Update $\tau_h$ for $h = 1, \ldots, H$ from $q^*(\tau_h) = \mathcal{B}(\hat{a}_{\tau_h}, \hat{b}_{\tau_h})$, where

$$\hat{a}_{\tau_h} = \frac{\hat{a}_\alpha^\Lambda}{\hat{b}_\alpha^\Lambda H} + r_h,$$
$$\hat{b}_{\tau_h} = 2 - r_h.$$

7: Update $\alpha_\Lambda$ from $q^*(\alpha_\Lambda) = \mathcal{G}(\hat{a}_\alpha^\Lambda, \hat{b}_\alpha^\Lambda)$, where

$$\hat{a}_\alpha^\Lambda = a_\alpha^\Lambda + H,$$
$$\hat{b}_\alpha^\Lambda = b_\alpha^\Lambda - \frac{1}{H}\sum_{h=1}^{H} \mathbb{E}\log\tau_h,$$

with $\mathbb{E}\log\tau_h = \psi(\hat{a}_{\tau_h}) - \psi(\hat{a}_{\tau_h} + \hat{b}_{\tau_h})$.

8: Update $\phi_{ih}$ for $i = 1, \ldots, p$ and $h = 1, \ldots, H$ from $q^*(\phi_{ih}) = \mathcal{G}^{-1}(\hat{a}_\phi, \hat{b}_{\phi_{ih}})$, where

$$\hat{a}_\phi = a_\phi + \frac{1}{2},$$
$$\hat{b}_{\phi_{ih}} = \frac{\hat{a}_{\beta_\phi}}{\hat{b}_{\beta_{i\phi}}} + \frac{1}{2}\frac{\hat{a}_{\kappa_\Lambda}}{\hat{b}_{\kappa_\Lambda}}\mathbb{E}_{mix}\left(\frac{1}{\theta_h}\right)\mathbb{E}\lambda_{ih}^2,$$

with $\mathbb{E}_{mix}\left(\frac{1}{\theta_h}\right) = r_h\frac{\hat{a}_{\theta_h slab}}{\hat{b}_{\theta_h slab}} + (1-r_h)\frac{\hat{a}_{\theta_h spike}}{\hat{b}_{\theta_h spike}}$ and $\mathbb{E}\lambda_{ih}^2 = \left[\boldsymbol{\mu}_{ih}^\lambda\right]^2 + \boldsymbol{\Sigma}_{i[h,h]}^\lambda$.

9: Update $\kappa_\Lambda$ from $q^*(\kappa_\Lambda) = \mathcal{G}^{-1}(\hat{a}_{\kappa_\Lambda}, \hat{b}_{\kappa_\Lambda})$, where

$$\hat{a}_{\kappa_\Lambda} = a^\kappa + \frac{1}{2}pH,$$
$$\hat{b}_{\kappa_\Lambda} = b^\kappa + \frac{1}{2}\sum_{i=1}^{H}\left[\mathbb{E}_{mix}\left(\frac{1}{\theta_h}\right)\tilde{S}_h^\Lambda\right],$$

with $\mathbb{E}_{mix}\left(\frac{1}{\theta_h}\right) = r_h\frac{\hat{a}_{\theta_h slab}}{\hat{b}_{\theta_h slab}} + (1-r_h)\frac{\hat{a}_{\theta_h spike}}{\hat{b}_{\theta_h spike}}$ and $\tilde{S}_h^\Lambda = \sum_{i=1}^{p}\left[\left(\left[\boldsymbol{\mu}_{ih}^\lambda\right]^2 + \boldsymbol{\Sigma}_{i[h,h]}^\lambda\right)\frac{\hat{a}_\phi}{\hat{b}_{\phi_{ih}}}\right]$.

10: Update $\lambda_i^\top$ for $i = 1, \ldots, p$ from $q^*(\lambda_i) = \mathcal{N}_H(\boldsymbol{\mu}_i^\lambda, \boldsymbol{\Sigma}_i^\lambda)$, where

$$\boldsymbol{\Sigma}_i^\lambda = \left(\boldsymbol{\Psi}_i^{-1} + \sum_{s=1}^{S}\frac{\hat{a}_s^\xi}{\hat{b}_{is}^\xi}\sum_{n=1}^{N_s}\left(\boldsymbol{\mu}_{ns}^f\left[\boldsymbol{\mu}_{ns}^f\right]^\top + \boldsymbol{\Sigma}_{ns}^f\right)\right)^{-1},$$
$$\boldsymbol{\mu}_i^\lambda = \boldsymbol{\Sigma}_i^\lambda\sum_{s=1}^{S}\frac{\hat{a}_s^\xi}{\hat{b}_{is}^\xi}\left(\boldsymbol{\mu}_{1s}^f, \ldots, \boldsymbol{\mu}_{N_s s}^f\right)\left(\boldsymbol{y}_{is}^\top - \left(\boldsymbol{\mu}_{1s}^\eta, \ldots, \boldsymbol{\mu}_{N_s s}^\eta\right)\left[\boldsymbol{\mu}_{is}^\lambda\right]^\top - \boldsymbol{X}_s^\top\left[\boldsymbol{\mu}_i^{\beta^X}\right]^\top\right),$$

and where $\boldsymbol{\Psi}_i^{-1} = diag\left(\frac{\hat{a}_{\kappa_\Lambda}}{\hat{b}_{\kappa_\Lambda}}\frac{\hat{a}_\phi}{\hat{b}_{\phi_{i1}}}\mathbb{E}_{mix}\left(\frac{1}{\theta_1}\right), \ldots, \frac{\hat{a}_{\kappa_\Lambda}}{\hat{b}_{\kappa_\Lambda}}\frac{\hat{a}_\phi}{\hat{b}_{\phi_{iH}}}\mathbb{E}_{mix}\left(\frac{1}{\theta_H}\right)\right)$ with $\mathbb{E}_{mix}\left(\frac{1}{\theta_h}\right) = r_h\frac{\hat{a}_{\theta_h slab}}{\hat{b}_{\theta_h slab}} + (1-r_h)\frac{\hat{a}_{\theta_h spike}}{\hat{b}_{\theta_h spike}}$.

11: Update $b_\Lambda$ from $q^*(b_\Lambda) = \mathcal{G}(\hat{a}_{b_\theta}, \hat{b}_{b_\theta})$, where

$$\hat{a}_{b_\theta} = a_1^\Lambda + a_\Lambda\sum_{h=1}^{H} r_h,$$
$$\hat{b}_{b_\theta} = b_1^\Lambda + \sum_{h=1}^{H} r_h\frac{\hat{a}_{\theta_h slab}}{\hat{b}_{\theta_h slab}}.$$

12: Update $b_0^\Lambda$ from $q^*(b_0^\Lambda) = \mathcal{G}(\hat{a}_{b_0^\Lambda}, \hat{b}_{b_0^\Lambda})$, where

$$\hat{a}_{b_0^\Lambda} = a_2^\Lambda + a_0^\Lambda \sum_{h=1}^{H}(1 - r_h),$$
$$\hat{b}_{b_0^\Lambda} = b_2^\Lambda + \sum_{h=1}^{H}(1 - r_h)\frac{\hat{a}_{\theta_h spike}}{\hat{b}_{\theta_h spike}}.$$

13: Update $\beta_{i\phi}$ for $i = 1, \ldots, p$ from $q^*(\beta_{i\phi}) = \mathcal{G}(\hat{a}_{\beta_\phi}, \hat{b}_{\beta_{i\phi}})$, where

$$\hat{a}_{\beta_\phi} = c + Ha_\phi,$$
$$\hat{b}_{\beta_{i\phi}} = d + \sum_{h=1}^{H}\frac{\hat{a}_\phi}{\hat{b}_{\phi_{ih}}}.$$

14: Update $\upsilon_h|\nu_h$ for $h = 1, \ldots, H$ from

- $q^*(\upsilon_h|\nu_h = 1) = \mathcal{G}^{-1}(\hat{a}_{\upsilon_h slab}, \hat{a}_{\upsilon_h slab})$, with

$$\hat{a}_{\upsilon_h slab} = a_F + \frac{N}{2},$$
$$\hat{b}_{\upsilon_h slab} = \frac{\hat{a}_{b_\upsilon}}{\hat{b}_{b_\upsilon}} + \frac{1}{2}\frac{\hat{a}_{\kappa_F}}{\hat{b}_{\kappa_F}}\tilde{S}_h^F,$$

- $q^*(\upsilon_h|\nu_h = 0) = \mathcal{G}^{-1}(\hat{a}_{\upsilon_h spike}, \hat{a}_{\upsilon_h spike})$ , with

$$\hat{a}_{\upsilon_h spike} = a_0^F + \frac{N}{2},$$
$$\hat{b}_{\upsilon_h spike} = \frac{\hat{a}_{b_0^F}}{\hat{b}_{b_0^F}} + \frac{1}{2}\frac{\hat{a}_{\kappa_F}}{\hat{b}_{\kappa_F}}\tilde{S}_h^F,$$

where $\tilde{S}_h^F = \sum_{s=1}^{S}\sum_{n=1}^{N_s}\left[\mathbb{E}f_{hns}^2\mathbb{E}\left(\frac{1}{\zeta_{hns}}\right)\right] = \sum_{s=1}^{S}\sum_{n=1}^{N_s}\left[\left(\left[\mu_{hns}^f\right]^2 + \mathbf{\Sigma}_{ns[h,h]}^f\right)\frac{\hat{a}_\zeta}{\hat{b}_{\zeta_{hns}}}\right]$, with $h$ and $[h, h]$ indicating the $h$th element of the mean vector and the $[h, h]$th element of the covariance matrix $\mathbf{\Sigma}_{ns}^f$, correspondingly.

Calculate weighted mixture of expectations for $h = 1, \ldots, H$

$$\mathbb{E}_{mix}\left(\frac{1}{\upsilon_h}\right) = z_h\frac{\hat{a}_{\upsilon_h slab}}{\hat{b}_{\upsilon_h slab}} + (1 - z_h)\frac{\hat{a}_{\upsilon_h spike}}{\hat{b}_{\upsilon_h spike}}.$$

15: Update responsibilities for $h = 1, \ldots, H$ from $q^*(\nu_h) = Ber(z_h)$. First calculate

$$m_{h\,slab} = \mathbb{E}\log\pi_h + a_F\mathbb{E}\log b_F - \log\Gamma(a_F) - (a_F + \frac{N}{2} + 1)\mathbb{E}_{slab}\log\upsilon_h - \mathbb{E}_{slab}\left(\frac{1}{\upsilon_h}\right)\left(\mathbb{E}b_F + \frac{1}{2}\mathbb{E}\left(\frac{1}{\kappa_F}\right)\tilde{S}_h^F\right),$$

$$m_{h\,spike} = \mathbb{E}\log(1 - \pi_h) + a_0^F\mathbb{E}\log b_0^F - \log\Gamma(a_0^F) - (a_0^F + \frac{N}{2} + 1)\mathbb{E}_{spike}\log\upsilon_h - \mathbb{E}_{spike}\left(\frac{1}{\upsilon_h}\right)\left(\mathbb{E}b_0^F + \frac{1}{2}\mathbb{E}\left(\frac{1}{\kappa_F}\right)\tilde{S}_h^F\right),$$

where

$$\mathbb{E}\log\pi_h = \psi(\hat{a}_{\pi_h}) - \psi(\hat{a}_{\pi_h} + \hat{b}_{\pi_h}),$$
$$\mathbb{E}\log(1-\pi_h) = \psi(\hat{b}_{\pi_h}) - \psi(\hat{a}_{\pi_h} + \hat{b}_{\pi_h}),$$
$$\mathbb{E}\log b_F = \psi(\hat{a}_{b_\upsilon}) - \ln\hat{b}_{b_\upsilon},$$
$$\mathbb{E}\log b_0^F = \psi(\hat{a}_{b_0^F}) - \ln\hat{b}_{b_0^F},$$
$$\mathbb{E}_{slab}\log\upsilon_h = \ln\hat{b}_{\upsilon_h slab} - \psi(\hat{a}_{\upsilon_h slab}),$$
$$\mathbb{E}_{spike}\log\upsilon_h = \ln\hat{b}_{\upsilon_h spike} - \psi(\hat{a}_{\upsilon_h spike}),$$
$$\mathbb{E}_{slab}\left(\frac{1}{\upsilon_h}\right) = \frac{\hat{a}_{\upsilon_h slab}}{\hat{b}_{\upsilon_h slab}},$$
$$\mathbb{E}_{spike}\left(\frac{1}{\upsilon_h}\right) = \frac{\hat{a}_{\upsilon_h spike}}{\hat{b}_{\upsilon_h spike}},$$
$$\mathbb{E}\left(\frac{1}{\kappa_F}\right) = \frac{\hat{a}_{\kappa_F}}{\hat{b}_{\kappa_F}},$$
$$\tilde{S}_h^F = \sum_{s=1}^{S}\sum_{n=1}^{N_s}\left[\mathbb{E}f_{hns}^2\mathbb{E}\left(\frac{1}{\zeta_{hns}}\right)\right] = \sum_{s=1}^{S}\sum_{n=1}^{N_s}\left[\left(\left[\mu_{hns}^f\right]^2 + \mathbf{\Sigma}_{ns[h,h]}^f\right)\frac{\hat{a}_\zeta}{\hat{b}_{\zeta_{hns}}}\right].$$

Calculate the responsibility of slab $z_h$ as

$$z_h = \frac{1}{1+\exp(m_{h\,spike} - m_{h\,slab})}.$$

16: Update $\pi_h$ for $h = 1,\ldots,H$ from $q^*(\pi_h) = \mathcal{B}(\hat{a}_{\pi_h}, \hat{b}_{\pi_h})$, where

$$\hat{a}_{\pi_h} = \frac{\hat{a}_\alpha^F}{\hat{b}_\alpha^F H} + z_h,$$
$$\hat{b}_{\tau_h} = 2 - z_h.$$

17: Update $\alpha_F$ from $q^*(\alpha_F) = \mathcal{G}(\hat{a}_\alpha^F, \hat{b}_\alpha^F)$, where

$$\hat{a}_\alpha^F = a_\alpha^F + H,$$
$$\hat{b}_\alpha^F = b_\alpha^F - \frac{1}{H}\sum_{h=1}^{H}\mathbb{E}\log\pi_h,$$

with $\mathbb{E}\log\pi_h = \psi(\hat{a}_{\pi_h}) - \psi(\hat{a}_{\pi_h} + \hat{b}_{\pi_h})$.

18: Update $\zeta_{hns}$ for $s = 1,\ldots,S$, $n = 1,\ldots,N_s$ and $h = 1,\ldots,H$ from $q^*(\zeta_{hns}) = \mathcal{G}^{-1}(\hat{a}_\zeta, \hat{b}_{\zeta_{hns}})$, where

$$\hat{a}_\zeta = a_\zeta + \frac{1}{2},$$
$$\hat{b}_{\zeta_{hns}} = \frac{\hat{a}_{\beta_\zeta}}{\hat{b}_{\beta_{ns\zeta}}} + \frac{1}{2}\frac{\hat{a}_{\kappa_F}}{\hat{b}_{\kappa_F}}\mathbb{E}_{mix}\left(\frac{1}{\upsilon_h}\right)\mathbb{E}f_{hns}^2,$$

with $\mathbb{E}_{mix}\left(\frac{1}{\upsilon_h}\right) = z_h\frac{\hat{a}_{\upsilon_h slab}}{\hat{b}_{\upsilon_h slab}} + (1-z_h)\frac{\hat{a}_{\upsilon_h spike}}{\hat{b}_{\upsilon_h spike}}$ and $\mathbb{E}f_{hns}^2 = \left[\mu_{hns}^f\right]^2 + \mathbf{\Sigma}_{[h,h]ns}^f$.

19: Update $\kappa_F$ from $q^*(\kappa_F) = \mathcal{G}^{-1}(\hat{a}_{\kappa_F}, \hat{b}_{\kappa_F})$, where

$$\hat{a}_{\kappa_F} = c^{\kappa} + \frac{1}{2}HN,$$
$$\hat{b}_{\kappa_F} = d^{\kappa} + \frac{1}{2}\sum_{i=1}^{H}\left[\mathbb{E}_{mix}\left(\frac{1}{v_h}\right)\tilde{S}_h^F\right],$$

with $\mathbb{E}_{mix}\left(\frac{1}{v_h}\right) = z_h\frac{\hat{a}_{v_h slab}}{\hat{b}_{v_h slab}} + (1-z_h)\frac{\hat{a}_{v_h spike}}{\hat{b}_{v_h spike}}$ and $\tilde{S}_h^F = \sum_{s=1}^{S}\sum_{n=1}^{N_s}\left[\left(\left[\mu_{hns}^f\right]^2 + \mathbf{\Sigma}_{ns[h,h]}^f\right)\frac{\hat{a}_\zeta}{\hat{b}_{\zeta_{hns}}}\right]$.

20: Update $\mathbf{f}_{ns}$ for $n = 1, \ldots, N_s$ and $s = 1, \ldots, S$ from $q^*(\mathbf{f}_{ns}) = \mathcal{N}_H(\boldsymbol{\mu}_{ns}^f, \mathbf{\Sigma}_{ns}^f)$, where

$$\mathbf{\Sigma}_{ns}^f = \left(\mathbf{\Omega}_{ns}^{-1} + \sum_{i=1}^{p}\frac{\hat{a}_s^\xi}{\hat{b}_{is}^\xi}\left(\boldsymbol{\mu}_i^\lambda\left[\boldsymbol{\mu}_i^\lambda\right]^\top + \mathbf{\Sigma}_i^\lambda\right)\right)^{-1},$$
$$\boldsymbol{\mu}_{ns}^f = \mathbf{\Sigma}_{ns}^f\left(\boldsymbol{\mu}_1^\lambda, \ldots, \boldsymbol{\mu}_p^\lambda\right)^\top diag\left(\frac{\hat{a}_s^\xi}{\hat{b}_{1s}^\xi}, \ldots, \frac{\hat{a}_s^\xi}{\hat{b}_{ps}^\xi}\right)\left(\boldsymbol{y}_{ns} - \left(\boldsymbol{\mu}_{s1}^\lambda, \ldots, \boldsymbol{\mu}_{sP}^\lambda\right)\boldsymbol{\mu}_{ns}^\eta - \left(\boldsymbol{\mu}_1^{\beta^X}, \ldots, \boldsymbol{\mu}_P^{\beta^X}\right)\boldsymbol{x}_{ns}\right),$$

and where $\mathbf{\Omega}_{ns}^{-1} = diag\left(\frac{\hat{a}_{\kappa_F}}{\hat{b}_{\kappa_F}}\frac{\hat{a}_\zeta}{\hat{b}_{\zeta_{1ns}}}\mathbb{E}_{mix}\left(\frac{1}{v_1}\right), \ldots, \frac{\hat{a}_{\kappa_F}}{\hat{b}_{\kappa_F}}\frac{\hat{a}_\zeta}{\hat{b}_{\phi_{Hns}}}\mathbb{E}_{mix}\left(\frac{1}{v_H}\right)\right)$ with $\mathbb{E}_{mix}\left(\frac{1}{v_h}\right) = z_h\frac{\hat{a}_{v_h slab}}{\hat{b}_{v_h slab}} + (1-z_h)\frac{\hat{a}_{v_h spike}}{\hat{b}_{v_h spike}}$.

21: Update $b_F$ from $q^*(b_F) = \mathcal{G}(\hat{a}_{b_v}, \hat{b}_{b_v})$, where

$$\hat{a}_{b_v} = a_1^F + a_\Lambda\sum_{h=1}^{H}z_h,$$
$$\hat{b}_{b_v} = b_1^F + \sum_{h=1}^{H}z_h\frac{\hat{a}_{v_h slab}}{\hat{b}_{v_h slab}}.$$

22: Update $b_0^F$ from $q^*(b_0^F) = \mathcal{G}(\hat{a}_{b_0^F}, \hat{b}_{b_0^F})$, where

$$\hat{a}_{b_0^F} = a_2^F + a_0^F\sum_{h=1}^{H}(1-z_h),$$
$$\hat{b}_{b_0^F} = b_2^F + \sum_{h=1}^{H}(1-z_h)\frac{\hat{a}_{v_h spike}}{\hat{b}_{v_h spike}}.$$

23: Update $\beta_{ns\zeta}$ for $n = 1, \ldots, N_s$ and $s = 1, \ldots, S$ from $q^*(\beta_{ns\zeta}) = \mathcal{G}(\hat{a}_{\beta_\zeta}, \hat{b}_{\beta_{ns\zeta}})$, where

$$\hat{a}_{\beta_\zeta} = e + Ha_\zeta,$$
$$\hat{b}_{\beta_{ns\zeta}} = w + \sum_{h=1}^{H}\frac{\hat{a}_\zeta}{\hat{b}_{\zeta_{hns}}}.$$

24: Update $\theta_{hs}|\gamma_{hs}$ for $h = 1, \ldots, H_s$ and $s = 1, \ldots, S$ from

- $q^*(\theta_{hs}|\gamma_{hs} = 1) = \mathcal{G}^{-1}(\hat{a}_{\theta_{hs}slab}, \hat{a}_{\theta_{hs}slab})$, with

$$\hat{a}_{\theta_{hs}slab} = a_\Lambda^S + \frac{p}{2},$$
$$\hat{b}_{\theta_{hs}slab} = \frac{\hat{a}_{b_\Lambda^S}}{\hat{b}_{b_\Lambda^S}} + \frac{1}{2}\tilde{S}_{hs}^\Lambda,$$

- $q^*(\theta_{hs}|\gamma_{hs}=0)=\mathcal{G}^{-1}(\hat{a}_{\theta_{hs}spike},\hat{a}_{\theta_{hs}spike})$, with

$$\hat{a}_{\theta_{hs}spike}=a_0^{S\Lambda}+\frac{p}{2},$$
$$\hat{b}_{\theta_{hs}spike}=\frac{\hat{a}_{b_0^{S\Lambda}}}{\hat{b}_{b_0^{S\Lambda}}}+\frac{1}{2}\tilde{S}_{hs}^{\Lambda},$$

where $\tilde{S}_{hs}^{\Lambda}=\sum_{i=1}^{p}\left[\mathbb{E}\lambda_{ihs}^2\mathbb{E}\left(\frac{1}{\phi_{ihs}}\right)\right]=\sum_{i=1}^{p}\left[\left(\left[\boldsymbol{\mu}_{ihs}^{\lambda}\right]^2+\boldsymbol{\Sigma}_{is[h,h]}^{\lambda}\right)\frac{\hat{a}_{\phi s}}{\hat{b}_{\phi_{ihs}}}\right]$, with $h$ and $[h,h]$ indicating the $h$th element of the mean vector and the $[h,h]$th element of the covariance matrix $\boldsymbol{\Sigma}_{is}^{\Lambda}$, correspondingly.

Calculate weighted mixture of expectations for $h=1,\ldots,H_s$ and $s=1,\ldots,S$

$$\mathbb{E}_{mix}\left(\frac{1}{\theta_{hs}}\right)=r_{hs}\frac{\hat{a}_{\theta_{hs}slab}}{\hat{b}_{\theta_{hs}slab}}+(1-r_{hs})\frac{\hat{a}_{\theta_{hs}spike}}{\hat{b}_{\theta_{hs}spike}}.$$

25: Update responsibilities for $h=1,\ldots,H_s$ and $s=1,\ldots,S$ from $q^*(\gamma_{hs})=Ber(r_{hs})$. First calculate

$$l_{hs\,slab}=\mathbb{E}\log\tau_{hs}+a_{\Lambda}^{S}\mathbb{E}\log b_{\Lambda}^{S}-\log\Gamma(a_{\Lambda}^{S})-(a_{\Lambda}^{S}+\frac{p}{2}+1)\mathbb{E}_{slab}\log\theta_{hs}-\mathbb{E}_{slab}\left(\frac{1}{\theta_{hs}}\right)\left(\mathbb{E}b_{\Lambda}^{S}+\frac{1}{2}\tilde{S}_{hs}^{\Lambda}\right),$$

$$l_{hs\,spike}=\mathbb{E}\log(1-\tau_{hs})+a_0^{S\Lambda}\mathbb{E}\log b_0^{S\Lambda}-\log\Gamma(a_0^{S\Lambda})-(a_0^{S\Lambda}+\frac{p}{2}+1)\mathbb{E}_{spike}\log\theta_{hs}-\mathbb{E}_{spike}\left(\frac{1}{\theta_{hs}}\right)\left(\mathbb{E}b_0^{S\Lambda}+\frac{1}{2}\tilde{S}_{hs}^{\Lambda}\right),$$

where

$$\mathbb{E}\log\tau_{hs}=\psi(\hat{a}_{\tau_{hs}})-\psi(\hat{a}_{\tau_{hs}}+\hat{b}_{\tau_{hs}}),$$
$$\mathbb{E}\log(1-\tau_{hs})=\psi(\hat{b}_{\tau_{hs}})-\psi(\hat{a}_{\tau_{hs}}+\hat{b}_{\tau_{hs}}),$$
$$\mathbb{E}\log b_{\Lambda}^{S}=\psi(\hat{a}_{b_{\Lambda}^{S}})-\ln\hat{b}_{b_{\Lambda}^{S}},$$
$$\mathbb{E}\log b_0^{S\Lambda}=\psi(\hat{a}_{b_0^{S\Lambda}})-\ln\hat{b}_{b_0^{S\Lambda}},$$
$$\mathbb{E}_{slab}\log\theta_{hs}=\ln\hat{b}_{\theta_{hs}slab}-\psi(\hat{a}_{\theta_{hs}slab}),$$
$$\mathbb{E}_{spike}\log\theta_{hs}=\ln\hat{b}_{\theta_{hs}spike}-\psi(\hat{a}_{\theta_{hs}spike}),$$
$$\mathbb{E}_{slab}\left(\frac{1}{\theta_{hs}}\right)=\frac{\hat{a}_{\theta_{sh}slab}}{\hat{b}_{\theta_{hs}slab}},$$
$$\mathbb{E}_{spike}\left(\frac{1}{\theta_{hs}}\right)=\frac{\hat{a}_{\theta_{hs}spike}}{\hat{b}_{\theta_{hs}spike}},$$
$$\tilde{S}_{hs}^{\Lambda}=\sum_{i=1}^{p}\left[\mathbb{E}\lambda_{ihs}^2\mathbb{E}\left(\frac{1}{\phi_{ihs}}\right)\right]=\sum_{i=1}^{p}\left[\left(\left[\boldsymbol{\mu}_{ihs}^{\lambda}\right]^2+\boldsymbol{\Sigma}_{is[h,h]}^{\lambda}\right)\frac{\hat{a}_{\phi s}}{\hat{b}_{\phi_{ihs}}}\right].$$

Calculate the responsibility of slab $r_{hs}$ as

$$r_{hs}=\frac{1}{1+\exp(l_{hs\,spike}-l_{hs\,slab})}.$$

26: Update $\tau_{hs}$ for $h = 1, \ldots, H_s$ and $s = 1, \ldots, S$ from $q^*(\tau_{hs}) = \mathcal{B}(\hat{a}_{\tau_{hs}}, \hat{b}_{\tau_{hs}})$, where

$$\hat{a}_{\tau_{hs}} = \frac{\hat{a}_\alpha^{S\Lambda}}{\hat{b}_\alpha^{S\Lambda} H_s} + r_{hs},$$
$$\hat{b}_{\tau_{hs}} = 2 - r_{hs}.$$

27: Update $\alpha_\Lambda^S$ from $q^*(\alpha_\Lambda^S) = \mathcal{G}(\hat{a}_\alpha^{S\Lambda}, \hat{b}_\alpha^{S\Lambda})$, where

$$\hat{a}_\alpha^{S\Lambda} = a_\alpha^{S\Lambda} + H_s S,$$
$$\hat{b}_\alpha^{S\Lambda} = b_\alpha^{S\Lambda} - \frac{1}{H_s} \sum_{s=1}^{S} \sum_{h=1}^{H_s} \mathbb{E} \log \tau_{hs},$$

with $\mathbb{E} \log \tau_{hs} = \psi(\hat{a}_{\tau_{hs}}) - \psi(\hat{a}_{\tau_{hs}} + \hat{b}_{\tau_{hs}})$.

28: Update $\phi_{ihs}$ for $i = 1, \ldots, p$, $h = 1, \ldots, H$ and $s = 1, \ldots, S$ from $q^*(\phi_{ihs}) = \mathcal{G}^{-1}(\hat{a}_{\phi s}, \hat{b}_{\phi_{ihs}})$, where

$$\hat{a}_{\phi s} = a_\phi^S + \frac{1}{2},$$
$$\hat{b}_{\phi_{ihs}} = \frac{\hat{a}_{\beta_{\phi s}}}{\hat{b}_{\beta_{\phi is}}} + \frac{1}{2} \mathbb{E}_{mix} \left( \frac{1}{\theta_{hs}} \right) \mathbb{E} \lambda_{ihs}^2,$$

with $\mathbb{E}_{mix} \left( \frac{1}{\theta_{hs}} \right) = r_{hs} \frac{\hat{a}_{\theta_{hs} slab}}{\hat{b}_{\theta_{hs} slab}} + (1 - r_{hs}) \frac{\hat{a}_{\theta_{hs} spike}}{\hat{b}_{\theta_{hs} spike}}$ and $\mathbb{E} \lambda_{ihs}^2 = \left[ \boldsymbol{\mu}_{ihs}^\lambda \right]^2 + \boldsymbol{\Sigma}_{i[h,h]}^\lambda$.

29: Update $\lambda_{is}^\top$ for $i = 1, \ldots, p$ and $s = 1, \ldots, S$ from $q^*(\lambda_{is}) = \mathcal{N}_{H_s}(\boldsymbol{\mu}_{is}^\lambda, \boldsymbol{\Sigma}_{is}^\lambda)$, where

$$\boldsymbol{\Sigma}_{is}^\lambda = \left( \boldsymbol{\Psi}_{is}^{-1} + \frac{\hat{a}_s^\xi}{\hat{b}_{is}^\xi} \sum_{n=1}^{N_s} \left( \boldsymbol{\mu}_{ns}^\eta \left[ \boldsymbol{\mu}_{ns}^\eta \right]^\top + \boldsymbol{\Sigma}_{ns}^\eta \right) \right)^{-1},$$
$$\boldsymbol{\mu}_{is}^\lambda = \boldsymbol{\Sigma}_{is}^\lambda \frac{\hat{a}_s^\xi}{\hat{b}_{is}^\xi} \left( \boldsymbol{\mu}_{1s}^\eta, \ldots, \boldsymbol{\mu}_{N_s s}^\eta \right) \left( \boldsymbol{y}_{is}^\top - \left( \boldsymbol{\mu}_{1s}^f, \ldots, \boldsymbol{\mu}_{N_s s}^f \right) \left[ \boldsymbol{\mu}_i^\lambda \right]^\top - \boldsymbol{X}_s^\top \left[ \boldsymbol{\mu}_i^{\beta^X} \right]^\top \right),$$

and where $\boldsymbol{\Psi}_{is}^{-1} = diag \left( \frac{\hat{a}_{\phi s}}{\hat{b}_{\phi_{i1s}}} \mathbb{E}_{mix} \left( \frac{1}{\theta_{1s}} \right), \ldots, \frac{\hat{a}_{\phi s}}{\hat{b}_{\phi_{iH_s s}}} \mathbb{E}_{mix} \left( \frac{1}{\theta_{H_s s}} \right) \right)$ with $\mathbb{E}_{mix} \left( \frac{1}{\theta_{hs}} \right) = r_{hs} \frac{\hat{a}_{\theta_{hs} slab}}{\hat{b}_{\theta_{hs} slab}} + (1 - r_{hs}) \frac{\hat{a}_{\theta_{hs} spike}}{\hat{b}_{\theta_{hs} spike}}$.

30: Update $b_\Lambda^S$ from $q^*(b_\Lambda^S) = \mathcal{G}(\hat{a}_{b_\Lambda^S}, \hat{b}_{b_\Lambda^S})$, where

$$\hat{a}_{b_\Lambda^S} = a_1^{S\Lambda} + a_\Lambda^S \sum_{s=1}^{S} \sum_{h=1}^{H_s} r_{hs},$$
$$\hat{b}_{b_\Lambda^S} = b_1^{S\Lambda} + \sum_{s=1}^{S} \sum_{h=1}^{H_s} r_{hs} \frac{\hat{a}_{\theta_{hs} slab}}{\hat{b}_{\theta_{hs} slab}}.$$

31: Update $b_0^{S\Lambda}$ from $q^*(b_0^{S\Lambda}) = \mathcal{G}(\hat{a}_{b_0^{S\Lambda}}, \hat{b}_{b_0^{S\Lambda}})$, where

$$\hat{a}_{b_0^{S\Lambda}} = a_2^{S\Lambda} + a_0^{S\Lambda} \sum_{s=1}^{S} \sum_{h=1}^{H_s} (1 - r_{hs}),$$
$$\hat{b}_{b_0^{S\Lambda}} = b_2^{S\Lambda} + \sum_{s=1}^{S} \sum_{h=1}^{H_s} (1 - r_{hs}) \frac{\hat{a}_{\theta_{hs} spike}}{\hat{b}_{\theta_{hs} spike}}.$$

32: Update $\beta_{\phi is}$ for $i = 1, \ldots, p$ and $s = 1, \ldots, S$ from $q^*(\beta_{\phi_{is}}) = \mathcal{G}(\hat{a}_{\beta_{\phi s}}, \hat{b}_{\beta_{\phi is}})$, where

$$\hat{a}_{\beta_{\phi s}} = c^S + H_s a_\phi^S$$
$$\hat{b}_{\beta_{\phi is}} = d^S + \sum_{h=1}^{H_s} \frac{\hat{a}_{\phi s}}{\hat{b}_{\phi_{ihs}}}.$$

33: Update $\eta_{ns}$ for $n = 1, \ldots, N_s$ and $s = 1, \ldots, S$ from $q^*(\eta_{ns}) \sim \mathcal{N}_{H_s}(\boldsymbol{\mu}_{ns}^{\eta}, \boldsymbol{\Sigma}_{ns}^{\eta})$, where

$$\boldsymbol{\Sigma}_{ns}^{\eta} = \left( \boldsymbol{I}_{H_s} + \sum_{i=1}^{p} \frac{\hat{a}_s^{\xi}}{\hat{b}_{is}^{\xi}} \left( \boldsymbol{\mu}_{is}^{\lambda} \left[ \boldsymbol{\mu}_{is}^{\lambda} \right]^\top + \boldsymbol{\Sigma}_{is}^{\lambda} \right) \right)^{-1},$$
$$\boldsymbol{\mu}_{ns}^{\eta} = \boldsymbol{\Sigma}_{ns}^{f} \left( \boldsymbol{\mu}_{1s}^{\lambda}, \ldots, \boldsymbol{\mu}_{ps}^{\lambda} \right)^\top diag\left( \frac{\hat{a}_s^{\xi}}{\hat{b}_{1s}^{\xi}}, \ldots, \frac{\hat{a}_s^{\xi}}{\hat{b}_{ps}^{\xi}} \right) \left( \boldsymbol{y}_{ns} - \left( \boldsymbol{\mu}_1^{\lambda}, \ldots, \boldsymbol{\mu}_P^{\lambda} \right) \boldsymbol{\mu}_{ns}^{f} - \left( \boldsymbol{\mu}_1^{\beta^X}, \ldots, \boldsymbol{\mu}_P^{\beta^X} \right) \boldsymbol{x}_{ns} \right).$$

34: Update $\boldsymbol{\beta}_i^X$ for $i = 1, \ldots, p$ from $q^*(\beta_i^X) \sim \mathcal{N}_{p_x}(\boldsymbol{\mu}_i^{\beta^X}, \boldsymbol{\Sigma}_i^{\beta^X})$, where

$$\boldsymbol{\Sigma}_i^{\beta^X} = \left( \boldsymbol{Q}_i^{-1} + \sum_{s=1}^{S} \frac{\hat{a}_s^{\xi}}{\hat{b}_{is}^{\xi}} \boldsymbol{X}_s \left[ \boldsymbol{X}_s \right]^\top \right)^{-1},$$
$$\boldsymbol{\mu}_i^{\beta^X} = \boldsymbol{\Sigma}_i^{\beta^X} \left( \sum_{s=1}^{S} \frac{\hat{a}_s^{\xi}}{\hat{b}_{is}^{\xi}} \left( \boldsymbol{x}_{1s}, \ldots, \boldsymbol{x}_{N_s s} \right) \left( \boldsymbol{y}_{is}^\top - \left( \boldsymbol{\mu}_{1s}^{f}, \ldots, \boldsymbol{\mu}_{N_s s}^{f} \right) \left[ \boldsymbol{\mu}_i^{\lambda} \right]^\top - \left( \boldsymbol{\mu}_{1s}^{\eta}, \ldots, \boldsymbol{\mu}_{N_s s}^{\eta} \right) \left[ \boldsymbol{\mu}_{is}^{\lambda} \right]^\top \right) \right),$$

where $\boldsymbol{Q}_i = diag(q_{i1}, \ldots, q_{ip_x})$.

## A.2 SVI algorithm for MSFRB

The SVI algorithm requires some modifications compared to CAVI. First, the local parameters are updated only for a subset of the data. Second, the global parameters need to be reformulated in terms of natural parameters and updated as a weighted average of the value at the current and previous iterations. The variational updates for other parameters, namely, $b_{\xi is}$, $r_h$, $\tau_h$, $\theta_h$, $\alpha_\Lambda$, $\phi_{ih}$, $\kappa_\Lambda$, $b_\Lambda$, $b_0^\Lambda$, $\beta_{i\phi}$, $\pi_h$, $\alpha_F$, $b_F$, $b_0^F$, $r_{hs}$, $\tau_{hs}$, $\theta_{hs}$, $\alpha_\Lambda^S$, $b_\Lambda^S$ and $b_0^{S\Lambda}$, are performed like in Algorithm A.1 and the updated global variables are reparameterised into their original forms according to Table A.1.

**Algorithm A.2:** Details of the SVI algorithm steps for the MSFRB model

1: Set initial variational parameters
2: Compute iteration-independent variational parameters:
for $s = 1, \ldots, S$ update $\hat{a}_s^{\xi} = a_\xi + \frac{N_s}{2}$,
update $\hat{a}^{b_\xi} = a_g + a_\xi$,
update $\hat{a}_{\theta_h slab} = a_\Lambda + \frac{p}{2}$,
update $\hat{a}_{\theta_h spike} = a_0^\Lambda + \frac{p}{2}$,
update $\hat{a}_\phi = a_\phi + \frac{1}{2}$,
update $\hat{a}_{\kappa_\Lambda} = a^\kappa + \frac{1}{2} pH$,
update $\hat{a}_{\beta_\phi} = c + H a_\phi$,
update $\hat{a}_{\upsilon_h slab} = a_F + \frac{N}{2}$,
update $\hat{a}_{\upsilon_h spike} = a_0^F + \frac{N}{2}$,
update $\hat{a}_\zeta = a_\zeta + \frac{1}{2}$,
update $\hat{a}_{\kappa_F} = c^\kappa + \frac{1}{2} HN$,

| Variable | Natural parameter | Original parameter |
|---|---|---|
| $\xi_{is}^2$ | $\eta_{is}^{\xi} = -\hat{b}_{is}^{\xi}$ | $\hat{b}_{is}^{\xi} = -\eta_{is}^{\xi}$ |
| $\kappa_F$ | $\eta_{\kappa_F} = -\hat{b}_{\kappa_F}$ | $\hat{b}_{\kappa_F} = -\eta_{\kappa_F}$ |
| $\upsilon_h$ | $\eta_{h2}^{\upsilon} = -\hat{b}_{\upsilon_h}$ | $\hat{b}_{\upsilon_h} = -\eta_{h2}^{\upsilon}$ |
| $z_h$ | $\eta_{z_h} = \log \frac{z_h}{1-z_h}$ | $z_h = \frac{1}{1+\exp(-\eta_{z_h})}$ |
| $\boldsymbol{\lambda}_i$ | $\boldsymbol{\eta}_{i1}^{\lambda} = -\frac{1}{2}\left(\boldsymbol{\Sigma}_i^{\lambda}\right)^{-1}$<br>$\boldsymbol{\eta}_{i2}^{\lambda} = \left(\boldsymbol{\Sigma}_i^{\lambda}\right)^{-1}\boldsymbol{\mu}_i^{\lambda}$ | $\boldsymbol{\Sigma}_i^{\lambda} = -\frac{1}{2}\left(\boldsymbol{\eta}_{i1}^{\lambda}\right)^{-1}$<br>$\boldsymbol{\mu}_i^{\lambda} = -\frac{1}{2}\left(\boldsymbol{\eta}_{i1}^{\lambda}\right)^{-1}\boldsymbol{\eta}_{i2}^{\lambda}$ |
| $\boldsymbol{\lambda}_{is}$ | $\boldsymbol{\eta}_{is1}^{\lambda} = -\frac{1}{2}\left(\boldsymbol{\Sigma}_{is}^{\lambda}\right)^{-1}$<br>$\boldsymbol{\eta}_{is2}^{\lambda} = \left(\boldsymbol{\Sigma}_{is}^{\lambda}\right)^{-1}\boldsymbol{\mu}_{is}^{\lambda}$ | $\boldsymbol{\Sigma}_{is}^{\lambda} = -\frac{1}{2}\left(\boldsymbol{\eta}_{is1}^{\lambda}\right)^{-1}$<br>$\boldsymbol{\mu}_{is}^{\lambda} = -\frac{1}{2}\left(\boldsymbol{\eta}_{is1}^{\lambda}\right)^{-1}\boldsymbol{\eta}_{is2}^{\lambda}$ |
| $\boldsymbol{\beta}_i^X$ | $\boldsymbol{\eta}_{i1}^{\beta^X} = -\frac{1}{2}\left(\boldsymbol{\Sigma}_i^{\beta^X}\right)^{-1}$<br>$\boldsymbol{\eta}_{i2}^{\beta^X} = \left(\boldsymbol{\Sigma}_i^{\beta^X}\right)^{-1}\boldsymbol{\mu}_i^{\beta^X}$ | $\boldsymbol{\Sigma}_i^{\beta^X} = -\frac{1}{2}\left(\boldsymbol{\eta}_{i1}^{\beta^X}\right)^{-1}$<br>$\boldsymbol{\mu}_i^{\beta^X} = -\frac{1}{2}\left(\boldsymbol{\eta}_{i1}^{\beta^X}\right)^{-1}\boldsymbol{\eta}_{i2}^{\beta^X}$ |

Table A.1: *Reparametrisation formulas to convert natural parameters of the global variables into the conventional form and vice versa*

update $\hat{a}_{\beta_\zeta} = e + Ha_\zeta$,
update $\hat{a}_{\theta_{hs}slab} = a_\Lambda^S + \frac{p}{2}$,
update $\hat{a}_{\theta_{hs}spike} = a_0^{S\Lambda} + \frac{p}{2}$,
update $\hat{a}_{\phi s} + \frac{1}{2}$,
update $\hat{a}_{\beta_{\phi s}} = c^S + H_s a_\phi^S$,
update $\hat{a}_\alpha^\Lambda = a_\alpha^\Lambda + H$,
update $\hat{a}_\alpha^F = a_\alpha^F + H$,
update $\hat{a}_\alpha^{S\Lambda} = a_\alpha^{S\Lambda} + H_s S$.

**for** $t = 1, \ldots$ **until convergence is reached**

3: Set the step size $\rho_t = (t + \tau^\rho)^{-\kappa^\rho}$
4: For $s = 1, \ldots, S$ draw $\tilde{N}_s = \lfloor b_s \times N_s \rfloor$ data points from $1, \ldots, N_s$ without replacement and call it $\mathcal{I}_s(t)$. Set $\tilde{\boldsymbol{Y}}_s^t = \{\boldsymbol{y}_{ns} \in \mathcal{I}_s(t)\}$
5: For $s = 1, \ldots, S$ and $n \in \mathcal{I}_s(t)$ calculate the updates for the local variational parameters for each observation in $\tilde{\boldsymbol{Y}}_s^t$:
update $\hat{b}_{\zeta hns}$ for $h = 1, \ldots, H$ as in Step 18 of Algorithm A.1
update $\boldsymbol{\Sigma}_{ns}^f$ and $\boldsymbol{\mu}_{ns}^f$ as in Step 20 of Algorithm A.1
update $\hat{b}_{\beta_{ns\zeta}}$ as in Step 23 of Algorithm A.1
update $\boldsymbol{\Sigma}_{ns}^\eta$ and $\boldsymbol{\mu}_{ns}^\eta$ as in Step 33 of Algorithm A.1
6: Calculate the updates for global parameters using $\tilde{\boldsymbol{Y}}_s^t$ and the current step size $\rho_t$:
(a) For $s = 1, \ldots, S$ and $i = 1, \ldots, p$ update

$$\hat{\eta}_{is}^{\xi} = -\frac{\hat{a}^{b_\xi}}{\hat{b}_{is}^{b_\xi}} - \frac{1}{2}\frac{N_s}{\tilde{N}_s}\sum_{n\in\mathcal{I}_s(t)}\Big[y_{ins}^2 - 2y_{ins}\left(\boldsymbol{\mu}_i^\lambda\left[\boldsymbol{\mu}_{ns}^f\right]^\top + \boldsymbol{\mu}_{is}^{\lambda^s}\left[\boldsymbol{\mu}_{ns}^\eta\right]^\top + \boldsymbol{\mu}_i^{\beta^X}\boldsymbol{x}_{ns}\right) +$$
$$2\left(\boldsymbol{\mu}_i^\lambda\left[\boldsymbol{\mu}_{ns}^f\right]^\top\boldsymbol{\mu}_{is}^{\lambda^s}\left[\boldsymbol{\mu}_{ns}^\eta\right]^\top + \boldsymbol{\mu}_i^\lambda\left[\boldsymbol{\mu}_{ns}^f\right]^\top\boldsymbol{\mu}_i^{\beta^X}\boldsymbol{x}_{ns} + \boldsymbol{\mu}_{is}^{\lambda^s}\left[\boldsymbol{\mu}_{ns}^\eta\right]^\top\boldsymbol{\mu}_i^{\beta^X}\boldsymbol{x}_{ns}\right) +$$

$$\boldsymbol{\mu}_i^{\beta^X}\boldsymbol{x}_{ns}\left[\boldsymbol{x}_{ns}\right]^\top\left[\boldsymbol{\mu}_i^{\beta^X}\right]^\top+\boldsymbol{\mu}_i^{\lambda}\left[\boldsymbol{\mu}_{ns}^{f}\right]^\top\boldsymbol{\mu}_{ns}^{f}\left[\boldsymbol{\mu}_i^{\lambda}\right]^\top+\boldsymbol{\mu}_{is}^{\lambda^s}\left[\boldsymbol{\mu}_{ns}^{\eta}\right]^\top\boldsymbol{\mu}_{ns}^{\eta}\left[\boldsymbol{\mu}_{is}^{\lambda^s}\right]^\top+$$
$$\boldsymbol{\mu}_i^{\lambda}\boldsymbol{\Sigma}_{ns}^{f}\left[\boldsymbol{\mu}_i^{\lambda}\right]^\top+\left[\boldsymbol{\mu}_{ns}^{f}\right]^\top\boldsymbol{\Sigma}_i^{\lambda}\boldsymbol{\mu}_{ns}^{f}+\boldsymbol{\mu}_{is}^{\lambda^s}\boldsymbol{\Sigma}_{ns}^{\eta}\left[\boldsymbol{\mu}_{is}^{\lambda^s}\right]^\top+\left[\boldsymbol{\mu}_{ns}^{\eta}\right]^\top\boldsymbol{\Sigma}_{is}^{\lambda^s}\boldsymbol{\mu}_{ns}^{\eta}+Tr\left(\boldsymbol{\Sigma}_i^{\lambda}\boldsymbol{\Sigma}_{ns}^{f}\right)+Tr\left(\boldsymbol{\Sigma}_{is}^{\lambda^s}\boldsymbol{\Sigma}_{ns}^{\eta}\right)\Big]$$

and set $\eta_{is}^{\xi}$ as the weighted average of the values at this and the previous iterations

$$\eta_{is}^{\xi}(t)=(1-\rho_t)\eta_{is}^{\xi}(t-1)+\rho_t\hat{\eta}_{is}^{\xi}$$

(b) For $i=1,\ldots,p$ update

$$\hat{\eta}_{i1}^{\lambda}=-\frac{1}{2}\left(\boldsymbol{\Psi}_i^{-1}+\sum_{s=1}^{S}\frac{\hat{a}_s^{\xi}}{\hat{b}_{is}^{\xi}}\frac{N_s}{\tilde{N}_s}\sum_{n\in\mathcal{I}_s(t)}\left(\boldsymbol{\mu}_{ns}^{f}\left[\boldsymbol{\mu}_{ns}^{f}\right]^\top+\boldsymbol{\Sigma}_{ns}^{f}\right)\right),$$

$$\hat{\eta}_{i2}^{\lambda}=\sum_{s=1}^{S}\frac{\hat{a}_s^{\xi}}{\hat{b}_{is}^{\xi}}\frac{N_s}{\tilde{N}_s}\sum_{n\in\mathcal{I}_s(t)}\left(\boldsymbol{\mu}_{ns}^{f}\left(\boldsymbol{y}_{ns}^\top-\boldsymbol{\mu}_{ns}^{\eta}\left[\boldsymbol{\mu}_{is}^{\lambda}\right]^\top-\boldsymbol{x}_{ns}^\top\left[\boldsymbol{\mu}_i^{\beta^X}\right]^\top\right)\right),$$

where $\boldsymbol{\Psi}_i^{-1}$ is as defined in Step 10 of Algorithm A.1.
Set $\eta_{i1}^{\lambda}$ and $\eta_{i2}^{\lambda}$ as the weighted average of the values at this and the previous iterations

$$\eta_{i1}^{\lambda}(t)=(1-\rho_t)\eta_{i1}^{\lambda}(t-1)+\rho_t\hat{\eta}_{i1}^{\lambda},$$
$$\eta_{i2}^{\lambda}(t)=(1-\rho_t)\eta_{i2}^{\lambda}(t-1)+\rho_t\hat{\eta}_{i2}^{\lambda}.$$

(c) For $i=1,\ldots,p$ and $s=1,\ldots,S$ update

$$\hat{\eta}_{is1}^{\lambda}=-\frac{1}{2}\left(\boldsymbol{\Psi}_{is}^{-1}+\frac{\hat{a}_s^{\xi}}{\hat{b}_{is}^{\xi}}\frac{N_s}{\tilde{N}_s}\sum_{n\in\mathcal{I}_s(t)}\left(\boldsymbol{\mu}_{ns}^{\eta}\left[\boldsymbol{\mu}_{ns}^{\eta}\right]^\top+\boldsymbol{\Sigma}_{ns}^{\eta}\right)\right),$$

$$\hat{\eta}_{is2}^{\lambda}=\frac{\hat{a}_s^{\xi}}{\hat{b}_{is}^{\xi}}\frac{N_s}{\tilde{N}_s}\sum_{n\in\mathcal{I}_s(t)}\left(\boldsymbol{\mu}_{ns}^{\eta}\left(\boldsymbol{y}_{ns}^\top-\boldsymbol{\mu}_{ns}^{f}\left[\boldsymbol{\mu}_i^{\lambda}\right]^\top-\boldsymbol{x}_{ns}^\top\left[\boldsymbol{\mu}_i^{\beta^X}\right]^\top\right)\right),$$

where $\boldsymbol{\Psi}_{is}^{-1}$ is as defined in Step 29 of Algorithm A.1.
Set $\eta_{i1}^{\lambda}$ and $\eta_{i2}^{\lambda}$ as the weighted average of the values at this and the previous iterations

$$\eta_{is1}^{\lambda}(t)=(1-\rho_t)\eta_{is1}^{\lambda}(t-1)+\rho_t\hat{\eta}_{is1}^{\lambda},$$
$$\eta_{is2}^{\lambda}(t)=(1-\rho_t)\eta_{is2}^{\lambda}(t-1)+\rho_t\hat{\eta}_{is2}^{\lambda}.$$

(d) For $i=1,\ldots,p$ update

$$\hat{\eta}_{i1}^{\beta^X}=-\frac{1}{2}\left(\boldsymbol{Q}_i^{-1}+\sum_{s=1}^{S}\frac{\hat{a}_s^{\xi}}{\hat{b}_{is}^{\xi}}\frac{N_s}{\tilde{N}_s}\sum_{n\in\mathcal{I}_s(t)}\boldsymbol{x}_{ns}\left[\boldsymbol{x}_{ns}\right]^\top\right),$$

$$\hat{\eta}_{i2}^{\beta^X}=\sum_{s=1}^{S}\frac{\hat{a}_s^{\xi}}{\hat{b}_{is}^{\xi}}\frac{N_s}{\tilde{N}_s}\sum_{n\in\mathcal{I}_s(t)}\left(\boldsymbol{x}_{ns}\left(\boldsymbol{y}_{ns}^\top-\boldsymbol{\mu}_{ns}^{f}\left[\boldsymbol{\mu}_i^{\lambda}\right]^\top-\boldsymbol{\mu}_{ns}^{\eta}\left[\boldsymbol{\mu}_{is}^{\lambda}\right]^\top\right)\right),$$

where $\boldsymbol{Q}_i=\mathrm{diag}(q_{i1},\ldots,q_{ip_x})$.

Set $\eta_{i1}^{\beta^X}$ and $\eta_{i2}^{\beta^X}$ as the weighted average of the values at this and the previous iterations

$$\eta_{i1}^{\beta^X}(t) = (1-\rho_t)\eta_{i1}^{\beta^X}(t-1) + \rho_t\hat{\eta}_{i1}^{\beta^X},$$
$$\eta_{i2}^{\beta^X}(t) = (1-\rho_t)\eta_{i2}^{\beta^X}(t-1) + \rho_t\hat{\eta}_{i2}^{\beta^X}.$$

(e) For $h = 1, \ldots, H$ update

$$\hat{\eta}_{\upsilon_{hslab}} = -\frac{\hat{a}_{b_\upsilon}}{\hat{b}_{b_\upsilon}} - \frac{1}{2}\frac{\hat{a}_{\kappa_F}}{\hat{b}_{\kappa_F}}\breve{S}_h^F,$$

$$\hat{\eta}_{\upsilon_{hspike}} = -\frac{\hat{a}_{b_0^F}}{\hat{b}_{b_0^F}} - \frac{1}{2}\frac{\hat{a}_{\kappa_F}}{\hat{b}_{\kappa_F}}\breve{S}_h^F,$$

where $\breve{S}_h^F = \sum_{s=1}^S \frac{N_s}{\bar{N}_s}\sum_{n\in\mathcal{I}_s(t)}\left[\left(\left[\mu_{hns}^f\right]^2 + \mathbf{\Sigma}_{ns[h,h]}^f\right)\frac{\hat{a}_\zeta}{\hat{b}_{\zeta_{hns}}}\right]$, with $h$ and $[h,h]$ indicating the $h$th element of the mean vector and the $[h,h]$th element of the covariance matrix $\mathbf{\Sigma}_{ns}^f$, correspondingly.
Set $\eta_{\upsilon_{hslab}}$ and $\eta_{\upsilon_{hspike}}$ as the weighted average of the values at this and the previous iterations

$$\eta_{\upsilon_{hslab}}(t) = (1-\rho_t)\eta_{\upsilon_{hslab}}(t-1) + \rho_t\hat{\eta}_{\upsilon_{hslab}},$$
$$\eta_{\upsilon_{hspike}}(t) = (1-\rho_t)\eta_{\upsilon_{hspike}}(t-1) + \rho_t\hat{\eta}_{\upsilon_{hspike}}.$$

(f) Update

$$\hat{\eta}_{\kappa_F} = -d^\kappa - \frac{1}{2}\sum_{h=1}^H\left[\mathbb{E}_{mix}\left(\frac{1}{\upsilon_h}\right)\breve{S}_h^F\right],$$

where $\mathbb{E}_{mix}\left(\frac{1}{\upsilon_h}\right)$ is as defined in Step 19 of Algorithm A.1 and $\breve{S}_h^F = \sum_{s=1}^S \frac{N_s}{\bar{N}_s}\sum_{n\in\mathcal{I}_s(t)}\left[\left(\left[\mu_{hns}^f\right]^2 + \mathbf{\Sigma}_{ns[h,h]}^f\right)\frac{\hat{a}_\zeta}{\hat{b}_{\zeta_{hns}}}\right]$.
Set $\eta_{\kappa_F}$ as the weighted average of the values at this and the previous iterations

$$\eta_{\kappa_F}(t) = (1-\rho_t)\eta_{\kappa_F}(t-1) + \rho_t\hat{\eta}_{\kappa_F}.$$

(g) For $h = 1, \ldots, H$ update

$$\hat{\eta}_{z_h} = m_{hslab} - m_{hspike},$$

where

$$m_{h\,slab} = \mathbb{E}\log\pi_h + a_F\mathbb{E}\log b_F - \log\Gamma(a_F) - (a_F + \frac{N}{2} + 1)\mathbb{E}_{slab}\log\upsilon_h - \mathbb{E}_{slab}\left(\frac{1}{\upsilon_h}\right)\left(\mathbb{E}b_F + \frac{1}{2}\mathbb{E}\left(\frac{1}{\kappa_F}\right)\breve{S}_h^F\right),$$

$$m_{h\,spike} = \mathbb{E}\log(1-\pi_h) + a_0^F\mathbb{E}\log b_0^F - \log\Gamma(a_0^F) - (a_0^F + \frac{N}{2} + 1)\mathbb{E}_{spike}\log\upsilon_h - \mathbb{E}_{spike}\left(\frac{1}{\upsilon_h}\right)\left(\mathbb{E}b_0^F + \frac{1}{2}\mathbb{E}\left(\frac{1}{\kappa_F}\right)\breve{S}_h^F\right),$$

with $\breve{S}_h^F$ defined in the previous step (f) and other quantities defined in Step 15 of Algorithm A.1.
Set $\eta_{z_h}$ as the weighted average of the values at this and the previous iterations

$$\eta_{z_h}(t) = (1-\rho_t)\eta_{z_h}(t-1) + \rho_t\hat{\eta}_{z_h}.$$

7: Update the remaining parameters which do not depend on local variables according to Algorithm A.1.

# B Initial values, convergence check and hyperparameters

## B.1 Initialisations

In a VI algorithm, the result can be rather sensitive to initial parameters and unfortunate starting values can lead to the algorithm being stuck at a local optimum. Therefore a careful strategy for robust initialisation of the model parameters is usually required. We initialise out algorithm in a similar way to the procedure described in Hansen et al. (2025), with the difference that at first we run the regression on the known covariates to get the initial estimates of the mean values of $\boldsymbol{\beta}^X$ and then proceed using the residuals from the regression. This is done according to the following step-by-step process:

1. First, the regression model $\boldsymbol{y}_{ns} = \boldsymbol{\beta}^X \boldsymbol{x}_{ns} + \boldsymbol{e}_{ns}$ is solved for all $n = 1, \ldots, N_s$ and all $s = 1, \ldots, S$, which gives the initial values for $\mu_{il}^{\beta^X} = \hat{\beta}_{il}$ for $i = 1, \ldots, p$ and $l = 1, \ldots, p_x$. Note that the $P \times N$ dimensional data matrix $\boldsymbol{Y} = \{y_{ins}\}$ is constructed by stacking the subject-specific data matrices $\boldsymbol{Y}_s$ one after another for $s = 1, \ldots, S$ and setting $N = \sum_{s=1}^{S} N_s$.

2. Initial values for $\boldsymbol{\mu}_i^{\lambda}$ and $\boldsymbol{\mu}_{ns}^{f}$ are obtained from decomposing the $P \times N$ error matrix from the regression of the previous step $\boldsymbol{E} = \{e_{ins}\}$ as $\boldsymbol{E}^\top = \tilde{\boldsymbol{f}} \tilde{\boldsymbol{\Lambda}}^\top$, where the common loadings matrix $\tilde{\boldsymbol{\Lambda}}$ of dimensions $P \times H$ and factor scores $\tilde{\boldsymbol{f}}$ of dimensions $N \times H$ are estimated via an unconstrained sparse principal component analysis (PCA) (Erichson et al., 2020) available in R-package `sparsepca`, with $H$ components Then we set $\boldsymbol{\mu}_i^{\lambda} = \tilde{\lambda}_{ih}$, where $\tilde{\lambda}_{ih}$ are elements $(i, h)$ of $\tilde{\boldsymbol{\Lambda}}$, and $\boldsymbol{\mu}_{ns}^{f} = \tilde{f}_{nsh}$, where $\tilde{f}_{nsh}$ refers to the element of the stacked matrix of common factor scores $\tilde{\boldsymbol{f}}$ corresponding to sample $n$ of subject $s$ and factor $h$.

3. As the next step, to define the study-specific residuals $\tilde{\boldsymbol{E}}_s$ as

$$\tilde{\boldsymbol{E}}_s^\top = \boldsymbol{E}_s^\top \left( \boldsymbol{I}_p - \tilde{\boldsymbol{\Lambda}} \left( \tilde{\boldsymbol{\Lambda}}^\top \tilde{\boldsymbol{\Lambda}} \right)^+ \tilde{\boldsymbol{\Lambda}}^\top \right),$$

   where $\boldsymbol{E}_s$ is the $p \times N_s$ matrix of regression residuals with the indices $1, \ldots, N_s$ corresponding to subject $s$ and the superscript "+" indicates the Moore-Penrose inverse. These are the sparse PCA residuals proposed by Camacho et al. (2020), which subtract the common part of the variability from the data.

4. Now the initial values for $\boldsymbol{\mu}_{is}^{\lambda}$ and $\boldsymbol{\mu}_{ns}^{\eta}$ are obtained from decomposing the $P \times N_s$ subject-specific error matrix $\tilde{\boldsymbol{E}}_s$ as $\tilde{\boldsymbol{E}}_s^\top = \tilde{\boldsymbol{\eta}}_s \tilde{\boldsymbol{\Lambda}}_s^\top$, with the constrained that $Cov(\tilde{\boldsymbol{\eta}}_s) = \boldsymbol{I}_{H_s}$. The subject-specific loadings matrix $\tilde{\boldsymbol{\Lambda}}_s$ of dimensions $P \times H_s$ and factor scores $\tilde{\boldsymbol{\eta}}_s$ of dimensions $N_s \times H_s$ are estimated via an unconstrained sparse principal component analysis (PCA) (Erichson et al., 2020) available in R-package `sparsepca` with $H_s$ components and are appropriately scaled. Then we set $\boldsymbol{\mu}_{is}^{\lambda} = \tilde{\lambda}_{ihs}$, where $\tilde{\lambda}_{ihs}$ is element $(i, h)$ of $\tilde{\boldsymbol{\Lambda}}_s$, and $\boldsymbol{\mu}_{ns}^{\eta} = \tilde{\eta}_{nsh}$, where $\tilde{\eta}_{ns}$ refers to the $(n, h)$ element of $\tilde{\boldsymbol{\eta}}_s$.

5. Initial values for the parameters of subject-specific errors $\xi_{is}^2$ are determined in the following way. For $s = 1, \ldots, S$, set $\hat{a}_{is}^{\xi} = a_{\xi} + N_s/2$, then let $(\tilde{\xi}_{1s}^2, \ldots, \tilde{\xi}_{ps}^2) = \mathrm{diag}(\tilde{\boldsymbol{E}}_s \tilde{\boldsymbol{E}}_s^\top - \tilde{\boldsymbol{\Lambda}}_s \tilde{\boldsymbol{\Lambda}}_s^\top)$ and set $\hat{b}_{is}^{\xi} = (\hat{a}_{is}^{\xi} - 1) \times \tilde{\xi}_{is}^2$. Set $\hat{b}_{is}^{b_\xi} = b_{gis} + \frac{(\hat{a}_{is}^{\xi} - 1)}{\hat{b}_{is}^{\xi}}$.

6. Initial values for $\mathbf{\Sigma}_i^{\beta^X}$ are defined by setting $\mathbf{\Sigma}_i^{\beta^X} = \left( \boldsymbol{Q}_i^{-1} + \sum_{s=1}^{S} \frac{(\hat{a}_{is}^{\xi} - 1)}{\hat{b}_{is}^{\xi}} \boldsymbol{X}_s \boldsymbol{X}_s^{\top} \right)^{-1}$, where $\boldsymbol{Q}_i = \text{diag}(q_{i1}, \ldots, q_{ip_X})$ and $\boldsymbol{X}_s$ is the $p_X \times N_s$ matrix of covariates for subject $s$.

7. Concentration parameters for slab probabilities are intialised as a mean of the prior distribution. As the prior distribution is gamma, the initial values for the slab concentration parameters are the following: for common loadings matrix $\alpha_\Lambda = a_\alpha^\Lambda / b_\alpha^\Lambda$, for the common factors $\alpha_F = a_\alpha^F / b_\alpha^F$, and for subject-specific factor loadings $\alpha_\Lambda^S = a_\alpha^{S\Lambda} / b_\alpha^{S\Lambda}$.

8. Initial values for the global sparsity parameters $\kappa_\Lambda$ and $\kappa_F$ are set equal for the prior hyperparameters, namely $\hat{a}_{\kappa_\Lambda} = a^\kappa$, $\hat{b}_{\kappa_\Lambda} = b^\kappa$, $\hat{a}_{\kappa_F} = c^\kappa$ and $\hat{b}_{\kappa_F} = d^\kappa$.

9. Initial values for the element-wise sparsity parameters $\phi_{ih}$, $\zeta_{hns}$ and $\phi_{ihs}$ are set equal to prior hyperparameters, with the rate parameters being equal to its prior mean, namely: $\hat{a}_{\beta_{i\phi}} = c$, $\hat{b}_{\beta_{i\phi}} = d$, $\hat{a}_{\phi_{ih}} = a_\phi$, $\hat{b}_{\phi_{ih}} = c/d$, $\hat{a}_{\beta_{ns\zeta}} = e$, $\hat{b}_{\beta_{ns\zeta}} = w$, $\hat{a}_{\zeta_{hns}} = a_\zeta$, $\hat{b}_{\zeta_{hns}} = e/w$, and $\hat{a}_{\beta_{\phi s}} = c^S$, $\hat{b}_{\beta_{\phi is}} = d^S$, $\hat{a}_{\phi_s} = a_\phi^S$, $\hat{b}_{\phi_{ihs}} = c^S/d^S$.

10. Initial values of the spike and the slab parameters of $\theta_h$ are set equal to prior hyperparameters, with the rate parameter being equal to its prior mean, namely: $\hat{a}_{b_0^\Lambda} = a_2^\Lambda$, $\hat{b}_{b_0^\Lambda} = b_2^\Lambda$, $\hat{a}_{\theta_h spike} = a_0^\Lambda$, $\hat{b}_{\theta_h spike} = a_2^\Lambda / b_2^\Lambda$, and $\hat{a}_{b_\theta} = a_1^\Lambda$, $\hat{b}_{b_\theta} = b_1^\Lambda$, $\hat{a}_{\theta_h slab} = a_\Lambda$, $\hat{b}_{\theta_h slab} = a_1^\Lambda / b_1^\Lambda$.

11. Initial values of the spike and the slab parameters of $\upsilon_h$ are set equal to prior hyperparameters, with the rate parameter being equal to its prior mean, namely: $\hat{a}_{b_0^F} = a_2^F$, $\hat{b}_{b_0^F} = b_2^F$, $\hat{a}_{\upsilon_h spike} = a_0^F$, $\hat{b}_{\upsilon_h spike} = a_2^F / b_2^F$, and $\hat{a}_{b_\upsilon} = a_1^F$, $\hat{b}_{b_\upsilon} = b_1^F$, $\hat{a}_{\upsilon_h slab} = a_F$, $\hat{b}_{\upsilon_h slab} = a_1^F / b_1^F$.

12. Initial values of the spike and the slab parameters of $\theta_{hs}$ are set equal to prior hyperparameters, with the rate parameter being equal to its prior mean, namely: $\hat{a}_{b_0^{S\Lambda}} = a_2^{S\Lambda}$, $\hat{b}_{b_0^{S\Lambda}} = b_2^{S\Lambda}$, $\hat{a}_{\theta_{hs} spike} = a_0^{S\Lambda}$, $\hat{b}_{\theta_{hs} spike} = a_2^{S\Lambda} / b_2^{S\Lambda}$, and $\hat{a}_{b_\Lambda^S} = a_1^{S\Lambda}$, $\hat{b}_{b_\Lambda^S} = b_1^{S\Lambda}$, $\hat{a}_{\theta_{hs} slab} = a_\Lambda^S$, $\hat{b}_{\theta_{hs} slab} = a_1^{S\Lambda} / b_1^{S\Lambda}$.

13. Initial assignments of $\theta_h$, $\upsilon_h$ and $\theta_{hs}$ to spike or slab is done by randomly sampling from the probabilities of the slab, when the probabilities of the slab is set as $\alpha_\Lambda/(\alpha_\Lambda + H)$, $\alpha_F/(\alpha_F + H)$, and $\alpha_\Lambda^S/(\alpha_\Lambda^S + H)$, correspondingly.

14. Initial values for common factor loading covariance matrix $\mathbf{\Sigma}_i^\lambda$ are set as $\mathbf{\Sigma}_i^\lambda = (\mathbf{\Psi}_i^{init})^{-1}$, where $\mathbf{\Psi}_i^{init} = diag\left( \frac{\hat{a}_{\kappa_\Lambda}}{\hat{b}_{\kappa_\Lambda}} \frac{\hat{a}_\phi}{\hat{b}_{\phi_{i1}}} \mathbb{E}_{mix}\left(\frac{1}{\theta_1}\right), \ldots, \frac{\hat{a}_{\kappa_\Lambda}}{\hat{b}_{\kappa_\Lambda}} \frac{\hat{a}_\phi}{\hat{b}_{\phi_{iH}}} \mathbb{E}_{mix}\left(\frac{1}{\theta_H}\right) \right)$ with $\mathbb{E}_{mix}\left(\frac{1}{\theta_h}\right) = r_h \frac{\hat{a}_{\theta_h slab}}{\hat{b}_{\theta_h slab}} + (1 - r_h) \frac{\hat{a}_{\theta_h spike}}{\hat{b}_{\theta_h spike}}$, and the parameters are initiated as described above.

15. Initial values for subject-specific factor loading covariance matrix $\mathbf{\Sigma}_{is}^\lambda$ are set as $\mathbf{\Sigma}_{is}^\lambda = (\mathbf{\Psi}_{is}^{init})^{-1}$, where $\mathbf{\Psi}_{is}^{init} = diag\left( \mathbb{E}_{mix}\left(\frac{1}{\theta_{1s}}\right), \ldots, \mathbb{E}_{mix}\left(\frac{1}{\theta_{Hs}}\right) \right)$ with $\mathbb{E}_{mix}\left(\frac{1}{\theta_{hs}}\right) = r_{hs} \frac{\hat{a}_{\theta_{hs} slab}}{\hat{b}_{\theta_{hs} slab}} + (1 - r_{hs}) \frac{\hat{a}_{\theta_{hs} spike}}{\hat{b}_{\theta_{hs} spike}}$, and the parameters are initiated as described above.

16. Subject-specific factor covariance matrix is initialised as $\mathbf{\Sigma}_{ns}^\eta = \left( \boldsymbol{I}_{H_s} + \sum_{i=1}^{p} \tilde{\xi}_{is}^2 \left( \boldsymbol{\mu}_{is}^\lambda \left[ \boldsymbol{\mu}_{is}^\lambda \right]^\top + \mathbf{\Sigma}_{is}^\lambda \right) \right)^{-1}$

17. Common factor covariance matrix is initialised as $\mathbf{\Sigma}_{ns}^{f} = \left(\mathbf{\Omega}_{ns}^{-1} + \sum_{i=1}^{p} \tilde{\xi}_{is}^{2} \left(\boldsymbol{\mu}_{i}^{\lambda} \left[\boldsymbol{\mu}_{i}^{\lambda}\right]^{\top} + \mathbf{\Sigma}_{i}^{\lambda}\right)\right)^{-1}$, where $\mathbf{\Omega}_{ns}^{-1} = diag\left(\frac{\hat{a}_{\kappa_F}}{\hat{b}_{\kappa_F}} \frac{\hat{a}_{\zeta}}{\hat{b}_{\zeta_{1ns}}} \mathbb{E}_{mix}\left(\frac{1}{\upsilon_1}\right), \ldots, \frac{\hat{a}_{\kappa_F}}{\hat{b}_{\kappa_F}} \frac{\hat{a}_{\zeta}}{\hat{b}_{\phi_{Hns}}} \mathbb{E}_{mix}\left(\frac{1}{\upsilon_H}\right)\right)$ with $\mathbb{E}_{mix}\left(\frac{1}{\upsilon_h}\right) = z_h \frac{\hat{a}_{\upsilon_h slab}}{\hat{b}_{\upsilon_h slab}} + (1 - z_h) \frac{\hat{a}_{\upsilon_h spike}}{\hat{b}_{\upsilon_h spike}}$, and the parameters are initialised as described above.

## B.2 Hyperparameter specification

We specify the hyperparameters of the sparse exchangeable shrinkage process prior (ESP) to induce structured sparsity in both the common and study-specific components while allowing sufficient flexibility for dense signals. The slab inclusion probabilities for the common factor loadings and factor scores are governed by Beta distributions with parameters $(a_{\alpha}^{\Lambda}, b_{\alpha}^{\Lambda}) = (8, 2)$ and $(a_{\alpha}^{F}, b_{\alpha}^{F}) = (8, 2)$, favouring inclusion and thus allowing for a moderate number of common factors. In contrast, the study-specific factor loadings are assigned $(a_{\alpha}^{S\Lambda}, b_{\alpha}^{S\Lambda}) = (1, 4)$, encouraging smaller number of study-specific effects.

Global sparsity for common loadings and scores is controlled via inverse-Gamma hyperpriors with parameters $(a^{\kappa}, b^{\kappa}) = (3, 1)$ and $(c^{\kappa}, d^{\kappa}) = (3, 1)$, which favour moderate shrinkage while still allowing flexibility when supported by the data. Element-wise sparsity is further regulated through hierarchical inverse-Gamma distributions on local shrinkage parameters, with parameters $(a_{\phi}, c, d) = (5, 1, 1)$ for common loadings, $(a_{\phi}^{S}, c^{S}, d^{S}) = (3, 1, 1)$ for study-specific loadings, and $(a_{\zeta}, e, w) = (5, 1, 1)$ for common factor scores, promoting a balance between sparse and dense patterns.

The spike-and-slab components are specified through inverse-Gamma priors on the column-specific variance parameters, with separate Gamma hyperpriors on their rate parameters. For the common loading variances, the slab component is assigned shape parameter $a_{\Lambda} = 3$, with rate parameter $b_{\Lambda}$ given the hyperprior $b_{\Lambda} \sim \mathcal{G}(a_1^{\Lambda}, b_1^{\Lambda})$ with $(a_1^{\Lambda}, b_1^{\Lambda}) = (2, 1)$. We use the same values for the corresponding common factor-score variances: $a_F = 3$ and $b_F \sim \mathcal{G}(a_1^{F}, b_1^{F})$, with $(a_1^{F}, b_1^{F}) = (2, 1)$. For the study-specific loading variances, we use $a_{\Lambda}^{S} = 2$ and $b_{\Lambda}^{S} \sim \mathcal{G}(a_1^{S\lambda}, b_1^{S\Lambda})$, with $(a_1^{S\Lambda}, b_1^{S\Lambda}) = (2, 1)$. The spike components are specified analogously, but with larger inverse-Gamma shape parameters (please see Grushanina and Frühwirth-Schnatter (2025) for the reasoning behind this choice), $a_0^{\Lambda} = 21$, $a_0^{F} = 21$, and $a_0^{S\Lambda} = 21$, together with Gamma hyperpriors on the corresponding spike rate parameters: $b_0^{\Lambda} \sim \mathcal{G}(a_2^{\Lambda}, b_2^{\Lambda})$ with $(a^{L}ambda_2, b_2^{\Lambda}) = (1, 2)$, $b_0^{F} \sim \mathcal{G}(a_2^{F}, b_2^{F})$ with $(a_2^{F}, b_2^{F}) = (1, 1)$, and $b_0^{S\Lambda} \sim \mathcal{G}(a_2^{S\lambda_S}, b_2^{S\Lambda})$ with $(a_2^{S\lambda}, b_2^{S\Lambda}) = (1, 12)$. This choice yields clearly separated spike and slab regimes, with the spike components favouring substantially stronger shrinkage of inactive factors.

Overall, this specification encourages a decomposition with flexible common structure and moderately dense study-specific contributions, while allowing the data to determine the degree of sparsity at both global and local levels.

## B.3 Convergence

Convergence is monitored using blockwise relative Frobenius distances between successive iterates of the variational parameters. For each parameter block, the Frobenius norm of the difference between iterations is scaled by the norm of the previous iterate. For Bernoulli parameters (responsibilities), distances are evaluated on the logit scale. The algorithm is deemed converged when the maximum relative Frobenius distance across all parameter blocks falls below a fixed tolerance level for several consecutive iterations.

# C Postprocessing

## C.1 Post hoc identification procedure

In this section, we describe the procedure used to check identifiability of latent factors based on their unique support, as outlined in Section 4.4 of the main text.

**Step 1:** *Compute activity energy for each factor*

In variational inference, the ESP prior does not yield a sharp separation between active and inactive factors, as the corresponding $0/1$ indicators are replaced by continuous responsibilities. While these quantities encode relative shrinkage, they are not guaranteed to concentrate at the extremes and thus may not provide a stable criterion for factor selection. This motivates the use of an alternative criterion for identifying active factors.

We identify active factors using posterior summaries of factor loadings and factor scores. We define activity energy for each factor as the sum of posterior second moments, combining posterior means and variances. This quantity corresponds to the expected squared magnitude of the factor under the variational posterior and is closely related to measures of factor strength used in sparse factor models (e.g. Hochreiter et al. (2010), Bhattacharya and Dunson (2011)).

For common factors, we compute activity energy separately for factor loadings and factor scores. The factor loading energy for factor $h$ is defined as

$$A_h^{(\lambda)} = \sum_{i=1}^{p} \left( \left[ \mu_{ih}^{\lambda} \right]^2 + \mathbf{\Sigma}_{i[h,h]}^{\lambda} \right),$$

and the factor score energy as

$$A_h^{(f)} = \sum_{s=1}^{S} \sum_{n=1}^{N_s} \left( \left[ \mu_{hns}^{f} \right]^2 + \mathbf{\Sigma}_{ns[h,h]}^{f} \right).$$

Activity energy for study-specific factors is computed separately for each study $s$ and factor $r$ in the following way:

$$A_r^{(\lambda_s)} = \sum_{i=1}^{p} \left( \left[ \mu_{irs}^{\lambda} \right]^2 + \mathbf{\Sigma}_{is[r,r]}^{\lambda} \right),$$

$$A_r^{(\eta_s)} = \frac{1}{N_s} \sum_{n=1}^{N_s} \left( [\mu_{rns}^{\eta}]^2 + \mathbf{\Sigma}_{ns[r,r]}^{\eta} \right).$$

The averaging by $N_s$ in $A_r^{(\eta_s)}$ ensures that the study-specific energy is not driven purely by study size. We then combine both metrics into an overall energy measure for study-specific factor $r$ in study $s$: $A_r^{(s)} = A_r^{(\lambda_s)} A_r^{(\eta_s)}$.

**Step 2:** *Select active factors by relative thresholding*

As the next step, we identify active factors by comparing the activity energy of each factor to its largest value in the corresponding set.

For the common component, define the active factor loading columns and the active factor score rows as

$$\mathcal{H}_{\lambda}^{+} = \left\{ h : A_h^{(\lambda)} > \delta \max_l A_l^{(\lambda)} \right\} \quad \text{and} \quad \mathcal{H}_{f}^{+} = \left\{ h : A_h^{(f)} > \delta \max_l A_l^{(f)} \right\},$$

where $h, l \in \{1, \ldots, H\}$.

The set of active common factors is then defined as $\mathcal{H}^+ = \mathcal{H}^+_\lambda \cap \mathcal{H}^+_f$. Thus, a common factor is retained as active only if it is active both in its loading column and in its factor score row.

For study-specific factors in study $s$ we define the set of active factors as

$$\mathcal{H}^+_s = \left\{ r : A^{(s)}_r > \delta \max_l A^{(\lambda_s)}_l \right\},$$

where $r, l \in \{1, \ldots, H_s\}$. In our simulation studies and application to real data we used $\delta = 0.1$.

**Step 3:** *Estimate supports of active factors*

After selecting the active factors, we recover their supports by thresholding posterior means in the following way. For each active factor $h \in \mathcal{H}^+$, define its feature support as

$$G_h = \left\{ i \in 1, \ldots, p : |\mu^\lambda_{ih}| > c_\lambda \max_{1 \le j \le p} |\mu^\lambda_{jh}| \right\},$$

and its sample support as

$$T_h = \left\{ n \in 1, \ldots, N : |\mu^f_{hn}| > c_f \max_{1 \le m \le N} |\mu^f_{hm}| \right\},$$

where $N = N_1 + \ldots + N_S$.

Thus, $G_h$ contains the features with sufficiently large posterior mean loadings on factor $h$ while $T_h$ contains the samples for which factor $h$ has sufficiently large posterior mean score. Then, for each active factor $h \in \mathcal{H}^+$ we define its total support as the Cartesian product of its feature and sample supports $B_h = G_h \times T_h$. Thus, $B_h$ represents the set of entries of the data matrix where both the loadings and scores of factor $h$ are non-negligible.

Then, for each active study-specific factor $r \in \mathcal{H}^+_s$, define its feature support as

$$G^s_r = \left\{ i \in 1, \ldots, p : |\mu^{\lambda_s}_{ir}| > c_{\lambda_s} \max_{1 \le j \le p} |\mu^{\lambda_s}_{jr}| \right\}.$$

The sample support of the study-specific factors is fixed by the model structure and is given by $T^s_r = \mathcal{I}_s$, where $\mathcal{I}_s$ denotes the set of indices corresponding to the samples belonging to study $s$. This reflects the fact that the study-specific factors are, by modeling choice, active across all samples within their corresponding study. Then, for each active study-specific factor $r \in \mathcal{H}^+_s$ in study $s$ we define its total support as $B^s_r = G^s_r \times \mathcal{I}_s$.

In our simulation studies and application to real data we used $c_\lambda = c_f = c_{\lambda_s} = 0.1$.

**Step 4:** *Calculate unique support for common factors*

For each active common factor $h$, the unique support is calculated by comparing its total support with the supports of all other active factors. Computationally, the total supports $B_h$ and $B^s_r$ can be represented as a binary mask of the data matrix, with entries equal to 1 if the factor is active at that location and 0 otherwise. With this in view, the unique support of common factor $h$ can be computed as follows:

1. Create a binary mask for each common factor $M_h$ such that $M_h(i, n) = 1$ if $(i, n) \in B_h$, otherwise $M_h(i, n) = 0$.

2. Create a binary mask for each study-specific factor in each study $s$ $M^s_r$ such that $M^s_r(i, n) = 1$ if $(i, n) \in B^s_r$, otherwise $M^s_r(i, n) = 0$, where $n = 1, \ldots, N_s$.

3. Construct the combined mask of all other factors by summing the binary masks of all other active common factors $l \neq h$ and all active study-specific factors in all studies $s = 1, \ldots, S$ as $M_{-h}(i,n) = \mathbf{1}\left\{\sum_{l \neq h} M_l(i,n) + \sum_{s=1}^{S} \sum_{r \in \mathcal{H}_s^+} M_r^s(i,n) > 0\right\}$.

4. Compute the unique support mask of common factor $h$ as $U_h = M_h \odot (1 - M_{-h})$, where $\odot$ denotes element-wise multiplication.

5. Calculate the size of the unique support as $|U_h| = \sum_{i,n} U_h(i,n)$.

A common factor $h$ is considered identified with respect to information switching if $|U_h| > 0$, that is, if its unique support is non-empty. We restrict the construction of unique supports to common factors, as study-specific factors are usually less sparse and, by design, active across all samples within a study, making the uniqueness criterion less likely to be satisfied. Moreover, our primary interest lies in the identification of common factors.

# References


Bhattacharya, A. and Dunson, D. (2011). Sparse bayesian infinite factor models. *Biometrika*, 98(2):291–306.

Camacho, J., Smilde, A., Saccenti, E., and Westerhuis, J. (2020). All sparse pca models are wrong, but some are useful. part i: Computation of scores, residuals and explained variance. *Chemometrics and Intelligent Laboratory Systems*, 196:103907.

Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of American Statistical Association*, 103(484):1438–1456.

Erichson, N. B., Zheng, P., Manohar, K., Brunton, S. L., Kutz, J. N., and Aravkin, A. Y. (2020). Sparse principal component analysis via variable projection. *SIAM Journal on Applied Mathematics*, 80(2):977–1002.

Hansen, B., Avalos-Pacheco, A., Russo, M., and Vito, R. D. (2025). Fast variational inference for bayesian factor analysis in single and multi-study settings. *Journal of Computational and Graphical Statistics*, 34(1):96–108.

Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., Khamiakova, T., Van Sanden, S., Lin, D., Talloen, W., Bijnens, L., Göhlmann, H. W. H., Shkedy, Z., and Clevert, D.-A. (2010). Fabia: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527.

Lopes, H. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–67.