

## RESEARCH ARTICLE

# Semi-supervised empirical Bayes group-regularized factor regression

Magnus M. Münch<sup>1,2</sup> | Mark A. van de Wiel<sup>1,3</sup> | Aad W. van der Vaart<sup>4</sup> |  
Carel F. W. Peeters<sup>1,5</sup>

<sup>1</sup>Department of Epidemiology & Data Science, Amsterdam UMC, Amsterdam, The Netherlands

<sup>2</sup>Mathematical Institute, Leiden University, Leiden, The Netherlands

<sup>3</sup>MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, UK

<sup>4</sup>Delft Institute of Applied Mathematics - TU Delft, Delft, The Netherlands

<sup>5</sup>Mathematical & Statistical Methods Group (Biometris), Wageningen University & Research, Wageningen, The Netherlands

## Correspondence

Carel F. W. Peeters, Mathematical & Statistical Methods group (Biometris), Wageningen University & Research, PO Box 16, 6700AA Wageningen, The Netherlands.

Email: [carel.peeters@wur.nl](mailto:carel.peeters@wur.nl)



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

## Abstract

The features in a high-dimensional biomedical prediction problem are often well described by low-dimensional latent variables (or factors). We use this to include unlabeled features and additional information on the features when building a prediction model. Such additional feature information is often available in biomedical applications. Examples are annotation of genes, metabolites, or  $p$ -values from a previous study. We employ a Bayesian factor regression model that jointly models the features and the outcome using Gaussian latent variables. We fit the model using a computationally efficient variational Bayes method, which scales to high dimensions. We use the extra information to set up a prior model for the features in terms of hyperparameters, which are then estimated through empirical Bayes. The method is demonstrated in simulations and two applications. One application considers influenza vaccine efficacy prediction based on microarray data. The second application predicts oral cancer metastasis from RNAseq data.

## KEYWORDS

empirical Bayes, factor regression, high-dimensional data, semisupervised learning

## 1 | INTRODUCTION

Modern biomedical research utilizes models based on large sets of omics features to predict outcomes such as categorical disease status, time-to-event, or continuous anthropomorphic measures. The number of omics features may run in the tens of thousands (in, e.g., genomics), but the number of samples may be low, due to high measurement costs, logistics, or the availability of subjects. The high-dimensionality of the data (i.e.,  $p > n$ ) complicates model estimation.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

Here, we propose a novel method, based on factor models, that enhances high-dimensional prediction in two ways. First, it can incorporate unlabeled samples, for which the predictor features are available, but not the response/outcome. Second, it allows incorporation of prior information on the features, for instance from previous studies, by automatically adapting prior modeling parameters. Both types of external data are often available in omics studies.

## 1.1 | Contributions and relation to the literature

Several authors have argued that the high-dimensional feature space in omics data arises from noisy observations on a lower dimensional latent space. West (2003) shows that gene expression data from breast cancer patients are indeed well described with a lower dimensional (linear) latent space. Moreover, Carvalho et al. (2008) improve the prediction of the mutant p53 gene versus the wild type in breast cancer patients with the lower dimensional structure of the gene expression data. West (2003) and Carvalho et al. (2008) use a Bayesian linear factor (regression) model approach to describe the latent space. Mes et al. (2020) is an example of a frequentist latent space approach (technically a hybrid between Bayes and frequentist) to a prediction from radiomic features. In this paper, we use this observation to include external information that can enhance the fitting of a high-dimensional prediction model.

Unlabeled feature data are one type of external information. Such data may, for example, come from online repositories or previous studies with the same set of features but with a different response. The inclusion of unlabeled data in prediction problems, termed semisupervised learning in the machine learning community, has received plenty of attention (see Zhu & Goldberg, 2009, for an introduction). The factor regression model can naturally exploit such unlabeled data, as shown in Bañbura and Modugno (2014) and Liu and Rubin (1998) and argued convincingly in Liang et al. (2007).

In addition, extra information on the features, termed codata, is often available. This may consist of a partitioning of the features, such as pathway membership of the genes, or continuous information, such as  $p$ -values from a previous study. Recently, several methods have been introduced that use the codata to improve prediction (see, e.g., Münch et al., 2021; te Beest et al., 2017; van Nee et al., 2020; van de Wiel et al., 2016).

Our contribution in the current paper is to combine these two types of external data. We extend the codata approach (more specifically, a group-adaptive empirical Bayes approach akin to that in Münch et al., 2021) to the Bayesian factor regression model that can include unlabeled data. We achieve this by developing a variational Bayes procedure that scales to high dimensions, augmented with an empirical Bayes procedure to estimate hyperparameters that encode codata. We also extend the method to a mixed-mode factor analysis, where the outcome is binary instead of continuous.

The current model differs from Liang et al. (2007) in two main ways: (i) it considers the more flexible factor model as opposed to the principal component regression model in Liang et al. (2007), and (ii) it integrates the factor and regression models into one framework, while Liang et al. (2007) take a heuristic two-step approach that separates the latent space analysis and regression. Our approach also differs from Avalos-Pacheco et al. (2022) in several ways: (i) we focus throughout on the prediction of a single outcome, using a prior that differentiates between features and outcome, whereas Avalos-Pacheco et al. (2022) focus more on dimension reduction and in the setting of predictive survival regression takes a more heuristic approach to factor regression through separate latent space analysis and regression steps. We (ii) also consider the binary response setting; (iii) provide full (variational) posteriors next to point estimates; (iv) include a grouping of the features, whereas Avalos-Pacheco et al. (2022) groups the observations (batch effect); (v) develop empirical Bayes estimation of hyperparameters, which is essential to our approach; (vi) allow for semisupervised learning.

## 1.2 | Overview

Simulations show that the approach is competitive or even outperforms classical approaches in some settings. Applications to influenza vaccine efficacy prediction and oral cancer lymph node metastasis prediction show that the approach has the potential to enhance predictive performance compared to existing methods. The remainder of the paper is organized as follows: Sections 2 and 3 describe the model and its estimation in detail. The approach is demonstrated in a simulated setting in Section 4 and two real data settings in Section 5. We conclude with a short discussion on the pros and cons of the method in Section 6.

## 2 | MODEL

### 2.1 | Observational model

The observations consist of centered  $p$ -dimensional feature vectors  $\mathbf{x}_i$  with corresponding outcomes  $y_i$ ,  $i = 1, \dots, n$  and (possibly) an additional sample of feature vectors  $\mathbf{x}_i$ ,  $i = n + 1, \dots, n + m$ . We assume that the observations for different  $i$  are independent, that all feature vectors are identically distributed, and that a typical pair  $(\mathbf{x}_i, y_i)$  follows a factor regression model (Liang et al., 2007) given for a single observation  $(\mathbf{x}, y)$  as

$$y|\boldsymbol{\lambda} \sim \mathcal{N}(\boldsymbol{\beta}^T \boldsymbol{\lambda}, \sigma^2), \quad (1a)$$

$$\mathbf{x}|\boldsymbol{\lambda} \sim \mathcal{N}_p(\mathbf{B}^T \boldsymbol{\lambda}, \boldsymbol{\Psi}), \quad (1b)$$

$$\boldsymbol{\lambda} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d). \quad (1c)$$

Here  $\boldsymbol{\lambda}$  are latent factor variables,  $\boldsymbol{\Psi} = \text{diag}(\psi_j)$ ,  $j = 1 \dots, p$ , are the uniqueness (residual variances),  $\sigma^2$  is the error variance, and  $\mathbf{B}$  and  $\boldsymbol{\beta}$  are the factor loadings. The latent factor dimension  $d$  is initially assumed to be fixed and known. The latent factors are determined only up to rotational invariance, but this does not play a major role in prediction; see Section 2 in the [Supporting Information for discussion](#).

Model (1) implies a joint multivariate Gaussian distribution for  $[\mathbf{x}^T y]^T$  (not conditioned on  $\boldsymbol{\lambda}$ ), and a prediction of a new outcome from observed new features  $\tilde{\mathbf{x}}$  is given by the conditional expectation:

$$\mathbb{E}(\tilde{y}|\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}^T (\mathbf{B}^T \mathbf{B} + \boldsymbol{\Psi})^{-1} \mathbf{B}^T \boldsymbol{\beta} =: \tilde{\mathbf{x}}^T \tilde{\boldsymbol{\beta}}. \quad (2)$$

We shall also develop the method for the case that the outcomes  $y_i$  are sums of  $N_i$  disjoint binary events with a shared probability of success. In this case, the linear outcome model (1a) is replaced with the logistic counterpart:

$$y|\boldsymbol{\lambda}, \boldsymbol{\beta}, \beta_0 \sim \mathcal{B}(N, \text{expit}(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{\lambda})), \quad (3)$$

where  $\mathcal{B}(N, \pi)$  denotes the binomial distribution with a number of trials  $N$  and success probability  $\pi$ . Note that the logistic model includes an intercept  $\beta_0$  to accommodate unbalanced data, whereas the linear model simply considers standardized data.

Feature and factor models (1b) and (1c), in combination with outcome model (3) result in a mixed-mode factor model, with Gaussian and binomially distributed features and outcomes, respectively. This mixed-mode extension is detailed in Section 5 of the [Supporting Information](#).

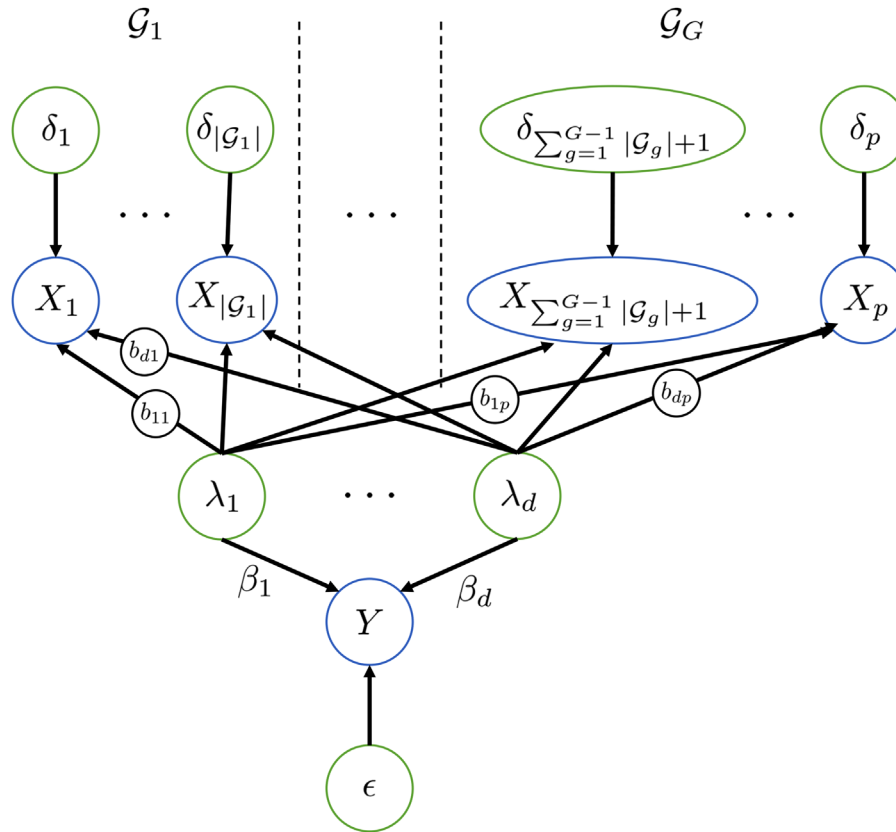
### 2.2 | Bayesian prior model

In the Bayesian version of the model, the parameters  $\theta := \{\mathbf{B}, \boldsymbol{\beta}, \psi_1, \dots, \psi_p, \sigma^2\}$  are endowed with conditionally conjugate prior distributions. For notational convenience, write  $\psi_{p+1} = \sigma^2$ , and let  $\mathbf{b}_j$  be the  $j$ th column of the matrix  $\mathbf{B}$ . Consider the priors:

$$\mathbf{b}_1, \dots, \mathbf{b}_p, \boldsymbol{\beta} | \psi_1, \dots, \psi_p, \psi_{p+1} \sim \prod_{j=1}^{p+1} \mathcal{N}_d(\mathbf{0}_d, \psi_j \gamma_j \mathbf{I}_d), \quad (4a)$$

$$\psi_1, \dots, \psi_{p+1} \sim \prod_{j=1}^{p+1} \Gamma^{-1}(\kappa_j, \nu_j), \quad (4b)$$

where  $\Gamma^{-1}(\kappa, \nu)$  denotes the inverse Gamma distribution with shape  $\kappa$  and scale  $\nu$ . The hyperparameters  $\gamma_j$  will be chosen to include prior information on the features (see the next section) and be estimated by an empirical Bayes method (see



**FIGURE 1** Model (1) with partitioned features as a Bayesian network, where the vertical dotted lines denote a partitioning of features  $X_1, \dots, X_p$  into groups  $g = 1, \dots, G$ . Green and blue circles denote latent and observed variables, respectively. Note that  $\delta_j$  and  $\epsilon$  are implicit in model (1) and omitted here for brevity. Here they denote the Gaussian, centered errors. That is, we have  $y = \beta^T \lambda + \epsilon$ ,  $\mathbf{x} = \mathbf{B}^T \lambda + \delta$ , with  $\epsilon \sim \mathcal{N}(0, \sigma^2)$  and  $\delta \sim \mathcal{N}_p(0, \Psi)$

Section 3.4). The prior variances of the  $\mathbf{b}_j$  and  $\beta$  scale with the uniqueness and error variance  $\psi_j$ , as is common in Bayesian (univariate) linear models. This is mostly for computational reasons but is often justified as a solution to scaling problems in multivariate regression problems (Leday et al., 2017).

In the Bayesian model, a prediction  $\tilde{y}$  from features  $\tilde{\mathbf{x}}$  is obtained by averaging over the posterior:

$$\mathbb{E}^*(\tilde{y}|\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}^T \mathbb{E}_{\mathbf{B}, \beta, \Psi | \tilde{\mathbf{x}}} [(\mathbf{B}^T \mathbf{B} + \Psi)^{-1} \mathbf{B}^T \beta] =: \tilde{\mathbf{x}}^T \beta^*. \quad (5)$$

In practice, this expectation is hard to compute. Here, we use a combination of variational Bayes for posterior computation and Monte Carlo simulation for approximation of (5). An alternative to Monte Carlo simulation is the Taylor approximation, as explained in Section 4.3 of the [Supporting Information](#).

### 2.3 | Feature-partitioning based on codata

In some applications, the features naturally come partitioned into groups  $\mathcal{G}_1, \dots, \mathcal{G}_G$ . Examples are the distinct functional networks of genes, features with significant versus features with nonsignificant association to the outcome in a previous study, and feature groups based on prior expert knowledge of feature importance (see, e.g., Münch et al., 2021). Figure 1 displays model (1) with partitioned features as a Bayesian network.

Partitioning can be included in the model through prior modeling of (i) the factor loadings  $\mathbf{B}$  or (ii) the residual variances  $\Psi$ . Here, we pursue option (i) and model the feature structure by considering groupwise constant (up to scaling by the uniqueness  $\psi_j$ ) prior variances, that is, in (4) we choose a single value of  $\gamma_j$  for every group of features:  $\forall j \in \mathcal{G}_g : \gamma_j = \gamma_g$ , for some common value  $\gamma_g$ . Thus feature effects are shrunk similarly in the same group. A small value of  $\gamma_g$  results in more shrinkage of feature effects in group  $\mathcal{G}_g$  compared to groups with a larger value.

Thus the prior expected relevance of a group's features is encoded in the model through  $\gamma_g$ . Setting the value of this variance parameter is not straightforward in most applications. Section 3.4 proposes an empirical Bayes approach to estimate these parameters from the data.

### 3 | ESTIMATION

To describe our model fitting procedure, it is convenient to recognize the mathematical symmetry between  $y$  and  $\mathbf{x}$  in model (1), and write  $\bar{\mathbf{x}} = [\mathbf{x}^T y]^T$ ,  $\bar{\mathbf{B}} = [\mathbf{B} \boldsymbol{\beta}]$ , and  $\bar{\boldsymbol{\Psi}} = \text{diag}(\psi, \dots, \psi_p, \sigma^2)$ . Setting the dimension to  $\bar{p} = p + 1$ , we can then consider the simplified, but an equivalent form of (1):

$$\bar{\mathbf{x}} | \boldsymbol{\lambda} \sim \mathcal{N}_{\bar{p}}(\bar{\mathbf{B}}^T \boldsymbol{\lambda}, \bar{\boldsymbol{\Psi}}) \quad (6a)$$

$$\boldsymbol{\lambda} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d). \quad (6b)$$

We first consider estimation in the model with observations  $\bar{\mathbf{X}} = (\mathbf{x}_i, y_i, i = 1, \dots, n)$  with only labeled features, and then indicate the changes to include unlabeled features in Section 3.2.

#### 3.1 | Variational Bayes

The maximum likelihood estimation of model (6) is straightforward when  $n > \bar{p}$ , and many algorithms are available in the literature. In the  $\bar{p} > n$  domain, estimation is possible through penalized likelihood maximization. In the current paper, the focus is on the Bayesian model, so we refer the reader to Sections 3.1 and 3.2 of the [Supporting Information](#) for details on the maximum (penalized) likelihood estimation of (6).

Bayesian posteriors are commonly approximated through Markov chain Monte Carlo (MCMC) sampling. Sampling from the posterior of models (6) and (4) is relatively straightforward (see [Supporting Information Section 4.1](#) for a Gibbs sampler). However, due to the high dimensionality of the parameters, sampling is relatively slow. In addition, the MCMC chain showed poor mixing in all investigated applications and simulations, thus requiring a prohibitive number of samples to properly explore the posterior. Here, we avoid MCMC sampling in favor of a mean-field variational Bayes approximation to the posterior.

Variational Bayes (VB) methods search for an approximation to the posterior distribution by minimizing the Kullback–Leibler divergence of the posterior distribution to a class of distributions of a given form. Mean-field variational Bayes takes the latter class as the set of product measures given some partitioning of the parameter. In our setting, the parameter includes the latent variables  $\boldsymbol{\Lambda} = [\lambda_1 \dots \lambda_n]^T$  and is partitioned as  $(\boldsymbol{\Lambda}, \bar{\mathbf{B}}, \bar{\boldsymbol{\Psi}})$ . Thus we are looking for an approximation to the posterior density of the form

$$p(\boldsymbol{\Lambda}, \bar{\mathbf{B}}, \bar{\boldsymbol{\Psi}}_1, \dots, \bar{\boldsymbol{\Psi}}_{\bar{p}} | \bar{\mathbf{X}}) \approx q(\boldsymbol{\Lambda})q(\bar{\mathbf{B}})q(\bar{\boldsymbol{\Psi}}_1, \dots, \bar{\boldsymbol{\Psi}}_{\bar{p}}), \quad (7)$$

where the  $q$  are arbitrary densities of the corresponding argument, with a slight abuse of notation denoted by the same symbol. Even though the mean-field method allows general forms of these densities, it can be shown that given exponential families with conjugate priors, the minimizing densities  $q$  remain in the family (Blei et al., 2017). In our case, minimization leads to (see [Supporting Information Section 4](#))

$$q(\boldsymbol{\Lambda}) \stackrel{D}{=} \prod_{i=1}^n \mathcal{N}_d(\boldsymbol{\phi}_i, \boldsymbol{\Xi}), \quad (8a)$$

$$q(\bar{\mathbf{B}}) \stackrel{D}{=} \prod_{j=1}^{\bar{p}} \mathcal{N}_d(\boldsymbol{\mu}_j, \boldsymbol{\Omega}_j), \quad (8b)$$

$$q(\bar{\boldsymbol{\Psi}}_1, \dots, \bar{\boldsymbol{\Psi}}_{\bar{p}}) \stackrel{D}{=} \prod_{j=1}^{\bar{p}} \Gamma^{-1}(n/2 + d/2 + \kappa_j, \zeta_j). \quad (8c)$$

The so-called variational parameters on the right-hand side are not given in the closed form. However, a coordinate ascent algorithm leads to an iterative scheme to update the parameters given the current values to the other parameters. Standard variational computations, detailed in [Supporting Information Section 4](#), give these updates as

$$\phi_i = \left\{ \sum_{j=1}^{\bar{p}} \mathbb{E}(\bar{\psi}_j^{-1}) \left[ \mathbb{V}(\bar{\mathbf{b}}_j) + \mathbb{E}(\bar{\mathbf{b}}_j) \mathbb{E}(\bar{\mathbf{b}}_j^T) \right] + \mathbf{I}_d \right\}^{-1} \mathbb{E}(\bar{\mathbf{B}}) \mathbb{E}(\bar{\Psi}^{-1}) \bar{\mathbf{x}}_i, i = 1, \dots, n, \quad (9a)$$

$$\Xi = \left\{ \sum_{j=1}^{\bar{p}} \mathbb{E}(\bar{\psi}_j^{-1}) \left[ \mathbb{V}(\bar{\mathbf{b}}_j) + \mathbb{E}(\bar{\mathbf{b}}_j) \mathbb{E}(\bar{\mathbf{b}}_j^T) \right] + \mathbf{I}_d \right\}^{-1}, \quad (9b)$$

$$\mu_j = \left[ \mathbb{E}(\Lambda^T) \mathbb{E}(\Lambda) + n \mathbb{V}(\lambda_i) + \gamma_j^{-1} \mathbf{I}_d \right]^{-1} \mathbb{E}(\Lambda^T) \bar{\mathbf{x}}_j, j = 1, \dots, \bar{p}, \quad (9c)$$

$$\Omega_j = \mathbb{E}(\psi_j^{-1})^{-1} \left[ \mathbb{E}(\Lambda^T) \mathbb{E}(\Lambda) + n \mathbb{V}(\lambda_i) + \gamma_j^{-1} \mathbf{I}_d \right]^{-1}, j = 1, \dots, \bar{p}, \quad (9d)$$

$$\begin{aligned} \zeta_j &= \bar{\mathbf{x}}_j^T \bar{\mathbf{x}}_j / 2 - \mathbb{E}(\bar{\mathbf{b}}_j^T) \mathbb{E}(\Lambda^T) \bar{\mathbf{x}}_j + \text{tr} \left[ \mathbb{E}(\Lambda^T) \mathbb{E}(\Lambda) \mathbb{V}(\bar{\mathbf{b}}_j) \right] / 2 + n \text{tr} \left[ \mathbb{V}(\lambda_i) \mathbb{V}(\bar{\mathbf{b}}_j) \right] / 2 \\ &\quad + \mathbb{E}(\bar{\mathbf{b}}_j^T) \mathbb{E}(\Lambda^T) \mathbb{E}(\Lambda) \mathbb{E}(\bar{\mathbf{b}}_j) / 2 + n \mathbb{E}(\bar{\mathbf{b}}_j^T) \mathbb{V}(\lambda_i) \mathbb{E}(\bar{\mathbf{b}}_j) / 2 + \gamma_j^{-1} \mathbb{E}(\bar{\mathbf{b}}_j^T) \mathbb{E}(\bar{\mathbf{b}}_j) / 2 \\ &\quad + \gamma_j^{-1} \text{tr} \left[ \mathbb{V}(\bar{\mathbf{b}}_j) \right] / 2 + \nu_j, j = 1, \dots, \bar{p}, \end{aligned} \quad (9e)$$

where we slightly abuse notation and let  $\bar{\mathbf{x}}_i$  and  $\bar{\mathbf{x}}_j$  denote the  $i$ th row and  $j$ th column of  $\bar{\mathbf{X}}$ , respectively. The expectations and variances are

$$\begin{aligned} \mathbb{E}(\bar{\psi}_j^{-1}) &= (n/2 + d/2 + \kappa_j) / \zeta_j, j = 1, \dots, \bar{p}, \\ \mathbb{E}(\bar{\mathbf{b}}_j) &= \mu_j, j = 1, \dots, \bar{p}, \\ \mathbb{V}(\bar{\mathbf{b}}_j) &= \Omega_j, j = 1, \dots, \bar{p}, \\ \mathbb{E}(\Lambda) &= [\phi_1 \dots \phi_n]^T =: \Phi, \\ \mathbb{V}(\lambda_i) &= \Xi, i = 1, \dots, n. \end{aligned}$$

These formulas contain cyclic dependencies and are updated until convergence. Point estimates of the prediction rule in (5) are calculated by averaging Monte Carlo draws from the estimated VB posterior. As an alternative (not used in the simulation and application sections), we introduce a slightly faster Taylor approximation to the prediction rule in [Section 4.3 of the Supporting Information](#).

Model (1) describes a covariance matrix  $\mathbf{B}^T \mathbf{B} + \Psi$  of general form for  $\mathbf{x}$ . However, standardized data are better described by a correlation matrix. In the frequentist setting, the general covariance model is easily extended to the correlation model by restriction of the likelihood to the space of correlation matrices, which in fact is the default setting in the R package `factanal`. In the Bayesian setting, this requires either more intricate prior modeling or post hoc corrections of the posterior distribution. Here, we opt for the latter. Details are given in [Supporting Information Section 4.4](#), together with a discussion of a possible future direct correlation modeling approach.

### 3.2 | Unlabeled observations

Next, we adapt the model and estimation procedure to allow for unlabeled observations, that is, a sample  $\mathbf{x}_i$ ,  $i = n + 1, \dots, n + m$  from the same distribution as  $\mathbf{x}_i$ ,  $i = 1, \dots, n$ , but without corresponding outcome variables. As detailed in Liang et al. (2007), these extra data points can greatly benefit prediction. This can also be seen by inspection of (2), which shows that the predictions  $\mathbb{E}(\bar{y} | \bar{\mathbf{x}})$  depend on the observational model for  $\mathbf{x}$  through  $\mathbf{B}$  and  $\Psi$ . We redefine the data as  $\bar{\mathbf{X}} = (\mathbf{x}_i, y_i, i = 1, \dots, n; \mathbf{x}_i, i = n + 1, \dots, n + m)$ .

To fit the model with the extended dataset, we treat the unobserved labels  $z_i$ ,  $i = n + 1, \dots, n + m$ , of the additional  $\mathbf{x}_i$  as missing data, and either apply the expectation-maximization (EM) algorithm for (penalized) maximum likelihood

estimation or treat the labels as additional parameters in the Bayesian approach. For  $\mathbf{z} = [z_{n+1} \cdots z_{n+m}]^T$  both procedures are based on the full data likelihood  $p(\mathbf{X}, \mathbf{z}, \mathbf{y} | \mathbf{B}, \Psi)$  (Bańbura & Modugno, 2014; Liu & Rubin, 1998). Section 3.3 in the [Supporting Information](#) describes the EM algorithm for (penalized) maximum likelihood estimation. Here we focus on the Bayesian model.

In the Bayesian model, the unobserved outcomes are now included in the posterior distribution. The variational Bayes posterior approximation (7) is augmented as

$$p(\Lambda, \mathbf{B}, \bar{\psi}_1, \dots, \bar{\psi}_{\bar{p}}, \mathbf{z} | \tilde{\mathbf{X}}) \approx q(\Lambda)q(\mathbf{B})q(\bar{\psi}_1, \dots, \bar{\psi}_{\bar{p}})q(\mathbf{z}),$$

where the extra factor resulting from the unobserved labels is given by

$$\begin{aligned} q(\mathbf{z}) &= \prod_{i=n+1}^{n+m} \mathcal{N}(v_i, \chi), \text{ with} \\ v_i &= \mathbb{E}(\bar{\mathbf{b}}_{\bar{p}}^T) \mathbb{E}(\lambda_i), \\ \chi &= \mathbb{E}(\bar{\psi}_{\bar{p}}^{-1})^{-1}. \end{aligned}$$

In addition, in the parameter updates of the VB algorithm, the term  $\mathbb{1}_{j=\bar{p}} m \mathbb{V}(z_i) / 2$  is added to (9e) and all occurrences of  $\bar{\mathbf{x}}_i$  and  $\bar{\mathbf{x}}_j$  in (9) are replaced with  $\tilde{\mathbf{x}}_i$  and  $\tilde{\mathbf{x}}_j$ , where

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} & \mathbf{y} \\ & \mathbb{E}(\mathbf{z}) \end{bmatrix}, \text{ with } \mathbb{E}(z_i) = v_i$$

[Supporting Information Section 4.1](#) contains more details on the inclusion of unlabeled observations in the (approximate) Bayesian posterior computations through MCMC. Although not shown here due to brevity, the unobserved outcome approach is straightforward to extend to an unobserved features approach.

### 3.3 | Latent dimension

Although we initially assumed  $d$  to be the true latent dimension, in general, it needs to be estimated. Methods for dimension estimation are plentiful in the literature (see, e.g., Preacher et al., 2013; Zwick & Velicer, 1986). Our modest aim of accurate prediction does not require correct estimation of the latent dimension, as even the true latent dimension does not always lead to optimal predictions (Goeman, 2006). Without this requirement of correct latent dimension estimation, we resort to the simple and fast Kaiser criterion. The Kaiser criterion selects  $d$  that retains dimensions with variance contributions larger than that of the average feature  $\mathbf{x}$ . This amounts to setting  $d = \sum_{j=1}^p \mathbb{1}\{v_j > 1\}$ , with  $v_j, j = 1, \dots, p$ , the eigenvalues of the correlation matrix. That is, we set  $d$  to the number of eigenvalues of the correlation matrix of  $\mathbf{X}$  larger than one.

### 3.4 | Hyperparameters and starting values

The Bayesian model requires a choice of hyperparameter  $\gamma_g$ , that is, the feature group-specific prior variances of the loadings, and the prior parameters of the uniqueness,  $\kappa_j$  and  $\nu_j$ . Choosing  $\gamma_g$  by hand requires intricate prior expert knowledge, which might not be available. An alternative is to estimate them from the data using empirical Bayes. Or, if we know the overall scale of  $\gamma_g$ , but not the group-specific deviations, we may reparameterize as  $\gamma_g = \gamma \gamma'_g$ , fix the overall scale  $\gamma$ , and estimate the group-specific multipliers  $\gamma'_g$ .

In both empirical Bayes settings, we maximize the marginal likelihood (constrained maximization for the second approach). Direct marginal likelihood maximization requires a calculation of a  $p$ -dimensional integral for which no closed form is available. With  $p$  large (i.e., the high-dimensional setup considered here), an EM algorithm with iterations

$$\boldsymbol{\gamma}^{(k+1)} = \underset{\boldsymbol{\gamma}}{\operatorname{argmax}} \mathbb{E}_{\theta | \mathbf{y}} [\log p(\mathbf{B} | \bar{\psi}_1, \dots, \bar{\psi}_{\bar{p}}) | \boldsymbol{\gamma}^{(k)}],$$

where  $\boldsymbol{\gamma} = [\gamma_1 \cdots \gamma_G]^\top$  is computationally much more feasible. With a variational Bayes approximation of the expectation, this results in

$$\boldsymbol{\gamma}^{(k+1)} = \underset{\boldsymbol{\gamma}}{\operatorname{argmax}} \left\{ -\frac{1}{2} \sum_{g=1}^G \gamma_g^{-1} \sum_{j \in \mathcal{G}_g} \mathbb{E}(\bar{\psi}_j^{-1}) \left\{ \operatorname{tr}[\mathbb{V}(\bar{\mathbf{b}}_j)] + \mathbb{E}(\bar{\mathbf{b}}_j^\top) \mathbb{E}(\bar{\mathbf{b}}_j) \right\} - \frac{d}{2} \sum_{g=1}^G |\mathcal{G}_g| \log \gamma_g \right\},$$

which renders empirical Bayes updates

$$\gamma_g^{(k+1)} = \frac{\sum_{j \in \mathcal{G}_g} \mathbb{E}(\bar{\psi}_j^{-1}) \left\{ \operatorname{tr}[\mathbb{V}(\bar{\mathbf{b}}_j)] + \mathbb{E}(\bar{\mathbf{b}}_j^\top) \mathbb{E}(\bar{\mathbf{b}}_j) \right\}}{d|\mathcal{G}_g|}.$$

For our default  $\gamma_g = \gamma \gamma'_g$  parameterization, the updates

$$\boldsymbol{\gamma}'^{(k+1)} = \underset{\boldsymbol{\gamma}'}{\operatorname{argmax}} \left\{ -\frac{1}{\gamma} \sum_{g=1}^G \gamma_g'^{-1} \sum_{j \in \mathcal{G}_g} \mathbb{E}(\bar{\psi}_j^{-1}) \left\{ \operatorname{tr}[\mathbb{V}(\bar{\mathbf{b}}_j)] + \mathbb{E}(\bar{\mathbf{b}}_j^\top) \mathbb{E}(\bar{\mathbf{b}}_j) \right\} - \frac{d}{2} \sum_{g=1}^G |\mathcal{G}_g| \log \gamma_g' \right\},$$

subject to  $\prod_{g=1}^G \gamma_g'^{|\mathcal{G}_g|} = 1,$

are not available in closed form but are still convex and easy to compute with standard numerical optimization tools.

Empirical Bayes estimation of the  $\gamma_g$  or  $\gamma'_g$  is data dependent and does not rely on subjective arguments. In addition, empirical Bayes estimation avoids (possibly complicated) hyperpriors on the  $\gamma_g$  and  $\gamma'_g$ . A drawback is that we lose the uncertainty propagation property of the full Bayesian approach.

Prior error variance/uniqueness shapes  $\kappa_j$  and scale  $\nu_j$ , and overall prior variance  $\gamma$  are set to default values to reflect a lack of prior knowledge. Our default choice of hyperparameters should take the standardization of the data into account. Three postulates are used to select the hyperparameters: (i) we ensure that the prior expectation describes a correlation matrix model, that is,  $\forall j : \mathbb{E}_{\bar{\mathbf{B}}, \bar{\psi}_1, \dots, \bar{\psi}_p}(\bar{\mathbf{b}}_j^\top \bar{\mathbf{b}}_j + \bar{\psi}_j) = 1$ . Furthermore, (ii) the prior contributions of the error and the latent structure to the data are assumed equal, that is,  $\forall j : \mathbb{E}_{\bar{\mathbf{B}}, \bar{\psi}_1, \dots, \bar{\psi}_p}(\bar{\mathbf{b}}_j^\top \bar{\mathbf{b}}_j) = \mathbb{E}_{\bar{\mathbf{B}}, \bar{\psi}_1, \dots, \bar{\psi}_p}(\bar{\psi}_j) = 1/2$ . Lastly, (iii) the prior uniqueness variance is set to  $\mathbb{V}_{\bar{\mathbf{B}}, \bar{\psi}_1, \dots, \bar{\psi}_p}(\bar{\psi}_j) = 1$ . These three postulates together result in  $\gamma = 1/d$ ,  $\forall j : \kappa_j = 9$ , and  $\forall j : \nu_j = 4$ . As a result,  $\forall j : \mathbb{V}_{\bar{\mathbf{B}}, \bar{\psi}_1, \dots, \bar{\psi}_p}(\bar{\mathbf{b}}_j^\top \bar{\mathbf{b}}_j) = 1 + 5/(2d)$ . For  $d$  large compared to  $5/2$  (as one expects in high-dimensional settings), we have  $\mathbb{V}_{\bar{\mathbf{B}}, \bar{\psi}_1, \dots, \bar{\psi}_p}(\bar{\mathbf{b}}_j^\top \bar{\mathbf{b}}_j) \approx 1 = \mathbb{V}_{\bar{\mathbf{B}}, \bar{\psi}_1, \dots, \bar{\psi}_p}(\bar{\psi}_j)$ , so that the contributions to the prior variance of latent structure and error are approximately equal.

The iterative algorithm requires setting a starting value for the variational parameters. As a default setting, we start the algorithm with the covariance matrices  $\Xi$  and  $\Omega$ , and the covariance  $\chi$  initialized to the respective sized identities. The elements of the mean parameters  $\boldsymbol{\phi}_i$  and  $\boldsymbol{\mu}_j$  are drawn from univariate random standard Gaussian,  $\nu_i$  is set to zero, and  $\zeta_j$  is initialized to the inverse of the prior mean of the inverse of the  $\phi_j$ :  $\zeta_j = \nu_j/\kappa_j$ .

The methods described in this section are implemented in the R package `bayesfactanal` available from <https://github.com/magnusmunch/bayesfactanal>.

## 4 | SIMULATIONS

### 4.1 | Setup

To assess the potential benefit of the proposed models in the prediction of outcome  $y$  from features  $\mathbf{x}$ , a simulation is set up. The simulation setting is meant to demonstrate the potential benefit of (i) the Bayesian factor regression model in general, (ii) the inclusion of the feature structure through the empirical Bayes estimation of the  $\gamma'_g$  as explained in Section 3.4, and (iii) the use of unlabeled features in the estimation.



To that end,  $n = 50$  labeled, and  $m \in \{0, 50, 100, 200, 500\}$  unlabeled observations are drawn from model (1) and standardized after simulation. Error variance and uniqueness are set to  $\sigma^2 = 1$  and  $\forall j : \psi_j = 1$ . The number of features is fixed to  $p = 100$ . Two scenarios for the model parameters  $b_{hj}$  and  $\beta_h$  are considered:

1. The number of factors is fixed to  $d = 10$ . The  $b_{hj}$  are set so that each feature loads on two factors, and the factor is a part of 20 features (see (10), where each  $b$  denotes 10 values and the empty cells are set to zero).  $\beta_h$  is set so that the outcome loads on all factors. The features are divided into two groups  $\mathcal{G}_1 = \{1, \dots, 50\}$  and  $\mathcal{G}_2 = \{51, \dots, 100\}$ . The nonzero  $b_{hj}$  values are drawn from independent univariate centered Gaussian distributions. The variances are  $\mathbb{V}(b_{hj}) = 0.1$  for  $j = 1, \dots, 50$ , and  $\mathbb{V}(b_{hj}) = 1$  for  $j = 51, \dots, 100$ . To ensure that the proportion of variance in  $y$  explained with the factors is 0.7, all  $\beta_h$  are set to  $\beta_h = 0.242$ .
2. The second scenario fixes  $d = 40$ . The model parameters  $b_{hj}$  are drawn from independent univariate centered Gaussian distributions. The variances are  $\mathbb{V}(b_{hj}) = 0.1$  for  $j = 1, \dots, 50$ , and  $\mathbb{V}(b_{hj}) = 1$  for  $j = 51, \dots, 100$ , that is, the features are structured into two groups:  $\mathcal{G}_1 = \{1, \dots, 50\}$  and  $\mathcal{G}_2 = \{51, \dots, 100\}$ . To ensure that the proportion of variance in  $y$  explained with the factors is 0.7, all  $\beta_h$  are set to  $\beta_h = 0.242$ .

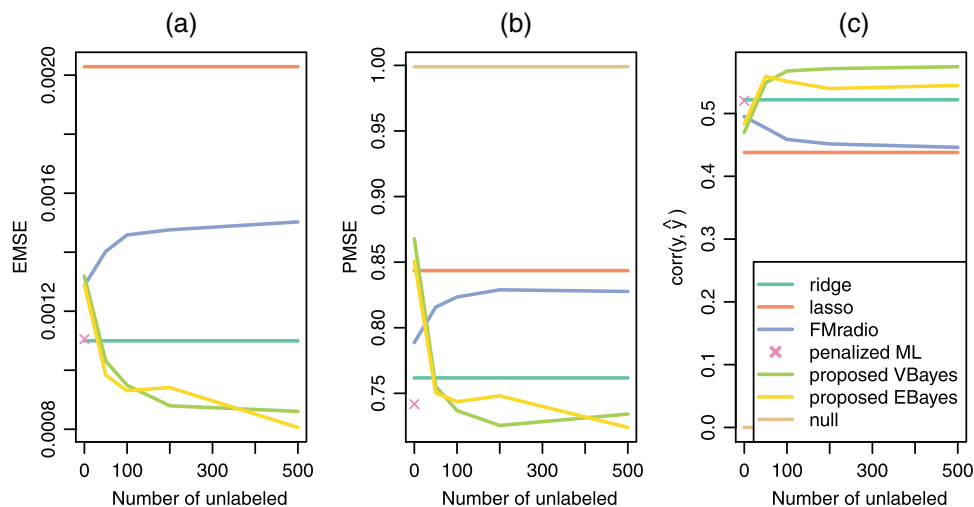
$$\mathbf{B} = \begin{bmatrix} b & & & & & & & & & b \\ & b & b & & & & & & & \\ & & b & b & & & & & & \\ & & & b & b & & & & & \\ & & & & b & b & & & & \\ & & & & & b & b & & & \\ & & & & & & b & b & & \\ & & & & & & & b & b & \\ & & & & & & & & b & b \\ & & & & & & & & & b & b \end{bmatrix}. \quad (10)$$

The first scenario models a situation where the features load on two factors only in such a way that the marginal correlation between features is weak. This might occur, for example, if genes are organized into nearly disjoint functional networks, but the outcome is related to all the networks. Ridge regression is expected to perform well here. With such a sparse loadings matrix, we have  $(\mathbf{B}^T \mathbf{B} + \Psi)^{-1} \approx \mathbf{I}_p$ . That is, the information in  $\mathbf{X}$  contributes little to the induced regression coefficients  $\tilde{\beta}$ . In addition, the induced regression coefficients become  $\tilde{\beta} \approx \mathbf{B}^T \beta = \text{Cov}(y, \mathbf{x})$ , a (rescaled version of the) quantity that standard linear regression methods aim to estimate.

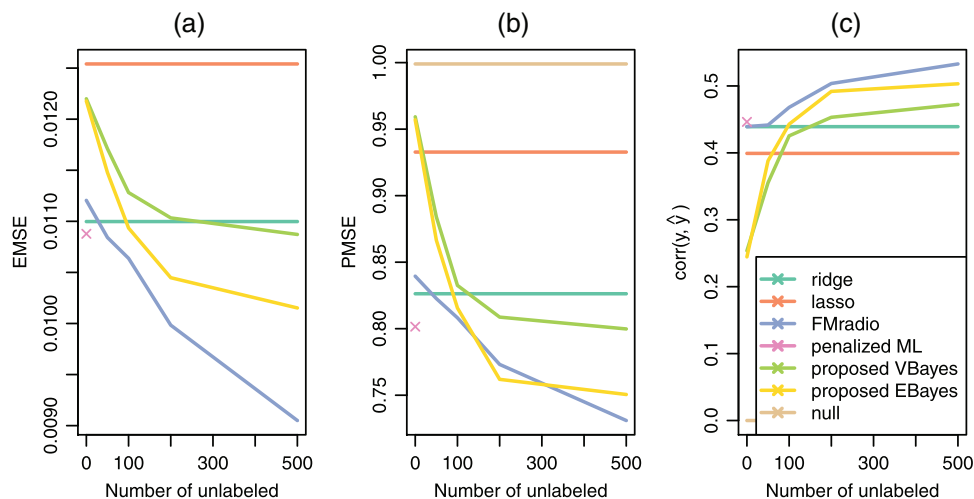
The second scenario models a setting where all features load on all factors, but the strength of the loading depends on the feature group. This might occur, for example, if genes are organized in several interconnected functional networks, but some networks have weak connections. The outcome is again related to all functional networks. In this setting, the factor regression methods are expected to perform well. In contrast to the first simulation,  $(\mathbf{B}^T \mathbf{B} + \Psi)^{-1} \neq \mathbf{I}_p$ , so information on the induced regression coefficients  $\tilde{\beta}$  is contained in  $\mathbf{X}$ . This results in increased efficiency due to the inclusion of data. Also,  $\tilde{\beta}$  is a weighted version of  $\text{Cov}(y, \mathbf{x})$  that is not straightforward to estimate with standard linear regression methods.

Six models are compared:

1. Ridge regression with a cross-validated penalty parameter with the R package `glmnet` (Friedman et al., 2010).
2. Lasso regression with a cross-validated penalty parameter with the R `glmnet` package (Friedman et al., 2010).
3. A two-step factor regression method: (i) a penalized factor model is estimated from the feature correlation matrix, with a cross-validated penalty parameter. Next, (ii) outcomes are regressed on the feature factor scores  $\hat{\mathbb{E}}(\lambda_i | \mathbf{x}_i)$  to obtain the prediction rule. This approach was shown to work in Peeters et al. (2019) and is implemented in the R `FMradio` package (Peeters et al., 2019).
4. A penalized factor regression model that includes unlabeled observations, with a cross-validated penalty parameter and estimated as described in [Supporting Information Section \(3\)](#).



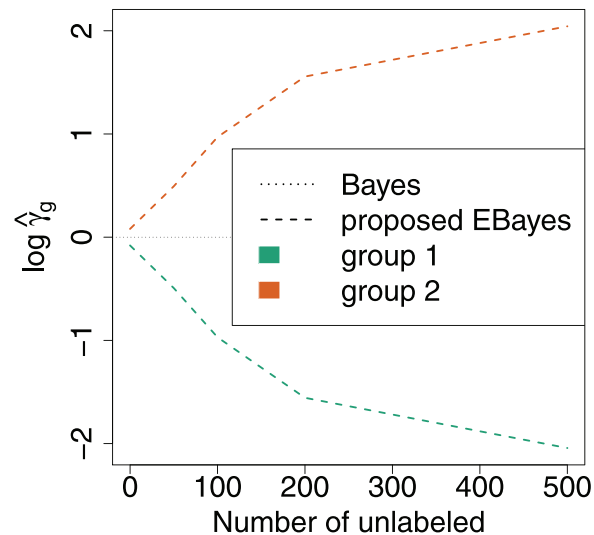
**FIGURE 2** Simulation results for Scenario 1 with median (a) EMSE, (b) PMSE, and (c) correlation between predictions and true values. The results are for several unlabeled sample sizes and consist of the methods: ridge, lasso, FMradio, penalized factor model (penalized), the proposed method without (VBAYes) and with (EBAYes) empirical Bayes, and the intercept-only model (null)



**FIGURE 3** Simulation results for scenario 2 with median (a) EMSE, (b) PMSE, and (c) correlation between predictions and true values. The results are for several unlabeled sample sizes and consist of the methods: ridge, lasso, FMradio, penalized factor model (penalized), the proposed method without (Bayes) and with (EBAYes) empirical Bayes, and the intercept-only model (null)

5. The proposed Bayesian factor regression model (4) is approximated with variational Bayes as in Section 3. The fixed hyperparameters are described in Section 3.4. Note that this model does not include an external feature structure and therefore does not estimate  $\gamma'_g$ .
6. The proposed empirical Bayesian factor regression model (4) is approximated with variational Bayes as in Section 3. The hyperparameters are described in Section 3.4, where we include the grouping of the features and estimate group-specific  $\gamma'_g$  by empirical Bayes.

For all models, the data are standardized before estimation, as is common in most real data applications. Models 3–6 allow for the inclusion of unlabeled features and are estimated for a range of numbers of unlabeled features. In addition, we fitted an intercept-only null model. We calculate the estimation mean squared error (EMSE) of  $\tilde{\beta}$ , prediction mean squared error (PMSE), and correlation between predictions and observations ( $\text{Cor}(y, \hat{y})$ ) on test data of size  $n_{\text{test}} = 1000$ . Lower PMSE and EMSE indicate better performance, while higher  $\text{Cor}(y, \hat{y})$  indicates better performance. The results, with the median taken over 50 simulation replications, are displayed in Figures 2 and 3, for scenarios 1 and 2, respectively.



**FIGURE 4** Simulation results for scenario 1 with median  $\log \hat{\gamma}'_g$  estimated with empirical Bayes according to the proposed method for the two feature groups. The VBayes method does not estimate  $\gamma'_g$ , so the resulting  $\log \hat{\gamma}'_g$  are denoted with the zero line (which corresponds to setting them to 1)

## 4.2 | Results

In both scenarios, the penalized factor regression model was not estimable with unlabeled data, due to nonconvergence. In both scenarios, the estimation (i.e., EMSE) and prediction calibration (i.e., PMSE) of the proposed Bayesian methods initially improve with more unlabeled data. However, in scenario 1 it starts deteriorating again after about  $m = 100$ . In scenario 2, where the performance continues to improve with more unlabeled data, the rate of improvement decreases with the number of unlabeled observations. This is unsurprising, as estimators generally converge at a similarly shaped  $\sqrt{n}$  rate. In both scenarios, discrimination (i.e.,  $\text{Cor}(y, \hat{y})$ ) keeps improving with the addition of unlabeled features. For scenario 1, this is surprising, considering the eventual deterioration in calibration and estimation.

In scenario 1, the proposed Bayesian methods outperform the frequentist methods for almost all  $m$  in terms of estimation and discrimination. For these methods, prediction calibration is better with medium  $m$  compared to smaller or larger  $m$ . The two-step factor regression model FMradio performs worse than the Bayesian factor regression methods and ridge, only outperforming lasso. In scenario 2, the frequentist methods outperform the proposed Bayesian method for small  $m$  in terms of estimation and calibration. For medium  $m$ , the Bayesian methods outperform ridge, and eventually, for large  $m$ , also lasso. FMradio outperforms all other methods in estimation, calibration, and discrimination. Scenario 2 simulates strong factors that explain much of the data. Extraction of these factors in step one of the FMradio approach is therefore relatively easy. Estimation of the prediction rule based on these strong factors in step two of FMradio then results in a strong predictor.

A comparison of full Bayes and empirical Bayes shows that the inclusion of the feature groupings helps in both estimation and prediction. In scenario 1, empirical Bayes estimation and calibration are comparable to full Bayes. Discrimination is slightly worse. In scenario 2, empirical clearly outperforms full Bayes in all three performance measures. Figures 4 and 5 display the estimated  $\log \hat{\gamma}'_g$  for the empirical Bayes model in scenarios 1 and 2, respectively. Both figures show a clear influence of the feature grouping on estimation, as the prior variances of the groups show a clear difference. Furthermore, the influence of the feature grouping grows with the number of unlabeled observations, as the diverging lines indicate.

## 5 | APPLICATIONS

### 5.1 | Influenza vaccine

The data described in this section are from Nakaya et al. (2011) and made publicly available through the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) archive (Barrett et al., 2012) with accession numbers

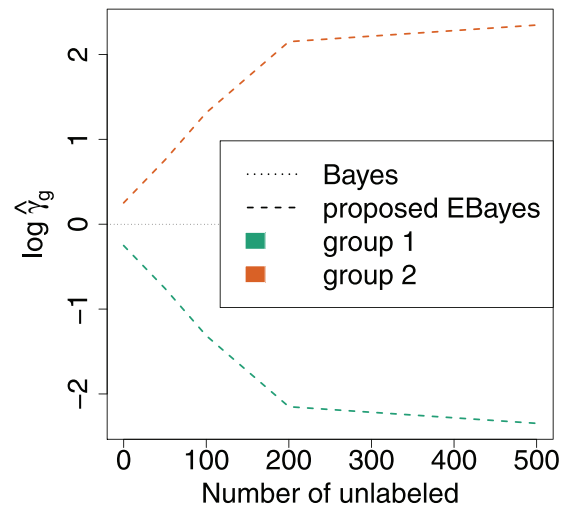


FIGURE 5 Simulation results for scenario 2 with median  $\log \hat{\gamma}'_g$  estimated with empirical Bayes according to the proposed method for the two feature groups. The VBayes method does not estimate  $\gamma'_g$ , so the resulting  $\log \hat{\gamma}'_g$  are denoted with the zero line (which corresponds to setting them to 1)

GSE29614 and GSE29617. The analysis mostly follows Van Deun et al. (2018), where the main aim was to predict vaccine efficacy with microarray gene expression data. Here follows a short description of the data; for more details, we refer the reader to Van Deun et al. (2018).

The data are from 9 and 26 subjects, observed during the 2007 and 2008 flu seasons, respectively. For all subjects, there are three efficacy measures available from a baseline measurement, just before the vaccination, and from 28 days after vaccination. The efficacy measurements are in the form of three different plasma hemagglutination inhibition antibody titers. The antibody titers were combined into one efficacy measure by first subtracting the log-transformed antibody titers at baseline from the measurement at 28 days after vaccination and subsequently taking the maximum of the three log-transformed differences. These steps were included to reduce the influence of subjects who started with high antibody concentrations due to previous infection. The scores were standardized to mean zero and variance one.

In addition to the vaccine efficacy measures, there are 54,675 microarray gene expression measurements available from a baseline just before vaccination and measurements 3 days after vaccination. The Robust multichip average algorithm (Irizarry, 2003) was used to preprocess the microarrays. After preprocessing, a change score was calculated by subtracting the baseline measurements from the measurement 3 days after vaccination. These scores were standardized to mean zero and variance one. Before the analysis, a preselection of 416 genes with the highest coefficient of variation was made. The selection of 416 genes follows the analysis results of Van Deun et al. (2018). Here, we consider the 2007 data as unlabeled and the 2008 data as labeled.

The application is an example of a difficult high-dimensional prediction problem, with little data available: a situation that regularly arises in practice. Here, the available unlabeled data potentially increase the predictive performance significantly. Additionally, genes are often considered to be organized in functional networks, so the factor model is an appropriate choice and we expect the factor regression methods to outperform classical linear regression methods.

We estimate the same models as in Section 4, with the exception of the empirical Bayes model, because there is no grouping of the features available. To assess performance, we calculated leave-one-out cross-validated PMSE and  $\text{Cor}(y, \hat{y})$  and display them in Table 1, where null refers to the intercept-only model. The penalized factor regression model did not converge, so it is not included in the results.

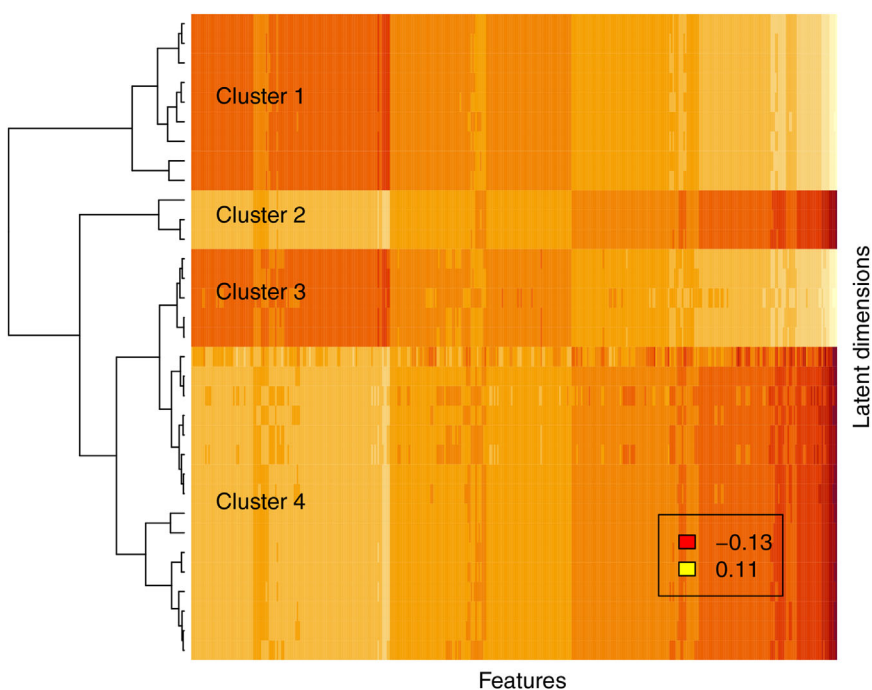
Table 1 shows that the variational Bayesian factor regression that includes the unlabeled data outperforms the other methods in terms of calibration (i.e., PMSE) and discrimination (i.e.,  $\text{Cor}(y, \hat{y})$ ), according to expectation. The other methods perform similarly in terms of PMSE, while lasso performance approaches the Bayesian factor regression in terms of  $\text{Cor}(y, \hat{y})$ . The estimated numbers of factors are 25 and 33, without and with unlabeled data, respectively.

The results indicate that, in general, prediction of vaccine efficacy from changes in gene expression is difficult. Among all methods, the largest correlation between observed and predicted efficacies is 0.341. We note that this may be due to a lack of data. There are only 26 labeled observations available.

**TABLE 1** Cross-validated PMSE and  $\text{Cor}(y, \hat{y})$  (best performing in bold) calculated on the influenza vaccine data, for the ridge, lasso, FMradio methods, the proposed method without empirical Bayes (VBayes), and the intercept-only model (null)

	PMSE	$\text{Cor}(y, \hat{y})$
Ridge	0.959	0.171
Lasso	0.929	0.339
FMradio	0.955	0.097
Proposed VBayes	<b>0.866</b>	<b>0.341</b>
Null	0.962	0

**FIGURE 6** Heatmap of the variational Bayesian posterior mean of the factor loadings for the 416 features in the influenza application. Clustering of the latent factors is based on the unweighted pair group method with arithmetic mean (UPGMA) clustering method (Sokal & Michener, 1958)



The proposed variational Bayesian factor regression results in 33 latent factors. The corresponding posterior means of the factor loadings for the features are displayed in Figure 6. The figure shows clear distinct clusters of features in terms of their factor loadings. Clustering of the latent features into four clusters leads to a clear relationship between the clusters and the posterior mean of the response variable factor loadings in Figure 7. This is a strong indication that the model results in meaningful and interpretable latent factors that relate to features and responses.

## 5.2 | Oral cancer lymph node metastasis

In this section, oral cancer lymph node metastasis is predicted with gene expression data. Sequenced ribonucleic acid (RNAseqs), taken from TCGA (The Cancer Genome Atlas Network, 2015), are measured on 133 human papilloma virus (HPV)-negative oral tumors taken from 76 and 57 oral cancer patients, with and without lymph node metastasis, respectively. For more details on these data, see te Beest et al. (2017). Additional gene expression data are available from an independent microarray study on 97 oral cancer patients in Mes et al. (2017). These microarrays are normalized to the same scale as the RNAseqs and included in the analysis as unlabeled data. A preselection of 871 genes is done using the cutoff for  $p$ -values  $p \leq 0.01$ , comparing the microarrays between metastatic and nonmetastatic patients. To investigate the empirical Bayes estimation of the  $\gamma_g$ , the genes are divided into three groups based on the cis-correlation between the RNAseq data and TCGA deoxyribonucleic acid (DNA) copy numbers on the same patients, quantified by Kendall's  $\tau$ .

As before, genes are assumed to be organized into functional modules, so we expect the factor regression methods to fit the data well. We expect features with a large positive correlation between RNAseq and DNA copy number, as quantified

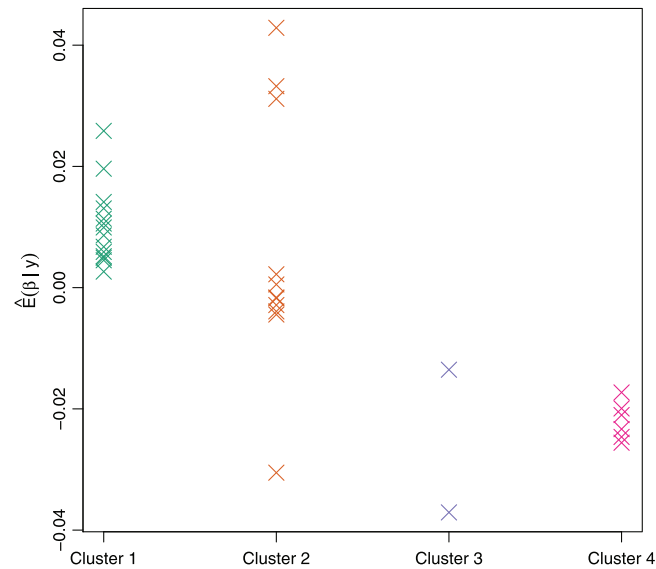


FIGURE 7 Posterior mean of the response variable factor loadings per cluster of response variable factor loadings, based on a clustering of the latent factors, for the influenza application

TABLE 2 BSS and AUC (best performing in bold) calculated on the oral cancer lymph node metastasis data, for the methods ridge, lasso, FMradio, and the proposed method without (VBayes) and with (EBayes) empirical Bayes

	BSS	AUC
Ridge	0.125	0.698
Lasso	<b>0.132</b>	0.708
FMradio	0.014	0.66
Proposed VBayes	0.099	0.746
Proposed EBayes	0.101	<b>0.748</b>

by Kendall's  $\tau$ , to be more important for metastasis prediction. We therefore expect to estimate larger  $\gamma'_g$  for the groups with higher Kendall's  $\tau$ .

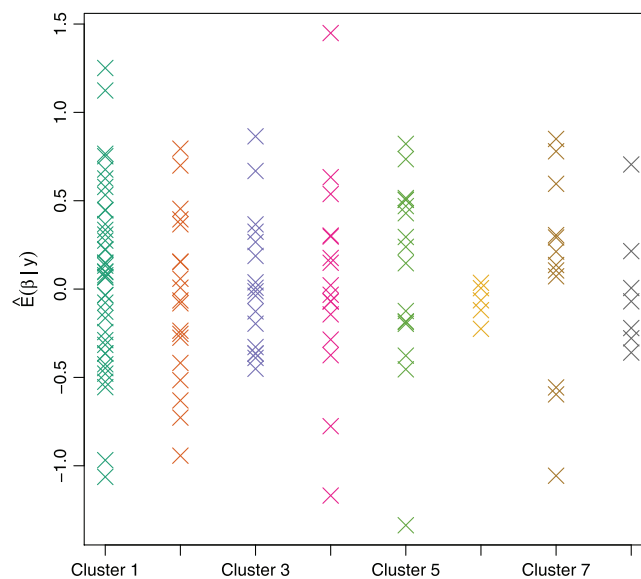
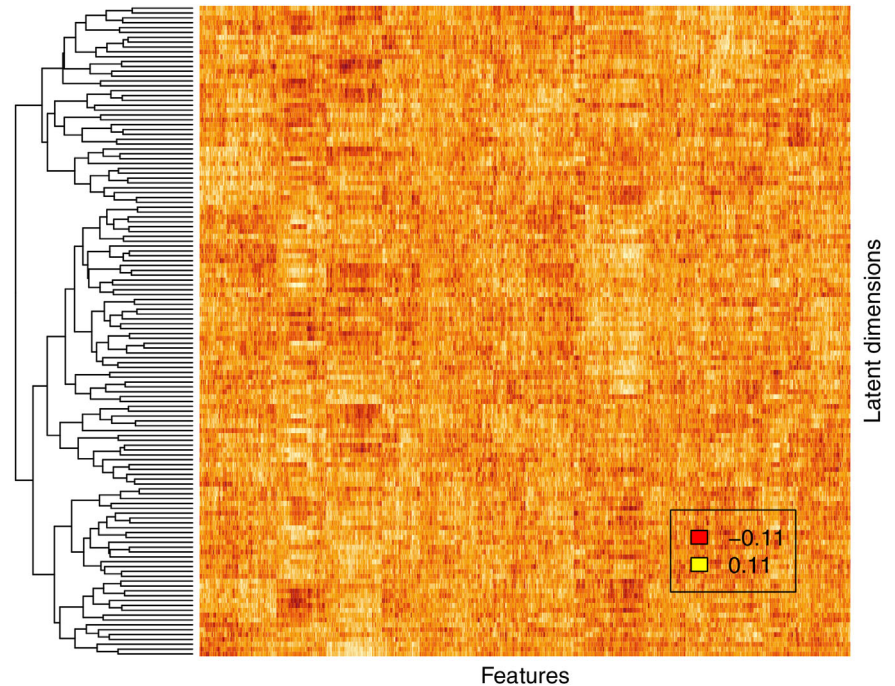
We estimate the logistic extensions of the models estimated in Section 4. To assess performance we calculated a calibration measure of the Brier skill score (BSS) and discrimination measure of the area under the receiver operator curve (AUC) on the unlabeled data and display them in Table 2. The penalized factor regression model did not converge, so it is not included in the results.

Here, the best performing model in terms of calibration (i.e., BSS) is the lasso. The Bayesian factor regression methods outperform the other methods in terms of discrimination (i.e., AUC). The estimated  $\gamma'_g$  are 0.97, 0.98, and 1.01 for the low-, medium-, and high cis-correlation groups, respectively. This small difference in shrinkage leads to a marginal increase in the predictive performance of the empirical Bayes method compared to the full Bayes version. The estimated numbers of factors are 113 and 134, without and with unlabeled data, respectively.

Generally, we see that AUC is relatively high (as compared to the random model with an AUC of 0.5), and we are able to discriminate reasonably well between oral cancer patients with and without lymph node metastasis using RNAseq data. In contrast, BSSs are generally low, indicating that calibration is difficult to improve on compared to the empty, intercept-only model.

The proposed variational Bayesian factor regression results in 134 latent factors. The corresponding posterior means of the factor loadings for the features are displayed in Figure 8. In contrast to the influenza application, no clear distinguishable clusters of features are found. If we set the number of clusters to eight (approximately the same ratio of clusters to latent features as in the influenza application), the relationship between the clusters and the posterior mean of the response variable factor loadings in Figure 9 is not evident. A clear interpretation of the results therefore requires more intricate biological knowledge of the problem, which is beyond the scope of this paper.

**FIGURE 8** Heatmap of the variational Bayesian posterior mean of the factor loadings for the 871 features in the oral cancer metastasis application. Clustering of the latent factors is based on the UPGMA clustering method (Sokal & Michener, 1958)



**FIGURE 9** Posterior mean of the response variable factor loadings per cluster of response variable factor loadings, based on a clustering of the latent factors, for the oral cancer metastasis application

## 6 | DISCUSSION

This paper investigates a Bayesian factor regression model for high-dimensional prediction and classification problems. It allows for the inclusion of unlabeled data and feature groupings to improve predictive performance. Estimation is done through a combination of variational and empirical Bayes techniques. The approach is competitive with classical ridge and lasso regression, as well as with more elaborate frequentist factor modeling approaches such as penalized factor regression and the two-step factor FMradio. Simulations show that the method is especially useful if the features are generated in dense, correlated networks. Two applications show that the method predicts just as well, or better, than existing methods in real data settings.

A technical advantage of the pursued factor modeling approach is the straightforward inclusion of unlabeled observations through the full likelihood approach. However, some caution regarding this approach is advised. For the full

likelihood approach to return unbiased estimates, the missing data mechanism is assumed to be missing at random (MAR). That is, the missingness possibly depends on the observed features but not on unobserved features. In the current setting, MAR implies that unobserved labels are not missing due to the value of the labels. We argue that in most applications, this is a reasonable assumption. In the examples above, observations are unlabeled because they come from independent studies. Due to the independence, it is reasonable to assume that no relation exists between not observing labels and the actual labels. Sections 4 and 5 show that the frequentist factor models suffer from convergence issues if the number of labeled and/or unlabeled samples becomes large. More investigation is required to determine when and why these convergence issues occur. An inherent benefit of Bayesian modeling is the uncertainty quantification that automatically comes with the Bayesian posterior. This allows for the straightforward calculation of prediction intervals. We should note, however, that uncertainty quantification in the current setting requires a more thorough investigation. A limitation of the method observed in the simulation results (Figure 2) is that despite the increasing discriminative performance of the model, calibration may deteriorate slightly with more extreme ratios of unlabeled to labeled number of observations.

More elaborate prior modeling of the factor loadings is possible through  $\gamma_j$ . For example, a more sparse lasso model for the factor loadings introduces the hyperpriors:  $\gamma_j \sim \text{Exp}(\lambda_j)$ . Feature grouping is then included by parameterizing  $\forall j \in \mathcal{G}_g : \lambda_j = \lambda_g$  and estimating  $\lambda_g$  with empirical Bayes. In general, such Gaussian-scale mixture extensions of the  $\bar{\mathbf{B}}$  prior require the addition of one or more extra layers to the prior and one or more extra variational parameters to update during estimation. Some existing examples of sparse Bayesian factor models are Ferrari and Dunson (2020) and Carvalho et al. (2008). Sparse factor models often simplify latent dimension estimation. In any case, latent dimension estimation is a topic that deserves more attention. Here, estimation is via a simple Kaiser criterion. More elaborate methods are available in the literature (see, e.g., Auerswald & Moshagen, 2019).

Lastly, we give some indication of computational times. The proposed factor regression approaches are slower to estimate compared to the other methods. Model estimation times for the influenza application are 0.87, 0.16, 5.27, and 36.46 seconds, for the ridge, lasso, FMradio, and Bayesian factor regression models, respectively. For the oral cancer metastasis application, we have 2.76, 0.98, and 54.82 seconds for the ridge, lasso and FMradio, respectively, and 124.68 and 245.18 minutes for the variational and empirical Bayesian models. Especially in the second application, the estimation is considerably slower. However, we argue that these times are still manageable and much faster than traditional MCMC estimation times.

## SOFTWARE

The code used in this paper is implemented as a (developmental) R package and is available from <https://github.com/magnusmunch/bayesfactanal>

## ACKNOWLEDGMENT

We thank the editor and referees for their useful suggestions and comments.


## CONFLICT OF INTEREST

The authors have declared no conflict of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the Supporting Information of this article.

## OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

## REFERENCES

Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods, 24*, 468–491.



- Avalos-Pacheco, A., Rossell, D., & Savage, R. S. (2022). Heterogeneous large datasets integration using Bayesian factor regression. *Bayesian Analysis*, 17(1), 33–66.
- Bañbura, M., & Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data: ML for factor models with missing data. *Journal of Applied Econometrics*, 29, 133–160.
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2012). NCBI GEO: archive for functional genomics data sets—Update. *Nucleic Acids Research*, 41, D991–D995.
- Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112, 859–877.
- Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., & West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103, 1438–1456.
- Ferrari, F., & Dunson, D. B. (2020). Bayesian factor analysis for inference on interactions. arXiv. <https://arxiv.org/abs/1904.11603>
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.
- Goeman, J. J. (2006). *Statistical methods for microarray data: Pathway analysis, prediction methods and visualization tools* (Ph.D. thesis). Leiden University, Leiden, The Netherlands.
- Irizarry, R. A. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4, 249–264.
- Leday, G. G. R., Gunst, M. C. M. d., Kpogbezan, G. B., Vaart, A. W. v. d., Wieringen, W. N. v., & van de Wiel, M. A. (2017). Gene network reconstruction using global-local shrinkage priors. *The Annals of Applied Statistics*, 11, 41–68.
- Liang, F., Mukherjee, S., & West, M. (2007). The use of unlabeled data in predictive modeling. *Statistical Science*, 22, 189–205.
- Liu, C., & Rubin, D. B. (1998). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Statistica Sinica*, 8, 729–747.
- Mes, S. W., te Beest, D., Poli, T., Rossi, S., Scheckenbach, K., van Wieringen, W. N., Brink, A., Bertani, N., Lanfranco, D., Silini, E. M., van Diest, P. J., Bloemena, E., Leemans, C. R., van de Wiel, M. A., & Brakenhoff, R. H. (2017). Prognostic modeling of oral cancer by gene profiles and clinicopathological co-variables. *Oncotarget*, 8, 59312–59323.
- Mes, S. W., van Velden, F. H. P., Peltenburg, B., Peeters, C. F. W., te Beest, D. E., van de Wiel, M. A., Mekke, J., Mulder, D. C., Martens, R. M., Castelijns, J. A., Pameijer, F. A., de Bree, R., Boellaard, R., Leemans, C. R., Brakenhoff, R. H., & de Graaf, P. (2020). Outcome prediction of head and neck squamous cell carcinoma by MRI radiomic signatures. *European Radiology*, 30, 6311–6321.
- Münch, M. M., Peeters, C. F. W., Van Der Vaart, A. W., & Van De Wiel, M. A. (2021). Adaptive group-regularized logistic elastic net regression. *Biostatistics*, 22, 723–737.
- Nakaya, H. I., Wrammert, J., Lee, E. K., Racioppi, L., Marie-Kunze, S., Haining, W. N., Means, A. R., Kasturi, S. P., Khan, N., Li, G.-M., McCausland, M., Kanchan, V., Kokko, K. E., Li, S., Elbein, R., Mehta, A. K., Aderem, A., Subbarao, K., Ahmed, R., & Pulendran, B. (2011). Systems biology of vaccination for seasonal influenza in humans. *Nature Immunology*, 12, 786–795.
- Peeters, C. F. W. (2019). FMradio: Factor modelling for radiomics data. <https://CRAN.R-project.org/package=FMradio>
- Peeters, C. F. W., Übelhör, C., Mes, S. W., Martens, R., Koopman, T., de Graaf, P., van Velden, F. H. P., Boellaard, R., Castelijns, J. A., Beest, D. E. T., Heymans, M. W., & van de Wiel, M. A. (2019). *Stable prediction with radiomics data*. arXiv. <https://arxiv.org/abs/1903.11696>
- Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, 48, 28–56.
- Sokal, R., & Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409–1438.
- te Beest, D. E., Mes, S. W., Wilting, S. M., Brakenhoff, R. H., & Van De Wiel, M. A. (2017). Improved high-dimensional prediction with Random Forests by the use of co-data. *BMC Bioinformatics*, 18, 584.
- The Cancer Genome Atlas Network. (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517, 576–582.
- van de Wiel, M. A., Lien, T. G., Verlaat, W., van Wieringen, W. N., & Wilting, S. M. (2016). Better prediction by use of co-data: Adaptive group-regularized ridge regression. *Statistics in Medicine*, 35, 368–381.
- Van Deun, K., Crompvoets, E. A. V., & Ceulemans, E. (2018). Obtaining insights from high-dimensional data: sparse principal covariates regression. *BMC Bioinformatics*, 19, Article 104.
- van Nee, M. M., Wessels, L. F., & van de Wiel, M. A. (2021). Flexible co-data learning for high-dimensional prediction. *Statistics in Medicine*, 40(26), 5910–5925.
- West, M. (2003). Bayesian Factor Regression Models in the “Large p, Small n” Paradigm. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (Eds.), *Bayesian Statistics 7* (pp. 723–732). Oxford University Press.
- Zhu, X., & Goldberg, A. B. (2009). *Introduction to semisupervised learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool.
- Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99, 432–442.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Münch, M. M., van de Wiel, M. A., van der Vaart, A. W., & Peeters, C. F. W. (2022). Semi-supervised empirical Bayes group-regularized factor regression. *Biometrical Journal*, 64, 1289–1306.  
<https://doi.org/10.1002/bimj.202100105>