INTEGRATIVE FACTOR MODELS FOR BIOMEDICAL APPLICATIONS

Alejandra Avalos-Pacheco [1] and Roberta De Vito [2]

[1] Applied Statistics Research Unit, TU Wien, Vienna, Austria, and
Harvard-MIT Center for Regulatory Science, Harvard University, Boston, MA,
(e-mail: alejandra.avalos@tuwien.ac.at)

[2] Dept. of Biostatistics, School of Public Health, Brown University, Providence, RI,
(e-mail: roberta_devito@brown.edu)

ABSTRACT: Data-integration of multiple studies is key to understanding and gaining knowledge in statistical research. However, such data present artifactual sources of variation, also known as covariate effects. Covariate effects can be complex and can lead to systematic biases. If not corrected, these biases may lead to unreliable inferences. Here, we will present novel sparse latent factor regression and multi-study factor regression models to integrate heterogeneous data.

KEYWORDS: factor regression, multi-study factor analysis, sparsity, non-local priors, scalable algorithms

1 Introduction

Data integration is crucial when separate data sources are collected on the same phenomenon. For instance, different economical studies may test the efficacy of several policy-making interventions; clinical trials may analyze various treatments using data gathered at different times. Integrative models provide gains in statistical power and help to take accurate decisions sooner. However, a lack of appropriate integration tools could lead to unreliable inference.

Data integration in biomedicine is particularly challenging as some measurements reappear across different studies. However, high throughput experiments display both biological and artifactual sources of variation. Here, we will present novel sparse factor regression and multi-study factor regression models to integrate such heterogeneous data.

The factor regression (FR) model (Avalos-Pacheco, 2018, Avalos-Pacheco *et al.*, 2022) provides a tool for data exploration via dimensionality reduction and sparse low-rank covariance estimation while correcting for a range of covariate, or artifactual, effects, such as batch effects. A limitation of FR models is the inability to isolate the study-specific latent structure.

The multi-study factor analysis (De Vito *et al.*, 2019, De Vito *et al.*, 2021) is able to handle multiple high-throughput experiments, simultaneously achieving two goals: a) to capture common component(s) across studies and b) to isolate the sources of variation that are unique of each study. We generalize the multi-study factor analysis by adopting a factor regression approach. Our proposed multi-study factor regression (MSFR) will enable us to jointly obtain the group-specific covariances and the common component.

In the conference presentation, we will discuss the use of several sparse priors, local and non-local (Johnson & Rossell, 2010), for learning the dimension of the latent factors. Our approaches provide a flexible methodology for sparse factor regression, which is not limited to data with covariate effects. Our models are fitted via scalable expectation–maximization (EM) algorithms.

We will also show the usefulness of our methods by presenting several examples, with a focus on bioinformatics applications. For all the examples, we give a visual representation of the latent factors of the data. Thereafter, in the case of cancer genomics data sets, we provide survival predictions leveraging the obtained factors; in the case of a Hispanic community health nutritional-data study, we obtain dietary patterns, associating each factor with a measure of overall diet quality related to cardiometabolic disease risk.

2 Model specification

We follow the model proposed in Avalos-Pacheco *et al.*, 2022. We consider vectors $\mathtt{x}_{is} = (x_{i1s}, x_{i2s}, \ldots, x_{ips})^\top \in \mathbb{R}^p$, observed for $i = 1, \ldots, n$ individuals in study $s, s = 1, \ldots, S$. The factor regression model defines $\mathtt{x}_{is}$ as a regression on p_v observed covariates $\mathtt{v}_{is} \in \mathbb{R}^{p_v}$, and q low-dimensional latent variables $\mathtt{f}_{is} \in \mathbb{R}^q$, also known as latent coordinates or factors $\mathtt{x}_{is} = \theta \mathtt{v}_{is} + \Phi \mathtt{f}_{is} + \mathtt{e}_{is}$, where $\theta \in \mathbb{R}^{p \times p_v}$ is the matrix of regression coefficients, $\Phi \in \mathbb{R}^{p \times q}, q \ll p$, is the loading matrix, $\mathtt{e}_{is} \in \mathbb{R}^p$ is the error, distributed as $\mathtt{e}_{is} \sim N(0, \mathcal{T}_s^{-1})$ independently across $i = 1, \ldots, n$, with $\mathcal{T}_s^{-1} = \text{diag}\{1/\tau_{ls}, l = 1, \ldots, p\}$ as the idiosyncratic precision matrix for study s. Factors are assumed to be standard normal, $\mathtt{f}_{is} \sim N(0, \mathbf{I})$, independent across $i = 1, \ldots, n$ and independent of $\mathtt{e}_{is}$.

We first set priors for the precisions $\tau_{ls} \mid \eta, \xi \sim$ Gamma$(\eta/2, \eta\xi/2)$,, and regression parameters $\theta \sim N(0, \psi \mathbf{I})$. The loadings $\Phi = \{\phi_{jk}, j = 1, \ldots, p, k = 1, \ldots, q\}$ play a key part in factor models as they allow us to improve shrinkage and simplify interpretation. Here, we set a non-local spike-and-slab prior on ϕ_{jk}, as in Avalos-Pacheco *et al.*, 2022. This prior distinguishes the loading elements that should be included, modelled by the slab component, from those that should be excluded, modelled by the spike component. We consider a mixture distribution with a product moment non-local prior (Johnson & Rossell, 2010) for the slab components and a normal prior for the spike components:

$$\mathsf{p}(\phi_{jk} \mid \gamma_{jk}) = (1-\gamma_{jk})\mathrm{N}(\phi_{jk};0,\lambda_0) + \gamma_{jk}\frac{\phi_{jk}^2}{\lambda_1}\mathrm{N}(\phi_{jk};0,\lambda_0). \tag{1}$$

We set a hierarchical prior over the latent indicator $\gamma_{jk} \mid \zeta_k \sim \text{Bernoulli}(\zeta_k)$, $\gamma_{jk} \mid \zeta_k \sim \text{Beta}\left(\frac{a_\zeta}{k}, b_\zeta\right), j = 1,\ldots,p, k = 1,\ldots,q$.

Inference is done by an efficient EM algorithm with closed-form expressions. We refer to Avalos-Pacheco *et al.*, 2022, for details, prior elicitation, parameter initialization, post-processing and description of the EM algorithm.

3 Pancreatic cancer

To quantify the effectiveness of our approach, we study an unpublished gene expression data set for individuals with pancreatic cancer. We analyze two studies collected under different experimental conditions and sizes ($n_1 = 27$ and $n_2 = 183$). We select the 5% genes with the highest total variance across all samples ($p = 1,177$ genes). We normalize the data to have zero mean and unit variance and included the type of tissue (normal or tumour) and a study indicator as covariates for our model. In order to evaluate the effect of the non-local prior, we compare our model (FR-NLSS) with methods that use a normal spike-and-slab prior (George & McCulloch, 1993) (FR-LSS), instead of our proposed non-local spike-and-slab prior, and that do not leverage any sparse inducing priors (FR-NS). Since the data generating ground truth is unknown, we assess the performance of our estimators by evaluating the cross-validated log likelihood. Table 1 presents the results from 10 independent runs of 10-fold cross-validation. It displays the selected number of factors $\widehat{q}$, the number of estimated non-zero loadings $||\widehat{\Phi}||_0$ and the cross-validated loglikelihood.

Table 1. *Cross-validated log-likelihood analysis for pancreatic cancer dataset.*

	$\widehat{q}$	$\|\|\widehat{\Phi}\|\|_0$	Log-likelihood
FR-NS	100.0	117,700	-1,644.8
FR-LSS	63.0	74,151	-1,622.0
FR-NLSS	19.0	22,363	-1,157.6

The results in Table 1 show that our proposed FR-NLSS obtained a better out-of-sample log-likelihood with fewer factors and sparser Φ than our competitors. Thus, we conclude that FR-NLSS reconstructed the data better than the other methods.

4 Extensions

We extend the FR Model to the Multi-study factor setting (De Vito *et al.*, 2019, De Vito *et al.*, 2021). We refer to this generalization as the Multi-study factor regression (MSFR) (De Vito & Avalos-Pacheco, 2023+).

Marginally, the underlying covariance of x_{is} of the FR Model is $\Sigma_s = \Phi\Phi^\top + \mathcal{T}_s^{-1}$. In the MSFR setting, the Σ_s becomes

$$\Sigma_s = \Phi\Phi^\top + \Lambda_s\Lambda_s^\top + \mathcal{T}_s^{-1}, \tag{2}$$

where $\Lambda_s \in \mathbb{R}^{p\times q_s}, q_s \ll p$, is the study-specific loading matrix. The new Σ_s allows to explain the total variance into the variance of the common factors, the variance of the study-specific factors and the idiosyncratic error. In the conference presentation, we will discuss the FR and MSFR in detail, and we will apply our models to different gene expression and nutritional epidemiology data sets. Both our FR and MSFR will be demonstrated to be valuable to visually depict the underlying factors of the data; and to make survival predictions or to identify dietary patterns and study the embedded risk of cardiometabolic disease. We refer to De Vito & Avalos-Pacheco, 2023+ for further details.

References

AVALOS-PACHECO, A. 2018. *Factor regression for dimensionality reduction and data integration techniques with applications to cancer data.* University of Warwick, PhD thesis.

AVALOS-PACHECO, A., ROSSELL, D., & SAVAGE, R. S. 2022. Heterogeneous large datasets integration using Bayesian factor regression. *Bayesian analysis*, **17**(1), 33–66.

DE VITO, R., & AVALOS-PACHECO, A. 2023+. Multi-study factor regression model: An application in nutritional epidemiology. *arXiv:2304.13077.*

DE VITO, R., BELLIO, R., TRIPPA, L., & PARMIGIANI, G. 2019. Multi-study factor analysis. *Biometrics*, **75**(1), 337–346.

DE VITO, R., BELLIO, R., TRIPPA, L., & PARMIGIANI, G. 2021. Bayesian multistudy factor analysis for high-throughput biological data. *The annals of applied statistics*, **15**(4), 1723–1741.

GEORGE, E., & MCCULLOCH, R. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**(423), 881–889.

JOHNSON, V. E., & ROSSELL, D. 2010. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **72**(2), 143–170.