Trial Version Wondershare PDFelement

arXiv:1806.09896v1 [stat.AP] 26 Jun 2018

Bayesian Multi-study Factor Analysis for High-throughput Biological Data

Roberta De Vito,
Department of Computer Science, Princeton University
Ruggero Bellio,
Department of Economics and Statistics, University of Udine
Lorenzo Trippa
Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute
Department of Biostatistics, Harvard T. H. Chan School of Public Health
and
Giovanni Parmigiani
Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute
Department of Biostatistics, Harvard T. H. Chan School of Public Health

Abstract

This paper presents a new modeling strategy for joint unsupervised analysis of multiple high-throughput biological studies. As in Multi-study Factor Analysis, our goals are to identify both common factors shared across studies and study-specific factors. Our approach is motivated by the growing body of high-throughput studies in biomedical research, as exemplified by the comprehensive set of expression data on breast tumors considered in our case study. To handle high-dimensional studies, we extend Multi-study Factor Analysis using a Bayesian approach that imposes sparsity. Specifically, we generalize the sparse Bayesian infinite factor model to multiple studies. We also devise novel solutions for the identification of the loading matrices: we recover the loading matrices of interest ex-post, by adapting the orthogonal Procrustes approach. Computationally, we propose an efficient and fast Gibbs sampling approach. Through an extensive simulation analysis, we show that the proposed approach performs very well in a range of different scenarios, and outperforms standard Factor analysis in all the scenarios identifying replicable signal in unsupervised

Trial Version Wondershare PDFelement

genomic applications. The results of our analysis of breast cancer gene expression across seven studies identified replicable gene patterns, clearly related to well-known breast cancer pathways. An R package is implemented and available on GitHub.

Keywords: Dimension Reduction; Factor Analysis; Gene Expression; Gibbs Sampling; Meta-analysis.

Trial Version Wondershare PDFelement

1 Introduction

High-throughput assays are transforming the study of biology, and are generating a rich, complex and diverse collection of high-dimensional data sets. Joint analyses combining data from different studies and technologies are crucial to improve accuracy of conclusions and to produce generalizable knowledge.

Most measurements from high-throughput experiments display variation arising from both biological and artifactual sources. Within a study, effects driven by unique issues with the experimental conditions of a specific laboratory or technology can be so large to surpass the biological signal for many biological features (Aach et al., 2000). In gene expression, for example, large systematic differences arising from different laboratories or technological platforms have been long recognized (Irizarry et al., 2003, Shi et al., 2006, Kerr, 2007). Systematic collections of gene expression data, collected with technologies that have evolved over time, are widely available, as exemplified by the breast cancer datasets that motivate our work, described in Section 2.

A strength of multi-study analyses is that, generally, genuine biological signal is more likely than spurious signal to be present in multiple studies, particularly when studies are collected from biologically similar populations. Thus, multi-study analyses offer the opportunity to learn replicable features shared among multiple studies. Discovering these features is, broadly speaking, more valuable than discovering signal in a single study. Joint analyses of multiple genomic datasets have begun more than a decade ago, they are now increasingly common, and can be highly successful (Rhodes et al., 2002, Huttenhower et al., 2006, Gao et al., 2014b, Pharoah et al., 2013, Riester et al., 2014, Ciriello et al., 2013). Many such analyses focus on identifying parameters that relate biological features measured at high throughput to phenotypes. These effects can be replicable, though signal extraction across studies can be challenging (Garrett-Mayer et al., 2008).

Trial Version Wondershare PDFelement

An important goal in high-dimensional data analysis is the unsupervised identification of latent components or factors. Despite the importance of this goal, the development of formal statistical approaches for unsupervised multi-study analyses is relatively unexplored.

In applications, joint unsupervised analyses of high-throughput biological studies often proceed by pooling all the data. Despite their success, these studies rely critically on simplified methods of analysis to capture common signal. For example Wang et al. (2011) and Edefonti et al. (2012) stack all studies and then perform standard analyses, such as factor analysis (FA) or Principal Component Analysis (PCA). The results will capture some common features, but the information about study-specific components will likely be lost, and ignoring it could compromise the accuracy of the common factors found.

Alternatively, it is also common to analyze each study separately and then heuristically explore common structures from the results (Hayes et al., 2006). Co-Inertia Analysis (CIA) (Dray et al., 2003) explores the common structure of two different sets of variables by first separately performing dimension reduction on each set to estimate factor scores, and then investigating the correlation between these factors. Multiple Co-Inertia Analysis (MCIA) is a generalization of CIA to more than two data sets, which projects different studies into a common hyperspace (Meng et al., 2014). Multiple Factor analysis (MFA) (Abdi et al., 2013) is an extension of PCA and consists of three steps. The first step applies PCA to each study. In the second step, each data set is normalized by dividing by the first singular value of the covariance matrix. In the third step, these normalized data are stacked by row creating a single data set to which PCA is then applied.

In practice, there is a need to automatically and rigorously model across studies the common signal that can reliably be identified, while at the same time modeling study-specific variation. A methodological tool for this task is Multi-Study Factor Analysis (MSFA), recently introduced in De Vito et al. (2016). Inspired by models used in the social sciences, MSFA extends FA to the joint analysis of multiple studies, separately estimating

Trial Version Wondershare PDFelement

signal reproducibly shared across multiple studies from study-specific components arising from artifactual and population-specific sources of variation. This dual goal clearly sets MSFA aside from earlier applications of FA to gene expression studies, such as Carvalho et al. (2008), Friguet et al. (2009), Blum et al. (2010), or Runcie and Mukherjee (2013).

The MSFA methodology in De Vito et al. (2016) is limited to settings where enough samples are available in each study, and no sparsity is expected or necessary. This is because model parameters are estimated by maximum likelihood (MLE) and model selection is performed by standard information criteria. In high-throughput biology, the sample size routinely exceeds the number of variables, and it is essential to employ regularization through priors or penalties.

In this paper we introduce a Bayesian generalization of Multi-study factor analysis. Bayesian approaches naturally provide helpful regularization, and offer further advantages, discussed later. We leverage the sparse Bayesian infinite factor model, and generalize the multiplicative gamma prior of Bhattacharya and Dunson (2011) to the MSFA setting, to induce sparsity on each loading matrix. We then sample from the posterior distribution via MCMC, without any ex-ante constraints on the loading matrices. This avoids the order dependence induced by the often-used assumption of a lower-triangular form of the loading matrices (Geweke and Zhou, 1996, Lopes and West, 2004), which was employed by the original MSFA proposal. Although useful inferences can be obtained with careful implementation of the constraint, removing it makes the application of FA much simpler and general. We regard this to be an important advantage of our proposal.

Our prior and parametrization also facilitate inference on the covariance matrices and precision matrices of the observed variables. These are often important goals. An important example is inference on gene networks, often implemented by first estimating the covariance matrix through FA (Zhao et al., 2014, Gao et al., 2016). Through the estimation of common factors implied by the decomposition of the covariance matrix described in §3.1,

Trial Version Wondershare PDFelement

the approach we propose allows to detect a common network across the studies, and also to recover the study-specific contributions to gene networks.

The original implementation of the sparse Bayesian infinite factor model Bhattacharya and Dunson (2011) truncates the dimension of the loading matrices at a fixed value. In MSFA, this point is even more important, since our model introduces $(S + 1)$ loading matrices if there are S studies. We suggest a pragmatic approach, where the number of dimensions is chosen based on a simple eigenvalue decomposition of covariance matrices obtained as output of the MCMC sampling from the posterior. The specific choice of prior makes the choice of the dimension less critical than would alternative approaches, as we discuss later.

A further strength of our proposal is the recovery of the loading matrices, which are not estimated in Bhattacharya and Dunson (2011). We leverage the recently proposed Orthogonal Procrustes (OP) method, introduced in Aßmann et al. (2016). OP performs an ex-post recovery of the estimated loadings by processing the MCMC output, after fitting the model without any restrictions. The method provides a satisfactory solution to the rotation invariance of FA. Our results show that the good properties of OP can be generalized to our multiple study setting.

The plan of the paper is as follows. Section 2 describes the data. Section 3 introduces the Bayesian Multi-study factor analysis (BMSFA) framework, describes our prior, our extension of OP, and our procedure for choosing the number of shared and study-specific factors. Section 4 presents extensive simulation studies, providing evidence on the performance of BMSFA and comparing it with standard methods. We also investigate determining the truncation level for latent factors. Section 5 applies BMSFA to the breast cancer data described in Section 2. Section 5 contains a discussion.

Trial Version Wondershare PDFelement

2 The Breast Cancer Data sets

Breast cancer is both a clinically diverse and a genetically heterogeneous disease (Perou et al., 2000, Planey and Gevaert, 2016). The complex nature of breast cancer has been clarified by classifying breast cancer into subtypes using gene expression measurements from tumor samples. Reliably identifying these subtypes has the potential of driving personalized patient treatment regimens (Masuda et al., 2013) and risk prediction models (Parker et al., 2009). Several groups (Sørlie et al., 2001, Sotiriou et al., 2003, Hu et al., 2006, Planey and Gevaert, 2016) have focused on finding replicable gene expression patterns across different studies, to better classify breast carcinomas into distinct subtypes.

A very valuable statistical approach is unsupervised clustering using different microarrays that query the same set of genes (Perou et al., 2000, Sørlie et al., 2001, 2003, Castro et al., 2016). A challenge is to characterize the extent to which variation in gene expression, and the resulting subtypes, are stable across different studies (Hayes et al., 2006). When different microarray studies are considered together, one is likely to encounter significant and unknown sources of study-to-study heterogeneity (Simon et al., 2009, Bernau et al., 2014). These sources include differences in design, hidden biases, technologies used for measurements, batch effects, and also variation in the populations studied —for example, differences in treatment or disease stage and severity. Quantifying these heterogeneities and dissecting their impact on the replicability of patterns is essential.

A typical bioinformatics analysis pipeline would attempt to remove variation attributable to experimental artifacts before further analysis. If information on batches of other relevant experimental factors is available, their effects can be addressed (Draghici et al., 2007). For example, Sørlie et al. (2001) use the SAM (significance analysis of microarrays) algorithm to detect genes not influenced by batch effect, and then use this set of genes to perform unsupervised cluster analysis. In general, it is challenging to fully remove artifactual effects,

Trial Version Wondershare PDFelement

Study	Adjuvant Therapy	N	N: ER+	3Q survival	Reference
CAL	Chemo, hormonal	118	75	42	Chin et al. (2006)
MAINZ	none	200	162	120	Schmidt et al. (2008)
MSK	combination	99	57	76	Minn et al. (2005)
EXPO	hormonal	517	325	126	Symmans et al. (2010)
TRANSBIG	none	198	134	143	Desmedt et al. (2007)
UNT	none	133	86	151	Sotiriou et al. (2006)
VDX	none	344	209	44	Minn et al. (2007)

Table 1: *The seven data sets considered in the illustration and their characteristics. N is the total number of samples; N: ER+ is the number of Estrogen Receptor positive patients. 3Q survival is the third quartile of the survival function for all patients in the study.*

particularly if they are related to unobserved confounders rather than known batches or factors (Draghici et al., 2007).

The joint analysis of multiple studies offers the opportunity to understand replicable variation across different studies. The overarching goal of this work is to improve the identification of a stable and replicable signal by simultaneously modeling both the components of variation shared across studies, and those that are study-specific. The latter could include artifacts and batch effects that were not addressed by the study specific preprocessing, as well as biological signal that may hard to replicate or genuinely unique to a study. An example of the latter would be the gene expression signature resulting from the administration of a treatment that is used in one study only.

In our case study, we consider a systematic collection of publicly available breast cancer microarray studies compiled by Haibe-Kains et al. (2012). Table 1 provides an overview of the studies, the corresponding references, sample size, Estrogen Receptor (ER) status prevalence, and survival time. Additional details about these studies, their preprocessing, curation, criteria for inclusion, and public availability are described in Haibe-Kains et al.

Trial Version Wondershare PDFelement

(2012). Four of these studies only include patients who did not receive hormone therapy or chemotherapy. Within the Affymetrix technology, genes can be represented by multiple probe-sets. Our analysis considers, for each gene, only the probe-set with maximum mean (Miller et al., 2011). As in Bernau et al. (2014), we only consider we only consider genes measured in all the seven studies and focus on the 50% of genes with higher variance.

3 A Bayesian Framework for multi-study analysis

This section provides details of our model, in four parts:

i) Definition of the multi-study factor model sampling distribution;

ii) Choice of the multiplicative gamma prior (Bhattacharya and Dunson, 2011), with shrinkage priors for the loading matrices to incorporate sparsity. Posterior sampling is carried out by Gibbs sampling, without any constraints on the model parameters;

iii) Choice of truncation level for the latent factor dimensions, determined by a suitable singular value decomposition;

iv) Recovery of the loading matrices, performed by the OP approach.

3.1 Model definition

We consider S studies, each with the same P genomic variables. Study s, $s = 1, \ldots, S$, has n_s subjects and P-dimensional data vector $\mathbf{x}_{is}$, $i = 1, \ldots, n_s$, centered at its sample mean. Our sampling distribution follows the multi-study factor model (De Vito et al., 2016). The variables in study s are decomposed into K factors shared among all studies, and J_s further factors specific to study s, as follows:

$$\mathbf{x}_{is} = \mathbf{\Phi}\mathbf{f}_{is} + \mathbf{\Lambda}_s\mathbf{l}_{is} + \mathbf{e}_{is}\,. \tag{1}$$

Trial Version Wondershare PDFelement

Here $\mathbf{f}_{is} \sim N_k(\mathbf{0}, \mathbf{I}_k)$ are the *shared* latent factors, $\mathbf{\Phi}$ is their $P \times K$ loading matrix; $\mathbf{l}_{is} \sim N_{j_s}(\mathbf{0}, \mathbf{I}_{j_s})$ are the *study-specific* latent factors and $\mathbf{\Lambda}_s$, $s = 1, \ldots, S$ are the corresponding $P \times J_s$ loading matrices; lastly, $\mathbf{e}_{is}$ is the $p \times 1$ Gaussian error vector with covariance $\mathbf{\Psi}_s = \text{diag}(\psi^2_{s_1}, \ldots, \psi^2_{s_p})$. The resulting marginal distribution of $\mathbf{x}_{is}$ is a multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_s = \mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Lambda}_s\mathbf{\Lambda}_s^\top + \mathbf{\Psi}_s$. The covariance matrix of study s can be rewritten as

$$\mathbf{\Sigma}_s = \mathbf{\Sigma}_\Phi + \mathbf{\Sigma}_{\Lambda_s} + \mathbf{\Psi}_s, \tag{2}$$

where $\mathbf{\Sigma}_\Phi = \mathbf{\Phi}\mathbf{\Phi}^\top$ is the covariance of the shared factors, and $\mathbf{\Sigma}_{\Lambda_s} = \mathbf{\Lambda}_s\mathbf{\Lambda}_s^\top$ is the covariance of the study-specific factors. A straightforward implication of (2) is that $\mathbf{\Sigma}_\Phi$ and $\mathbf{\Sigma}_{\Lambda_s}$ describe the variability of the P variables in study s that can be interpreted as shared across studies and specific to study s, respectively.

The decomposition of $\mathbf{\Sigma}_s$ is not unique, as there are infinite possibilities to represent it because $\mathbf{\Phi}^* = \mathbf{\Phi}\mathbf{Q}$ and $\mathbf{\Lambda}_s^* = \mathbf{\Lambda}_s\mathbf{Q}_s$ both satisfy (2) for any two orthogonal matrices $\mathbf{Q}$ and $\mathbf{Q}_s$. MSFA identifies the parameters by imposing constraints on the two factor loadings matrices, such as the lower triangular constraint used in Factor Analysis (FA) (Geweke and Zhou, 1996, Lopes and West, 2004) . This constraint generates an order dependence among the variables. Thus, as noted by Carvalho et al. (2008), the choice of the first $K + J_S$ variables becomes an important modeling choice.

Several approaches focus on the estimation of covariance matrix (Bhattacharya and Dunson, 2011) or precision matrix (Gao et al., 2014a, Zhao et al., 2014). These methods do not require identifiability of the loading matrix. Our approach is also based on this concept: we focus on the estimation of the common variation $\mathbf{\Sigma}_\Phi$ shared among the studies and the variation specific to each study $\mathbf{\Sigma}_{\Lambda_s}$. The two matrices $\mathbf{\Sigma}_\Phi$ and $\mathbf{\Sigma}_{\Lambda_s}$ are only assumed to be positive semidefinite normal matrices, i.e. symmetric matrices with a subset of positive non-null eigenvalues.

Trial Version Wondershare PDFelement

3.2 The multiplicative gamma shrinkage prior

We adapt a shrinkage prior from Bhattacharya and Dunson (2011) for both the common and study-specific factor loadings. The shrinkage priors favor sparsity by removing some entries of the loading matrix. When an element is close to zero, the variable corresponding to the row does not contribute to the common or study-specific latent factor corresponding to the column. In the genomic context, this sparsity models the biological reality that only a subset of the genes represented in a cell's transcriptome is participating in a specific biological function (Tegner et al., 2003). Another important property of the Bhattacharya and Dunson (2011) prior is that the shrinkage towards zero increasing with the column index of the loading matrix.

Our extension of the multiplicative gamma shrinkage prior to the multiple study setting is as follows. The prior for the elements of the shared factor loading matrix $\mathbf{\Phi}$ is

$$\phi_{pk} \mid \omega_{pk}, \tau_k \sim N(0, \omega_{pk}^{-1}\tau_k^{-1}), \quad p = 1, \ldots, P, \; k = 1, \ldots, \infty,$$

$$\omega_{pk} \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \qquad \tau_k = \prod_{l=1}^{k} \delta_l \qquad \delta_1 \sim \Gamma(a_1, 1) \qquad \delta_l \sim \Gamma(a_2, 1), \quad l \geq 2$$

where δ_l $(l = 1, 2, \ldots)$ are independent, τ_k is the global shrinkage parameter for the k-th column and ω_{pk} is the local shrinkage for the element p in column k. We then replicate this scheme to specify the prior for the elements of the study-specific factor loading matrix $\mathbf{\Lambda}_s$:

$$\lambda_{pj_s} \mid \omega^s_{pj_s}, \tau^s_{j_s} \sim N(0, \omega^{s^{-1}}_{pj_s}\tau^{s^{-1}}_{j_s}), \quad p = 1, \ldots, P, \; j_s = 1, \ldots, \infty, \quad \text{and } s = 1, \ldots, S,$$

$$\omega_{pj_s} \sim \Gamma\left(\frac{\nu^s}{2}, \frac{\nu^s}{2}\right) \qquad \tau^s_{j_s} = \prod_{l=1}^{j_s} \delta^s_l \qquad \delta^s_1 \sim \Gamma(a^s_1, 1) \qquad \delta^s_l \sim \Gamma(a^s_2, 1), \quad l \geq 2$$

where $\delta^s_l (l = 1, 2, \ldots)$ are independent, $\tau^s_{j_s}$ is the global shrinkage parameter for the j_s column and $\omega^s_{pj_s}$ is the local shrinkage for the element p in column j_s.

Trial Version Wondershare PDFelement

For each of the error variances ψ_{ps}, $p = 1, \ldots, P$ we assume an inverse gamma prior $\psi_{ps}^{-1} \sim \Gamma(a_\psi, b_\psi)$. This choice, made also by Bhattacharya and Dunson (2011), is common in standard FA (Lopes and West, 2004, Gao et al., 2013, Ročková and George, 2016). Sampling from the posterior distribution of the model parameters is carried out by Gibbs sampling. Details are in Supplementary Materials.

3.3 Choosing the number of latent factors

In practical applications, the number of important latent factors is likely to be small compared to the number of variables P. As suggested by Bhattacharya and Dunson (2011), the effective number of factors would be small when data are sparse. Our approach circumvents the need for pre-specifying the latent dimension since the shrinkage prior gives positive mass to an infinite number of them. However, we need a proper computational strategy for choosing accurate truncation levels K and J_s, $s = 1, \ldots, S$. Ideally, we would like to retain the relevant factors discarding the redundant ones.

An analogous task for FA is addressed in Bhattacharya and Dunson (2011) who truncate the number of factors to a finite value, usually far smaller than the number of variables P. This truncation level is chosen by checking the columns of the estimated loading matrix, to assess which ones are formed entirely by elements of negligible size. The fact that the shrinkage implied by the prior increases in later columns greatly simplifies this task, compared to what required by alternative shrinkage priors such as the spike and slab (Carvalho et al., 2008). We use the same idea, though computational details differ.

Our practical method to assess the numbers of shared factors K and study-specific factors J_s is based on singular value decomposition (SVD) and proceeds as follows. Starting from a considerable number of shared and study-specific factors, we seek $K \ll P$ and $J_s \ll P$. In the MSFA model, this implies that the two matrices $\boldsymbol{\Sigma}_\Phi$ and $\boldsymbol{\Sigma}_{\Lambda_s}$ are singular,

Trial Version Wondershare PDFelement

with ranks K and J_s, respectively. Since these matrices are symmetric, they have K and J_s non-null eigenvalues. Based on this, we compute the eigenvalues $\nu_1, \ldots, \nu_P$ of $\widehat{\boldsymbol{\Sigma}}_\Phi$, with $\nu_p \geq 0$, $p = 1, \ldots, P$, ordered in decreasing size. We then choose K as the number of eigenvalues larger than a pre-specified positive threshold, to achieve $\mathbf{U}\,\mathbf{N}_K\mathbf{U}^\top \doteq \widehat{\boldsymbol{\Sigma}}_\Phi$, where $\mathbf{N}_K = \text{diag}(\nu_1, \ldots, \nu_K)$, and the columns of $\mathbf{U}$, of size $P \times K$, are given by K (normalized) eigenvectors of $\widehat{\boldsymbol{\Sigma}}_\Phi$. We proceed in the same way for J_s, $s = 1, \ldots, S$.

3.4 Recovering loading matrices

The method of §3.1-3.3 provides a practical route to the estimation of $\boldsymbol{\Sigma}_\Phi$ and $\boldsymbol{\Sigma}_{\Lambda_s}$, but in many applications recovery of the loading matrices is also useful. Recently Aßmann et al. (2016) solved the identification issue in the context of FA by first generating an MCMC sample without any constraints, and then filtering out the possible effect of orthogonal rotations. They solve an Orthogonal Procrustes (OP) problem (Gower and Dijksterhuis, 2004) by building a sequence of orthogonal matrices defined from the MCMC output.

Here we extend this procedure to BMSFA. When the model parameters are not constrained, the Gibbs sampler is said to be orthogonally mixed (Aßmann et al., 2016), as each chain may produce different orthogonal transformations (represented by the matrices $\mathbf{Q}$ and $\mathbf{Q}_s$) for the factor loadings $\boldsymbol{\Phi}^*$ and $\boldsymbol{\Lambda}_s^*$. Starting from a sequence of R draws from the posterior distribution of $\boldsymbol{\Phi}(\boldsymbol{\Phi}^1, \ldots, \boldsymbol{\Phi}^R)$, the OP algorithm circumvents this problem by estimation the loading matrices via the following constrained optimization:

$$\left\{ \left\{ \tilde{\mathbf{Q}} \right\}_{r=1}^{R}, \tilde{\boldsymbol{\Phi}}^* \right\} = \underset{\mathbf{Q}^{(r)}, \boldsymbol{\Phi}^*}{\text{argmin}} \sum_{r=1}^{R} L_Q \left(\boldsymbol{\Phi}^*, \boldsymbol{\Phi}^{(r)}\mathbf{Q}^{(r)} \right) \quad \text{s.t.} \quad \mathbf{Q}^{(r)}\mathbf{Q}^{(r)\top} = \mathbf{I}_K,\ r = 1, \ldots, R \quad (3)$$

where L_Q is the loss function

$$L_Q \left(\boldsymbol{\Phi}^*, \boldsymbol{\Phi}^{(r)}\mathbf{Q}^{(r)} \right) = \text{tr} \left\{ \left(\boldsymbol{\Phi}^{(r)}\mathbf{Q}^{(r)} - \boldsymbol{\Phi}^* \right)^\top \left(\boldsymbol{\Phi}^{(r)}\mathbf{Q}^{(r)} - \boldsymbol{\Phi}^* \right) \right\}.$$

The optimization is carried out by iterating two steps:

Trial Version Wondershare PDFelement

1. Minimize equation (3), for a given $\mathbf{\Phi}^*$ by computing the SVD of $\mathbf{\Sigma}_{\Phi^*} = \mathbf{\Phi}^{(r)}\mathbf{\Phi}^{*\top}$ and setting $\tilde{\mathbf{Q}}^{(r)} = \mathbf{U}_r\mathbf{V}_r$, where $\mathbf{U}_r$ and $\mathbf{V}_r$ are the two orthogonal matrices obtained by the SVD at MCMC iteration r,

2. Compute $\tilde{\mathbf{\Phi}}^{*(r)} = \frac{1}{R}\sum_{r=1}^{R}\mathbf{\Phi}^{(r)}\tilde{\mathbf{Q}}^{(r)}$.

The algorithm is then iterated using the updated value of $\tilde{\mathbf{\Phi}}^*$ in place of $\mathbf{\Phi}^*$. The search stops when subsequent estimates of $\mathbf{\Phi}$ are close enough.

This algorithm requires a starting value for $\mathbf{\Phi}^*$. Aßmann et al. (2016) suggests the last iteration of the Gibbs sampler as initial value for $\mathbf{\Phi}^*$. The same procedure can be applied to each of the study-specific loading matrices. This algorithm provides an approximate solution to identifiability, since the posterior distribution of the loading matrices is only known in approximate form. Yet, Aßmann et al. (2016) show that it can be quite effective.

The OP procedure is iterative in nature. However, we verified that typically the first iteration is sufficient to get close to the final estimate. Since the OP algorithm is computationally demanding, the one-step version is recommendable. All the results of this paper have been obtained with a single iteration of the OP algorithm.

This point will be further examined for our setting in the following section.

4 Simulation Results

In this section we use simulation experiments to assess BMSFA's ability to recover common and study-specific latent dimensions, by itself and in comparison to standard FA applied to the merged datasets. We generate 50 datasets from the distributions specified in Table 2. We fixed $\mathbf{\Phi}$, $\mathbf{\Lambda}_s$ and $\mathbf{\Psi}_s$ and thus $\mathbf{\Sigma}_s$. We consider four scenarios differing in the number of studies, study sample sizes, and covariance structure (see Figure 1). Scenarios 1 and 2 are similar to Zhao et al. (2014): n_s is chosen to be smaller than P to mimic large P and

Trial Version Wondershare PDFelement

$\mathbf{X}_s \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma}_s)$
$\mathbf{\Sigma}_s = \mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Lambda}_s\mathbf{\Lambda}_s^\top + \mathbf{\Psi}_s$
fixed $\mathbf{\Phi}$ and $\mathbf{\Lambda}_s$: sparse matrices with ≈ 80 % of zeros
fixed $\mathbf{\Phi}$ and $\mathbf{\Lambda}_s$: non zero elements drawn once from $\text{U}(-1, 1)$
fixed $\mathbf{\Psi}_s$: diagonal elements drawn once from $\text{U}(0, 1)$

Table 2: *Distributions used to generate observations in study s, for simulation experiments.*

Common factor loadings : $\omega_{pk} \sim \Gamma\left(\frac{\nu=3}{2}, \frac{\nu=3}{2}\right)$
Study-Specific factor loadings : $\omega_{pj_s} \sim \Gamma\left(\frac{\nu^s=3}{2}, \frac{\nu^s=3}{2}\right)$
$\delta_1 \sim \Gamma(a_1 = 2.1, 1)$ and $\delta_l \sim \Gamma(a_2 = 3.1, 1)$ with $l \geq 2$
$\delta_1^s \sim \Gamma(a_1^s = 2.1, 1)$ and $\delta_l^s \sim \Gamma(a_2^s = 3.1, 1)$ with $l \geq 2$
$\mathbf{\Psi}_s^{-1} \sim \Gamma(a_\psi = 1, b_\psi = 0.3)$

Table 3: *Prior distributions used in the simulation experiments and real data analysis.*

small n_s conditions while operating with a manageable set of variables for visualization and summarization. In the Scenario 3 we wish to model a situation where not all the studies have $P \gg n$. Moreover, in this scenario, study-specific factor loadings are large. The motivation behind this scenario is to investigate if our method recovers the shared biological signal in the presence of large study-specific or batch effects, and if it can isolate these sources. In Scenario 4 we closely mimic the data in Table 1, choosing $S = 7$ and matching the sample sizes to those of Table 1. Moreover, in Scenarios 1, 2 and 4 we randomly allocate the zeros in each column of $\mathbf{\Phi}$ and $\mathbf{\Lambda}_S$ (Table 2), while in Scenario 3, we allocate zeros matching the central panel in the third row of Figure 1. We run the Gibbs sampler for 15000 iterations with a burn-in of 5000 iterations. We set priors as in Table 3.

We first evaluate, for fixed latent dimension K and J_s, BMSFA's ability to recover the covariance component $\mathbf{\Sigma}_\Phi$ determined by the shared factors, as well as the shared factors' loadings $\mathbf{\Phi}$. For one randomly selected simulation dataset, Figure 1 compares the true and estimated elements of $\mathbf{\Sigma}_\Phi$. We also present a summary of the analyses of 50 datasets. To quantify the similarity between $\mathbf{\Sigma}_\Phi^{true}$ and $\widehat{\mathbf{\Sigma}}_\Phi$ we use the RV coefficient (Robert and Escouffer, 1976) of similarity of two $P \times P$ matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$:

$$RV(\mathbf{S}_1, \mathbf{S}_2) = \frac{tr((\mathbf{\Sigma}_1\mathbf{\Sigma}_2^\top)(\mathbf{\Sigma}_1\mathbf{\Sigma}_2^\top)}{tr(\mathbf{\Sigma}_1\mathbf{\Sigma}_1^\top)^2 tr(\mathbf{\Sigma}_2\mathbf{\Sigma}_2^\top)^2}.$$

RV varies in $[0, 1]$. The closer RV is to 1 the more similar the two matrices are. Smilde

Trial Version Wondershare PDFelement

et al. (2008) argue that the RV coefficient can overestimate similarity between data sets in high-dimensions, and propose a modified version that addresses this problem. We use it in Scenario 4, though differences will not be pronounced. The red boxplots in the right column of Figure 1 show the RV distributions across 50 simulations in our four scenarios.

Figure 2 presents a similar analysis comparing the true factor loadings to their estimates obtained through posterior sampling and the OP procedure. The correlations between true

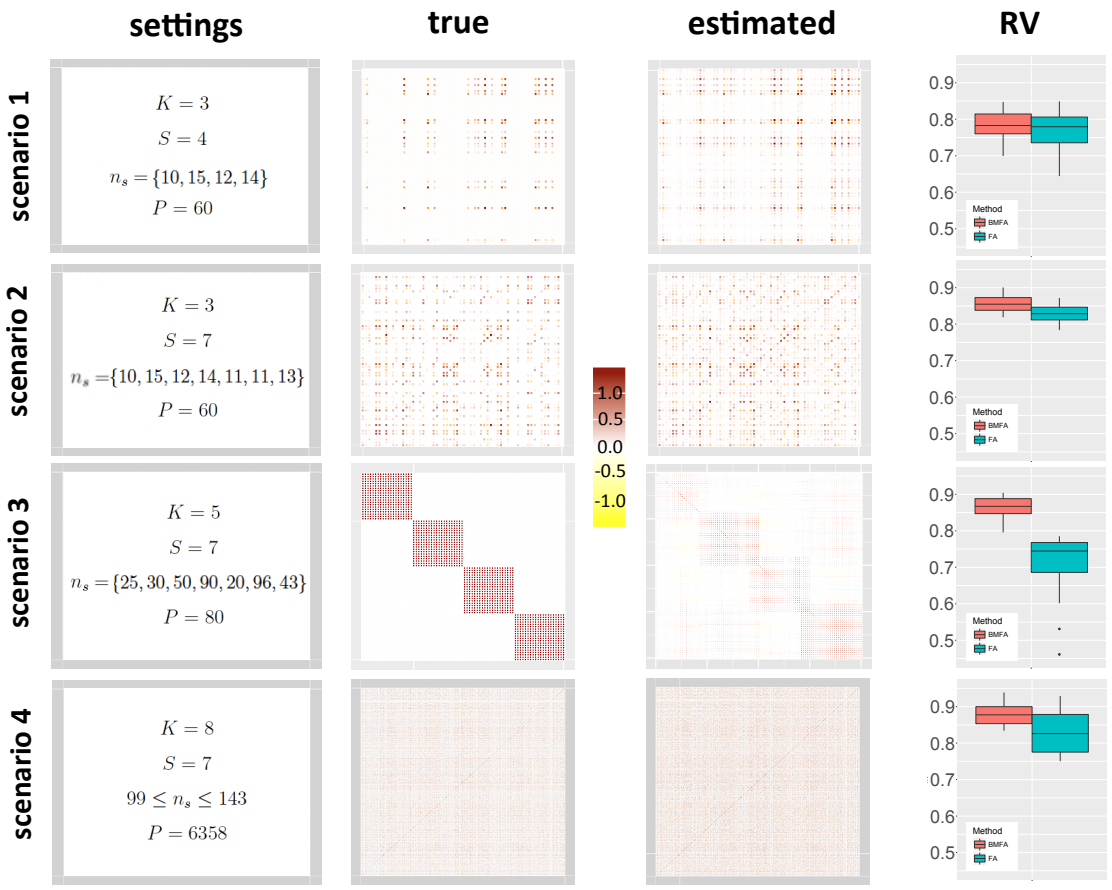

Figure 1: *Covariance matrices* $\mathbf{\Sigma}_\Phi$ *and their Bayesian estimates in four simulation scenarios. The right column shows the boxplots of RV coefficient between the true and the estimated* $\mathbf{\Sigma}_\Phi$.

Trial Version Wondershare PDFelement

and estimated values in both Figures 1 and 2) are very high, suggesting that our estimands are well identified and our sampling approaches are appropriate.

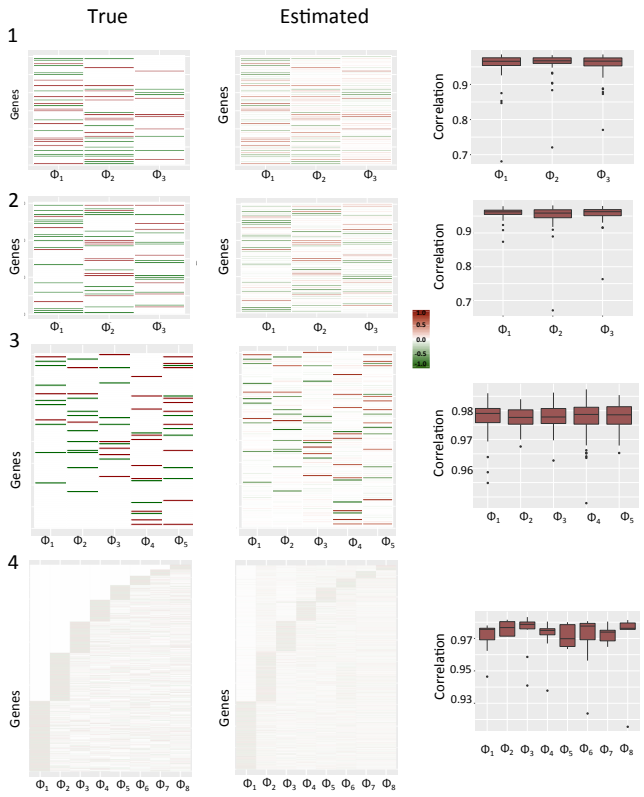

Figure 2: *Heatmap of the true (left) and estimated (center) shared factor loadings* $\mathbf{\Phi}$ *in the four scenarios of Figure 1. In Scenario 4 we only show common factor loadings* ≥ 0.5*. The right column displays boxplots of correlations between the true and estimated common factor loadings over 50 datasets for each scenario.*

Next we compare BMSFA to a Bayesian FA, using the same prior distribution. For Bayesian FA, we stacked all studies into a single dataset, ignoring that samples originate from distinct studies. The RV coefficients for BMSFA are systematically greater than FA's (Figure 1, right column), demonstrating that BMSFA recovers shared factors better than a

Trial Version Wondershare PDFelement

merged analysis. In Scenario 3 the gap is more pronounced, as study-specific factor loadings are large. In most simulations, FA captures study-specific effects that are not actually shared. BMSFA recovers the shared signal better. Also, the distribution of BMSFA's RV coefficient is narrower than FA's. This comparison illustrates that BMSFA identifies the shared signal across the studies and improves its estimation compared to standard Bayesian FA. Moreover, the BMSFA estimations are more efficient compared to the FA estimation, due to the beneficial effects of removing the study-specific components that lack cross-study reproducibility.

So far we took K, the number of shared factors, and J_s's, the numbers of study-specific factors, to be known. We next focus on the latent dimensions calculated via SVD of matrices $\mathbf{\Sigma}_\Phi$ and $\mathbf{\Sigma}_{\Lambda_s}$, as described earlier, and using an eigenvalue threshold of 0.05. The simple adaptive method described in §3.3 for latent factor selection, common K and specific J_s, proved to be extremely robust respect to the choice of this threshold. Conclusion with a threshold of 0.1 was the same. We choose a lower value as are more concerned to lose important shared biological factors than to include additional shared factors. Figure 3 shows the results obtained by fitting the model for 50 different data sets generated from the BMSFA with $K = 3$ in the four different scenarios. The vertical lines show the 50 estimated latent dimensions in each data set. Our method consistently selects the right dimensions for both the shared and the study-specific factors.

The simulation analysis highlights the merit of our method in a variety of scenarios, with improved performance over FA in terms of covariance matrices estimation in multi-study settings, estimation of the reproducible signal across studies, and identification issue for the factor loading matrix.

Trial Version Wondershare PDFelement

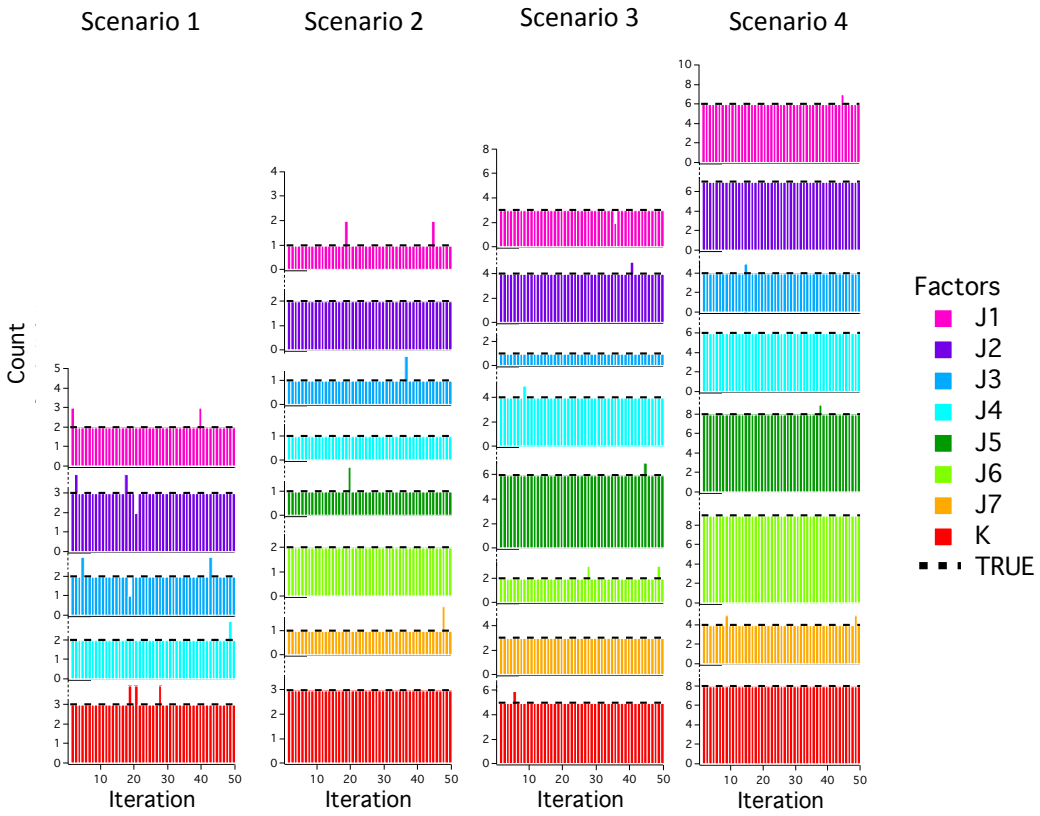

Figure 3: *Dimensions of shared and study specific factors in the four scenarios. Model selection procedure for the shared K and the study-specific J_s latent dimension via SVD of $\mathbf{\Sigma}_\Phi$ and $\mathbf{\Sigma}_{\Lambda_s}$. The true dimensions are visualized by the dashed lines.*

5 Breast Cancer Case Study

The aim of this analysis is to identify shared common factors describing the common correlation structure across the 7 breast cancer microarray studies listed in Table 1. Recovering shared gene co-expression patterns from different high-throughput studies is important to identify replicable genetic regulation. This case study considers a relatively well understood area of cancer biology and provides a realistic positive control for the BMSFA methodology.

Trial Version Wondershare PDFelement

We consider genes measured in all studies and remove the 50% of genes with the lowest variance. We use the prior of Table 3. Our method chooses a shared latent dimension of $K = 8$, through the SVD of $\mathbf{\Sigma}_\Phi$. We first summarize and visualize the shared co-expression patterns via a co-expression network (Figure 4) built on $\mathbf{\Sigma}_\Phi$, and thus representing all studies. A gene co-expression network is an undirected graph. Each node corresponds to a gene and each edge represents a high co-expression between genes. The importance of genes in a cluster is represented by the node size.

Our analysis identifies five larger clusters. Co-expressed genes tend to be members of the same, highly plausible, biological pathways. All clusters are associated with biological processes known for explaining heterogeneity of expression across breast cancers, lending credibility to BMSFA. The first cluster is driven by expression of the estrogen receptor (ESR1), which historically is one of the earliest cancer biomarkers to have been discovered, and plays a crucial role in the biology and treatment of breast cancer (Jordan, 2007, Robinson et al., 2013). High dimensional expression pattern are found in Sørlie et al. (2001). Many studies have shown the relation of ESR1 with growth of cancer (Osborne et al., 2001, Iorio et al., 2005, Toy et al., 2013). Levels of ESR1 expression are associated with different outcomes (Ross-Innes et al., 2012, Theodorou et al., 2013). Three other genes stand out: GATA3, XBP1 and FOXA1. These are ESR1-cooperating transcription factors altered in breast tumors (Lacroix and Leclercq, 2004, Theodorou et al., 2013). In breast cancer cell, many studies revealed strong and positive association of GATA3, XBP1 and FOXA1 with ESR1 (Hoch et al., 1999, Sotiriou et al., 2003, Sørlie et al., 2003, Lacroix and Leclercq, 2004, Lai et al., 2013, Theodorou et al., 2013). The second cluster is related to the cell cycle. One of the most important genes in this cluster is CCNB1, which encodes cyclin B. Cyclins are prime cell cycle regulators. Many analyses found a common pattern of overexpression of the mitotic cyclins A and B and their dependent kinase in the tumor cell of breast cancer (Keyomarsi and Pardee, 1993, Lin et al., 2000, Basso et al., 2002).

Trial Version Wondershare PDFelement

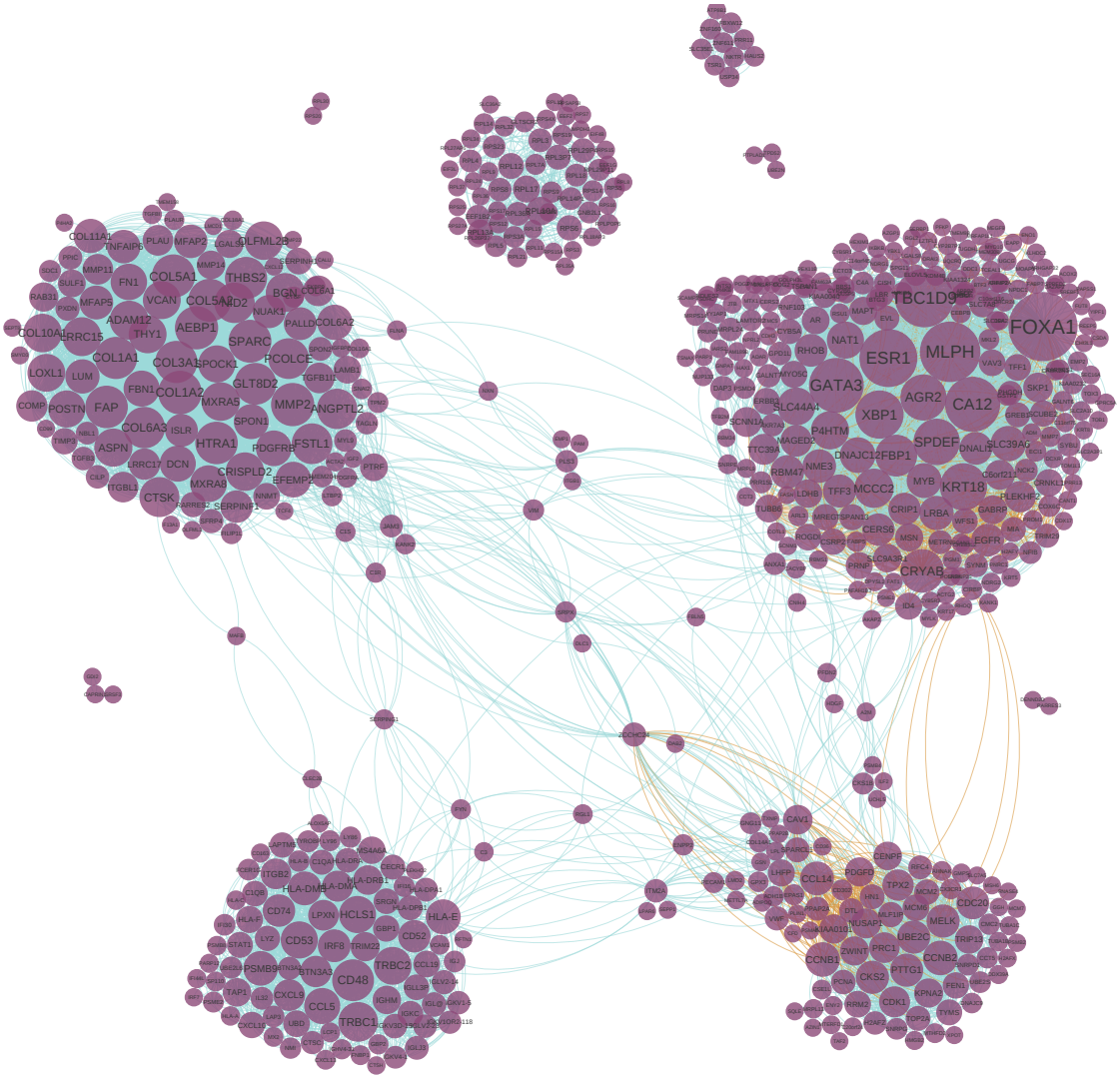

Figure 4: *Shared gene co-expression network across the 7 studies of Table 1. We include edges between two genes if the corresponding element in the shared part of the covariance matrix is greater than 0.5. Edges in blue (orange) represent positive (negative) associations.*

Two other important genes in this cluster are CDK1, a kinase dependent on cyclins, and CDC20, a gene related to the metaphase and anaphase of cell cycle. All genes in the third

Trial Version Wondershare PDFelement

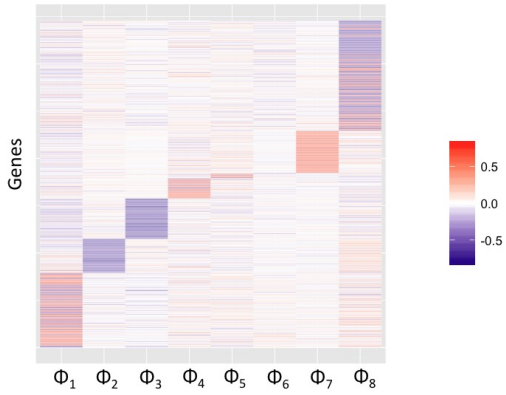

Figure 5: *Heatmap of the estimated shared factor loadings obtained with BMSFA across the 7 studies in Table 1. We only show common factor loadings* ≥ 0.5.

cluster are related to regulation of the immune response. The CD genes are important for the immune system pathway and the HLA genes are a crucial element for immune function. The fourth cluster includes several genes expressed by the connective tissue, including collagen genes (COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, COL10A1, COL11A1), previously associated with stromal cells (Ross et al., 2000, Ioachim et al., 2002). Of note are also ADAM, a protease related to the degradation of the connective tissue, and smooth muscle cell marker TAGLN, also previously found to play a role in breast cancer. Finally, all the RP genes in the fifth cluster codify the ribosome, which synthesizes proteins. Dysregulation of Ribosome function is related to tumor progression in breast cancer (Belin et al., 2009).

To further explore the patterns found in the shared gene co-expression network, we estimate the shared factor loadings after the post-process procrustes algorithm. The heatmap in Figure 5 depicts the estimates of the shared factor loadings that can be identified reproducibly across the studies. To extract biological insight from the shared factors, we explore whether specific gene sets are enriched among the loadings using Gene Set En-

Trial Version Wondershare PDFelement

richment Analysis (GSEA) (Mootha et al., 2003). We used the package `RTopper` in `R` in `Bioconductor`, following the method of Tyekucheva et al. (2011) and considering all the gene sets representing pathways from `reactome.org`. The resulting analysis shows concordant results with the pathways obtained with the shared gene co-expression network, further suggesting that we identify genuine biological signal. The first shared factor is significantly enriched with the "Cell communication" and "Cell cycle" pathways. The second factor is associated with the Immune system pathway and all the sub-pathway included in it, namely the "Adaptive Immune System", "Innate Immune System" and "Cytokine Signaling in Immune System". Factor 3 shows a significant association with cell cycle, namely with the pathway "Cell cycle", "Cell cycle mitotic", "Cell cycle checkpoints", "Regulation of mitotic cell cycle". The shared 5, 6 and 7 factors have protein production "Transport of ribonucleoproteins into the host nucleus", "Protein folding", "Mitochondrial protein import", "Metabolism of proteins", "NRIF signals cell death from the nucleus". Finally, factor 8 is related to the ER pathway, "ER phagosome pathway", "Interferon signaling", and "Interferon alpha beta signaling".

An important feature of BMSFA in this case study is regularization of the common factor loadings. To illustrate this in more detail, we conclude this section comparing BMSFA to the MSFA which uses MLE for parameter estimation. The data consists of 63 genes in the Immune System Pathway. Their loadings are compared in Figure 6. BMSFA regularizes common factor loadings by shrinking small and moderate MLE loadings to zero while systematically amplifying larger MLE loadings (Figure 6, left panel). This regularization behavior is somewhat unique to this setting, as it is far more common for regularization to only result in shrinkage. Here, the prior helps the posterior perform a factor rotation method which results in more sparse factors. To further illustrate we rotate the loadings obtained with the MLE using the varimax rotation (Kaiser, 1958) and we compare it with the BMSFA (Figure 6 right panel). The BMSFA loadings are far more similar to the

Trial Version Wondershare PDFelement

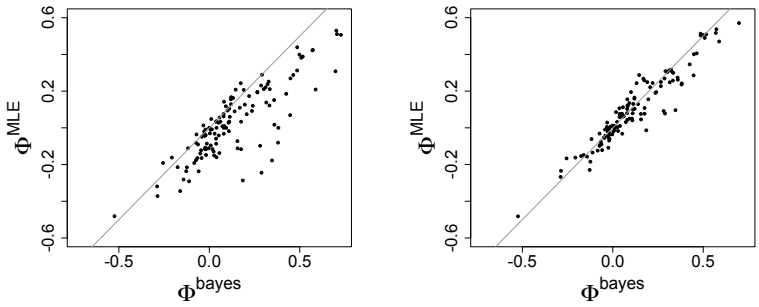

Figure 6: *Left: Comparison between the first two estimated common factor loadings with MLE and BMSFA. Right: Comparison between the first two estimated factor loadings with MLE, followed by varimax rotation, and BMSFA.*

varimax rotated MLE (correlation 0.97) than the original MLE (correlation 0.7).

6 Discussion

In this paper we propose a general Bayesian framework for the unsupervised analysis of high-dimensional biological data across multiple studies, building upon De Vito et al. (2016). We address the unmet need to rigorously model replicable signal across studies, while at the same time capturing study-specific variation. Our approach is not limited by $P \ll n$ and, in addition to replicability, shows considerable promise in modeling sparsity and enhancing interpretability via rotation-like shrinkage. Building on Bhattacharya and Dunson (2011) we propose a computationally efficient MCMC algorithm.

The work in this paper is motivated by identifying replicable signal in unsupervised genomic applications. The results of our analysis of breast cancer gene expression across seven studies identified shared gene patterns, which we also represented via clusters in a co-expression network. Both factors and clusters are clearly related to well-known breast

Trial Version Wondershare PDFelement

cancer pathways. Our analytic tools allows investigators to focus on the replicable shared signal, after properly accounting for, and separating, the influence of study-specific variation. While we focused on the shared signal, study-specific loadings can also be examined directly.

BMSFA may have broad applicability in a wide variety of genomic platforms, including microarrays, RNA-seq, SNP-based genome-wide association studies, proteomics, metabolomics, and epigenomics. Relevance is also immediate in other fields of biomedical research, such as those generated by exposome studies, Electronic Medical Record (EMR), or high dimensional epidemiological data. In dietary pattern analysis, it is important to find replicable dietary patterns in different populations (Castelló et al., 2016). Our analysis could be applied to check if there are shared dietary patterns across different populations and to detect the study-specific dietary patterns of a particular population. In this field generally, it is common to apply a varimax rotation to factor loading matrix, for a better interpretation. Specifically, the interpretation of a factor relies on loadings. The interpretation of the model is simplified if more of the loadings are shrunk towards zero and the factor is defined by few large loadings. In the frequentist analysis, this is possible by rotation methods, such as varimax. In our representation, the BMSFA embeds this step giving an immediate representation of the two sparse factor loading matrices through the shrinkage prior, as shown in Section 5.

Our Bayesian non-parametric approach offers more flexibility in the choice of the dimensionality of shared latent factors. Moreover, we provide shrinkage of the latent factor loadings, enhancing the role of the variables that are most important in each factor.

To address the choice of model dimension, we developed, building on Bhattacharya and Dunson (2011), a practical procedure based on separate SVD of the shared covariance part and the study-specific covariance parts. The choice of the number of factors remains an important open problem. The most common method for choosing latent dimension fits the

Trial Version Wondershare PDFelement

factor model for different choices of K and compares them using selection criteria such as BIC. This approach presents many problems especially in a $p \gg n$ setting where MLE is not duable. Lopes and West (2004) proposed a reversible jump MCMC to estimate the number of factors in standard FA, but this method is also often computationally intensive. Bhattacharya and Dunson (2011) developed an interesting adaptive scheme that dynamically changes the dimension of the latent factors as the Gibbs sampling progresses. In our approach, we develop a practical approach where we have a balance between retaining important factors and removing the redundant ones.

We also address identification. Identifiability remains a challenge in standard FA. In the Bayesian approach, constraints were proposed to tackle this issue, such as that of a block lower triangular matrix (Lopes and West, 2004, Carvalho et al., 2008). As Carvalho et al. (2008) noticed, in this constraint different ordering of variables could lead to different conclusions. In our work, we adopt a procrustes algorithm and demonstrate through a series of simulation analyses that this method applied to the BMSFA is effective. Ročková and George (2016) solves this problems in a Bayesian context by rotating the factor loadings matrix with the varimax rotation (Kaiser, 1958). We also compared the BMSFA estimates after the procrustes algorithm with the MLE after rotating the common factor loadings. The resulting analysis are close, demonstrating that the prior we adopt works similarly to a rotation.

We hope BMSFA will encourage joint analyses of multiple high-throughput studies in biology, and contribute to alleviating the current challenges in replicability of unsupervised analyses in this fields and across data science.

Trial Version Wondershare PDFelement

References

Aach, J., Rindone, W., and Church, G. M. (2000). Systematic management and analysis of yeast gene expression data. *Genome Research*, 10(4):431–445.

Abdi, H., Williams, L. J., and Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(2):149–179.

Aßmann, C., Boysen-Hogrefe, J., and Pape, M. (2016). Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem. *Journal of Econometrics*, 192(1):190–206.

Basso, A. D., Solit, D. B., Munster, P. N., and Rosen, N. (2002). Ansamycin antibiotics inhibit Akt activation and cyclin D expression in breast cancer cells that overexpress HER2. *Oncogene*, 21(8):1159–1166.

Belin, S. et al. (2009). Dysregulation of ribosome biogenesis and translational capacity is associated with tumor progression of human breast cancer cells. *PloS One*, 4(9):e7147.

Bernau, C. et al. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, 30(12):i105–i112.

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

Blum, Y., Le Mignon, G., Lagarrigue, S., and Causeur, D. (2010). A factor model to analyze heterogeneity in gene expression. *BMC Bioinformatics*, 11(1):368.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008).

Trial Version Wondershare PDFelement

High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456.

Castelló, A. et al. (2016). Reproducibility of data-driven dietary patterns in two groups of adult spanish women from different studies. *British Journal of Nutrition*, 116(4):734–742.

Castro, M. A. et al. (2016). Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nature Genetics*, 48(1):12.

Chin, K. et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541.

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45(10):1127–1133.

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2016). Multi-study factor analysis. *arXiv preprint arXiv:1611.06350*.

Desmedt, C. et al. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical Cancer Research*, 13(11):3207–3214.

Draghici, S. et al. (2007). A systems biology approach for pathway level analysis. *Genome Research*, 17(10):1537–1545.

Dray, S., Chessel, D., and Thioulouse, J. (2003). Co-inertia analysis and the linking of ecological data tables. *Ecology*, 84(11):3078–3089.

Trial Version Wondershare PDFelement

Edefonti, V. et al. (2012). Nutrient-based dietary patterns and the risk of head and neck cancer: a pooled analysis in the international head and neck cancer epidemiology consortium. *Annals of Oncology*, 23(7):1869–1880.

Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104(488):1406–1415.

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 preprint arXiv:1310.4792*.

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 Computational Biology*, 12(7):e1004791.

Gao, C., Zhao, S., McDowell, I. C., Brown, C. D., and Engelhardt, B. E. (2014a). Differential gene co-expression networks via bayesian biclustering models. *arXiv preprint arXiv:1411.1997*.

Gao, J., Ciriello, G., Sander, C., and Schultz, N. (2014b). Collection, integration and analysis of cancer genomic profiles: from data to insight. *Current Opinion in Genetics & Development*, 24:92–98.

Garrett-Mayer, E., Parmigiani, G., Zhong, X., Cope, L., and Gabrielson, E. (2008). Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics*, 9(2):333–354.

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies*, 9(2):557–587.

Trial Version Wondershare PDFelement

Gower, J. C. and Dijksterhuis, G. B. (2004). *Procrustes problems*. Number 30. Oxford University Press on Demand.

Haibe-Kains, B. et al. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*, 104(4):311–325.

Hayes, D. N. et al. (2006). Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *Journal of Clinical Oncology*, 24(31):5079–5090.

Hoch, R. V., Thompson, D. A., Baker, R. J., and Weigel, R. J. (1999). GATA-3 is expressed in association with estrogen receptor in breast cancer. *International Journal of Cancer*, 84(2):122–128.

Hu, Z. et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, 7(1):96.

Huttenhower, C., Hibbs, M., Myers, C., and Troyanskaya, O. G. (2006). A scalable method for integration and functional analysis of multiple microarray datasets. *BMC Bioinformatics*, 22(23):2890–2897.

Ioachim, E. et al. (2002). Immunohistochemical expression of extracellular matrix components tenascin, fibronectin, collagen type IV and laminin in breast cancer: their prognostic value and role in tumour invasion and progression. *European Journal of Cancer*, 38(18):2362–2370.

Iorio, M. V. et al. (2005). MicroRNA gene expression deregulation in human breast cancer. *Cancer Research*, 65(16):7065–7070.

Irizarry, R. A. et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

Trial Version Wondershare PDFelement

Jordan, V. C. (2007). Chemoprevention of breast cancer with selective oestrogen-receptor modulators. *Nature Reviews Cancer*, 7(1):46–53.

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

Kerr, K. F. (2007). Extended analysis of benchmark datasets for agilent two-color microarrays. *BMC Bioinformatics*, 8(1):371.

Keyomarsi, K. and Pardee, A. B. (1993). Redundant cyclin overexpression and gene amplification in breast cancer cells. *Proceedings of the National Academy of Sciences*, 90(3):1112–1116.

Lacroix, M. and Leclercq, G. (2004). About GATA3, HNF3A, and XBP1, three genes co-expressed with the oestrogen receptor-α gene (ESR1) in breast cancer. *Molecular and cellular endocrinology*, 219(1):1–7.

Lai, C.-F. et al. (2013). Co-regulated gene expression by oestrogen receptor α and liver receptor homolog-1 is a feature of the oestrogen response in breast cancer cells. *Nucleic Acids Research*, 41(22):10228–10240.

Lin, S.-Y. et al. (2000). β-catenin, a novel prognostic marker for breast cancer: its roles in cyclin D1 expression and cancer progression. *Proceedings of the National Academy of Sciences*, 97(8):4262–4266.

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

Masuda, H. et al. (2013). Differential response to neoadjuvant chemotherapy among 7 triple-negative breast cancer molecular subtypes. *Clinical Cancer Research*, 19(19):5533–5540.

Trial Version Wondershare PDFelement

Meng, C., Kuster, B., Culhane, A. C., and Gholami, A. M. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, 15(1):1.

Miller, J. A. et al. (2011). Strategies for aggregating gene expression data: the collapseRows R function. *Bioinformatics*, 12(1):322.

Minn, A. J. et al. (2005). Genes that mediate breast cancer metastasis to lung. *Nature*, 436(7050):518–524.

Minn, A. J. et al. (2007). Lung metastasis genes couple breast tumor size and metastatic spread. *Proceedings of the National Academy of Sciences*, 104(16):6740–6745.

Mootha, V. K. et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273.

Osborne, C. K., Schiff, R., Fuqua, S. A., and Shou, J. (2001). Estrogen receptor: current understanding of its activation and modulation. *Clinical Cancer Research*, 7(12):4338s–4342s.

Parker, J. S. et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167.

Perou, C. M. et al. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747.

Pharoah, P. D. P. et al. (2013). GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nature Genetics*, 45(4):362–370.

Planey, C. R. and Gevaert, O. (2016). Coincide: A framework for discovery of patient subtypes across multiple datasets. *Genome medicine*, 8(1):27.

Trial Version Wondershare PDFelement

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D., and Chinnaiyan, A. M. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research*, 62(15):4427–4433.

Riester, M. et al. (2014). Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *Journal of the National Cancer Institute*, 106(5).

Robinson, D. R. et al. (2013). Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nature Genetics*, 45(12):1446–1451.

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

Ross, D. T. et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24(3):227–235.

Ross-Innes, C. S. et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481(7381):389–393.

Runcie, D. E. and Mukherjee, S. (2013). Dissecting high-dimensional phenotypes with Bayesian sparse factor analysis of genetic covariance matrices. *Genetics*, 194(3):753–767.

Schmidt, M. et al. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Research*, 68(13):5405–5413.

Shi, L. et al. (2006). The microarray quality control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161.

Trial Version Wondershare PDFelement

Simon, R. M., Paik, S., and Hayes, D. F. (2009). Use of archived specimens in evaluation of prognostic and predictive biomarkers. *Journal of the National Cancer Institute*, 101(21):1446–1452.

Smilde, A. K., Kiers, H. A., Bijlsma, S., Rubingh, C., and Van Erk, M. (2008). Matrix correlations for high-dimensional data: the modified RV-coefficient. *Bioinformatics*, 25(3):401–405.

Sørlie, T. et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874.

Sørlie, T. et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, 100(14):8418–8423.

Sotiriou, C. et al. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences*, 100(18):10393–10398.

Sotiriou, C. et al. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, 98(4):262–272.

Symmans, W. F. et al. (2010). Genomic index of sensitivity to endocrine therapy for breast cancer. *Journal of Clinical Oncology*, 28(27):4111–4119.

Tegner, J., Yeung, M. S., Hasty, J., and Collins, J. J. (2003). Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proceedings of the National Academy of Sciences*, 100(10):5944–5949.

Trial Version Wondershare PDFelement

Theodorou, V., Stark, R., Menon, S., and Carroll, J. S. (2013). GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Research*, 23(1):12–22.

Toy, W. et al. (2013). ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nature Genetics*, 45(12):1439–1445.

Tyekucheva, S., Marchionni, L., Karchin, R., and Parmigiani, G. (2011). Integrating diverse genomic data using gene sets. *Genome Biology*, 12(10):R105.

Wang, X. V., Verhaak, R., Purdom, E., Spellman, P. T., and Speed, T. P. (2011). Unifying gene expression measures from multiple platforms using factor analysis. *PloS One*, 6(3).

Zhao, S., Gao, C., Mukherjee, S., and Engelhardt, B. E. (2014). Bayesian group latent factor analysis with structured sparsity. *arXiv preprint arXiv:1411.2698*.