The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1723–1741
https://doi.org/10.1214/21-AOAS1456

BAYESIAN MULTISTUDY FACTOR ANALYSIS FOR HIGH-THROUGHPUT BIOLOGICAL DATA

BY ROBERTA DE VITO[1], RUGGERO BELLIO[2], LORENZO TRIPPA[3] AND GIOVANNI PARMIGIANI[4]

[1]*Department of Biostatistics, Brown University, Roberta_DeVito@brown.edu*
[2]*Department of Economics and Statistics, University of Udine, ruggero.bellio@uniud.it*
[3]*Department of Data Science, Dana Farber Cancer Institute, ltrippa@jimmy.harvard.edu*
[4]*Department of Biostatistics, Harvard T. H. Chan School of Public Health, gp@jimmy.harvard.edu*

This paper analyzes breast cancer gene expression across seven studies to identify genuine and thus replicable gene patterns shared among these studies. Our premise is that genuine biological signal is more likely to be reproducibly present in multiple studies than spurious signal. Our analysis uses a new modeling strategy for the joint analysis of high-throughput biological studies which simultaneously identifies shared as well as study-specific signal. To this end, we generalize the multi-study factor analysis model to handle high-dimensional data and generalize the sparse Bayesian infinite factor model to this context. We provide strategies for the identification of the loading matrices, common and study-specific. Through extensive simulation analysis, we characterize the performance of the proposed approach in various scenarios and show that it outperforms standard factor analysis in identifying replicable signal in all scenarios considered. The analysis of breast cancer gene expression studies identifies clear replicable gene patterns. These patterns are related to well-known biological pathways involved in breast cancer, such as the ER, cell cycle, immune system, collagen, and metabolic pathways. Some of these patterns are also associated with existing breast cancer subtypes, such as LumA, Her2, and basal subtypes, while other patterns identify novel pathways active across subtypes and missed by hierarchical clustering approaches. The R package MSFA implementing the method is available on GitHub.

1. Introduction.

High-throughput molecular assays are now common in biology, resulting in a rich, complex, and diverse collection of high-dimensional datasets and transforming our approach to investigating many diseases, particularly cancer. Breast cancer molecular subtypes based on gene expression patterns, learned from high-throughput studies, have been among the most impactful discoveries brought about by this transformation (Perou et al. (2000)). Breast cancer presents heterogeneity in both clinical and genetic aspects, elucidated by classifying tumors into subtypes and defined by gene expression measurements, with implications for personalized treatment (Masuda et al. (2013)) and prediction (Parker et al. (2009)). To improve inference on subtypes, Sørlie et al. (2001) and Planey and Gevaert (2016), among others, sought to discover gene expression patterns using multiple studies.

Analyses combining high-dimensional biological data from different studies and technologies are crucial to improving the accuracy of conclusions and producing generalizable knowledge. Measurements from high-throughput experiments display variation arising from both biological and artifactual sources (Hicks, Teng and Irizarry (2015)). Within a study, effects driven by a specific laboratory or technology's experimental conditions can be so large to surpass the biological signal (Aach, Rindone and Church (2000), Irizarry et al. (2003), Shi

Received June 2020; revised January 2021.
Key words and phrases. Dimension reduction, factor analysis, gene expression, Gibbs sampling, meta-analysis.

et al. (2006), Kerr (2007)). In breast cancer gene expression studies, batch-related effects are also common, and addressing them improves the ability to learn from multiple studies simultaneously (Larsen et al. (2014)).

A strength of multi-study analyses is that, generally, genuine biological signal is more likely than spurious signal to be present in multiple studies, 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 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 (Huttenhower et al. (2006), Pharoah et al. (2013), Ciriello et al. (2013), Riester et al. (2014), Gao et al. (2014)).

An important goal in our breast cancer application, and in high-dimensional data analysis generally, is the unsupervised identification of latent 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 data from different studies. Despite their utility, these analyses rely, critically, on simplified methods of analysis to capture common signal (Meng et al. (2014), Kim et al. (2017)). For example, Wang et al. (2011) and Edefonti et al. (2012) stack all studies and then perform standard factor analysis (FA). The results capture some common features, but the information about study-specific components are lost, and ignoring theses study-specific characteristics can bias or compromise the conclusions about common signal. Alternatively, it is also common to analyze each study separately and then heuristically identify common characteristics (Hayes et al. (2006)).

Multiple co-inertia analysis (MCIA) (Meng et al. (2014)) is a generalization of co-inertia analysis (CIA) (Dray, Chessel and Thioulouse (2003)) to more than two datasets. CIA 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 factor analysis (MFA) (Abdi, Williams and Valentin (2013)) is an extension of principal component analyses (PCA) which starts with study-specific PCAs and further combines them. Kim et al. (2017) developed a meta-analytic PCA, combining separate study-specific principal components by identifying a common linear subspace through a squared cosine maximization. Other approaches focus on cluster analysis techniques considering multiple datasets, such as allocating subjects in different breast cancer subtypes (Planey and Gevaert (2016), Huo et al. (2016)).

In this work our aim is to systematically identify gene patterns that are replicable across multiple breast cancer gene expression studies. We explore if gene patterns common to all the studies are associate with known, or plausible, sources of biological variation. Then, we identify patterns specific to each study and explore whether they are likely to be batch effects A methodological tool for this task is the multi-study factor analysis (MSFA) (De Vito et al. (2019)), which extends FA to the joint analysis of multiple studies, separately estimating signal replicably 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), Blum et al. (2010), or Runcie and Mukherjee (2013).

The MSFA methodology in De Vito et al. (2019) is limited to low-dimensional 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 our breast cancer gene expression case study, the number of variables exceeds the sample size in all studies, and it is essential to employ regularization through priors or penalties. To address this challenge, we introduce

TABLE 1
The seven datasets considered in our case study. N is the total number of samples; N: ER+ is the number of ER positive patients. 3Q survival is the third quartile of the survival function, in months

Study	Adjuvant Therapy	N	N: ER+	3Q survival
CAL	Chemo, hormonal	118	75	42
MAINZ	none	200	162	120
MSK	combination	99	57	76
EXPO	hormonal	517	325	126
TRANSBIG	none	198	134	143
UNT	none	133	86	151
VDX	none	344	209	44

Bayesian multi-study factor analysis (BMSFA), an extension of MSFA using a Bayesian approach that imposes sparsity. Bayesian approaches naturally provide helpful regularization and offer further advantages, discussed later. We leverage the Bayesian infinite factor model and generalize the multiplicative gamma prior of Bhattacharya and Dunson (2011) to the MSFA setting. We then sample from the posterior distribution, via MCMC, without any ex-ante constraints, such as lower-triangularity (Lopes and West (2004)), on the loading matrices. Although useful inferences can be obtained with careful use of constraints on loading matrices, removing them makes the application of FA much simpler.

We employ two methods to recover loading matrices. The first is based on a matrix spectral decomposition (SD) and closely resembles *principal factor analysis* (Darton (1980)), applied to estimated covariance matrices. The second uses the recently proposed orthogonal procrustes (OP) method (Aßmann, Boysen-Hogrefe and Pape (2016)), which performs an ex post recovery of the estimated loadings by processing the MCMC output, after sampling from the posterior without any restrictions.

The plan of the paper is as follows. Section 2 describes the seven gene expression breast cancer studies that motivated our model. Section 3 introduces the BMSFA framework, describes our prior, our procedure for choosing the number of common and study-specific factors, and methods for estimating the factor loadings. Section 4 presents extensive simulation studies, providing evidence on the performance of BMSFA and comparing it with other methods. Section 5 applies BMSFA to the breast cancer data described in Section 2 and discusses the association between the common pattern we discovered with both biologically relevant pathway and known subtypes. Section 6 contains a discussion.

2. Breast cancer gene expression studies. We consider the collection of publicly available breast cancer microarray studies listed in Table 1 which provides an overview of the studies, the corresponding references, sample sizes, estrogen receptor (ER) status prevalences, and survival summaries. Additional details about these studies, their preprocessing, normalization, curation, criteria for inclusion, and public availability are described in Haibe-Kains et al. (2012). In the Affymetrix technology, genes can be represented by multiple probe-sets. Our analysis considers, for each gene, only the probe-set with maximum mean. As in Bernau et al. (2014), we only consider genes measured in all seven studies (6358 in total) and focus on the 50% of genes with the highest variance, computed combining all studies.

Our goal is to identify stable and replicable latent factors by simultaneously modeling both common components of variability shared across the breast cancer studies and those that are study-specific. The latter could include artifacts and batch effects not adequately removed by the initial data preprocessing as well as study-specific biological signal.

A widely used statistical approach for defining breast cancer subtypes is unsupervised clustering (Sørlie et al. (2001, 2003)). A challenge is to characterize to which extent gene expression distributions, and the resulting subtypes are stable across different studies using different technologies that query the same set of genes (Hayes et al. (2006)). When different studies are considered together, one is likely to encounter significant and unknown sources of study-to-study heterogeneity (Bernau et al. (2014), Zhang et al. (2020)). 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. It is essential to quantify this study-to-study heterogeneity and its impact on the replicability of single-study analyses.

A typical bioinformatics analysis pipeline would attempt to remove variation attributable to experimental artifacts before further analysis. For example, Sørlie et al. (2001) use the SAM (significance analysis of microarrays) algorithm to detect genes not influenced by batch effects and then use this set of genes to perform unsupervised cluster analysis. In general, it is challenging to entirely remove artifactual effects (Draghici et al. (2007)).

3. Bayesian multi-study factor analysis.

This section provides definitions and algorithms for Bayesian multi-study factor analysis (BMSFA), in five parts: (i) Definition of the multi-study factor model; (ii) Specification of the prior; (iii) Model identifiability; (iv) Estimation of the number of latent factors; (v) Estimation of the loading matrices.

3.1. *Model specification.*

We consider S studies, each with the same P variables. Study s, $s = 1, \dots, S$, has n_s subjects and P-dimensional data vectors $\mathbf{x}_{is}$, $i = 1, \dots, n_s$. In MSFA (De Vito et al. (2019)) the variables in study s are decomposed into K factors common to all the studies and J_s factors specific to study s, with $K + J_s \ll P, s = 1, \dots, S$,

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

Here, $\mathbf{f}_{is} \sim N_K(\mathbf{0}, \mathbf{I}_K)$ are the *common* 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 $\mathbf{\Lambda}_s, s = 1, \dots, S$ are the corresponding $P \times J_s$ loading matrices; lastly, $\mathbf{e}_{is}$ is the $p \times 1$ Gaussian noise with covariance $\mathbf{\Psi}_s = \operatorname{diag}(\psi_{1s}^2, \dots, \psi_{ps}^2)$. 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 $\mathbf{\Sigma}_\Phi = \mathbf{\Phi}\mathbf{\Phi}^\top$ represents the component of variability common to all studies, while $\mathbf{\Sigma}_{\Lambda_s} = \mathbf{\Lambda}_s\mathbf{\Lambda}_s^\top$ is the component of variability attributable to the study-specific factors.

We focus on the estimation of the common term $\mathbf{\Sigma}_\Phi$ as well as the components specific to each study $\mathbf{\Sigma}_{\Lambda_s}$, extending our previous work to estimate low-rank covariance matrices.

3.2. *The multiplicative gamma shrinkage prior.*

To specify a prior for the entries ϕ_{pk} of the common factor loading matrix $\mathbf{\Phi}$ and for the entries λ_{pj}^s of the study-specific factor loading matrices $\mathbf{\Lambda}_s$, we generalize the multiplicative gamma shrinkage prior from Bhattacharya and Dunson (2011) to the multi-study setting as follows. The prior for the elements of the common factor loading matrix $\mathbf{\Phi}$ is

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

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

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

$$\lambda^s_{pj} \mid \omega^s_{pj}, \tau^s_j \sim N\big(0, \omega^{s^{-1}}_{pj} \tau^{s^{-1}}_j\big), \quad p = 1, \dots, P, j = 1, 2, \dots, \infty, \text{ and } s = 1, \dots, S,$$

$$\omega^s_{pj} \sim \Gamma\left(\frac{\nu^s}{2}, \frac{\nu^s}{2}\right), \qquad \tau^s_j = \prod_{l=1}^{j} \delta^s_l, \quad \delta^s_1 \sim \Gamma\big(a^s_1, 1\big), \delta^s_l \sim \Gamma\big(a^s_2, 1\big), l \geq 2,$$

where δ^s_l ($l = 1, 2, \dots$) are independent, τ^s_j is the global shrinkage parameter for column j, and ω^s_{pj} is the local shrinkage for the element p in column j.

An important property of these priors is that they result in shrinkage of factor loadings toward zero in a way that tends to increase with the column index of the factor loading.

For each of the error variances ψ_{ps}, $p = 1, \dots, P$, we assume an inverse gamma prior $\psi^{-1}_{ps} \sim \Gamma(a_\psi, b_\psi)$, often used in standard FA (Lopes and West (2004), Bhattacharya and Dunson (2011), Ročková and George (2016)).

3.3. *Model identification and parameter estimation.* Sampling from the posterior distribution of the model parameters is carried out by Markov chain Monte Carlo (MCMC) simulations, fully described in the Supplementary Material (De Vito et al. (2021), Section A). In our setting we encounter two identifiability issues: (i) Label switching (switching a column in any of the factor loading matrices with another) and (ii) Orthogonal rotations of factor loading matrices. Both of these changes result in the same distribution for the data but may affect the samples produced by MCMC. Therefore, averaging unprocessed MCMC posterior samples of the loading matrices is not recommended, as it may lead to misleading summaries. For this reason, we avoid it and propose instead two different approaches to recovering loading matrices, described below.

3.4. *Choosing the number of latent factors.* In applications the number of latent factors is likely to be small compared to the number of variables P. We need to choose the factor dimensions to values suitable for the data at hand. We denote our choice the common latent dimension by K^* and those for the study-specific dimensions by J^*_s. An analogous task for FA is addressed in Bhattacharya and Dunson (2011) with truncation of the number of factors to a finite value, smaller than P. The fact that the shrinkage progressively increases in later columns simplifies this task, compared to alternative shrinkage priors such as the spike-and-slab prior (Carvalho et al. (2008)).

The approach we propose here starts from MCMC simulations from the BMSFA with large values K^{init} and J^{init}_s. Using these posterior samples, we select $K^* \ll P$ and $J^*_s \ll P$ as follows. In the BMSFA model in Section 3.1, the two matrices $\mathbf{\Sigma}_\Phi$ and $\mathbf{\Sigma}_{\Lambda_s}$ are singular with ranks K and J_s, respectively. Since these matrices are symmetric semi-definite, they have K and J_s non-null eigenvalues which are not affected by column label switching or orthogonal rotations. We then compute the estimate the posterior mean of the ordered eigenvalues $\nu_1, \dots, \nu_P$ of $\mathbf{\Sigma}_\Phi$ and choose K^* by setting a threshold on the proportion of the total variance explained, obtained by dividing each eigenvalue by the sum of eigenvalues. In our data analyses we choose a threshold of 5% of the total variance and retain the K^* factors exceeding this threshold. We proceed in the same way for each J_s, $s = 1, \dots, S$. Finally, we run the MCMC again with latent dimensions K^* and J^*_s.

3.5. *Recovering loading matrices.* We discuss two methods for recovering these matrices: spectral decomposition (SD) and orthogonal procrustes (OP).

The SD methods follow principal FA (e.g., Darton (1980)), adapting it to the BMSFA setting. It starts by decomposing the estimated common covariance component as

$$\widehat{\mathbf{\Sigma}}_\Phi = \mathbf{U}\mathbf{N}\mathbf{U}^\top = (\mathbf{U}\mathbf{N}^{1/2})(\mathbf{N}^{1/2}\mathbf{U})^\top.$$

Using this decomposition and for fixed K^* selected, as described above, the common factor loading matrix can be estimated as $\widehat{\mathbf{\Phi}} = \mathbf{U}_{K^*}\mathbf{N}_{K^*}^{1/2}$, where $\mathbf{U}_{K^*}$ corresponds to the matrix of the first K^* eigenvectors of $\mathbf{U}$ and $\mathbf{N}_K$ corresponds to the diagonal matrix of the first K^* eigenvalues. We proceed in the same way for the study-specific factor loading matrices.

The OP approach generalizes (Aßmann, Boysen-Hogrefe and Pape (2016)) which addresses identification by suitably postprocessing samples from the posterior distribution. We illustrate the idea by considering $\mathbf{\Phi}$, though the same procedure is also applied to $\mathbf{\Lambda}_s$. Let $\mathbf{\Phi}^{(1)}, \dots, \mathbf{\Phi}^{(R)}$ be posterior samples. The summary $\widetilde{\mathbf{\Phi}}^*$ is defined by the following constrained minimization:

$$\begin{aligned} \{\widetilde{\mathbf{\Phi}}^*, \widetilde{\mathbf{Q}}^{(r)}\} &= \operatorname*{argmin}_{\mathbf{\Phi}^*, \mathbf{Q}^{(r)}} \sum_{r=1}^{R} \operatorname{tr}\{(\mathbf{\Phi}^{(r)}\mathbf{Q}^{(r)} - \mathbf{\Phi}^*)^\top(\mathbf{\Phi}^{(r)}\mathbf{Q}^{(r)} - \mathbf{\Phi}^*)\}, \\ &\text{subject to } (\mathbf{Q}^{(r)})^\top\mathbf{Q}^{(r)} = \mathbf{I}_{K^*}, \end{aligned} \tag{3}$$

for $r = 1, \dots, R$. Aßmann, Boysen-Hogrefe and Pape (2016), Proposition 3.1, show that the minimization of (3) enforces invariance with respect to rotations that may take place during the MCMC sampling. However, it is possible to multiply both $\mathbf{\Phi}^*$ and $\mathbf{Q}^{(r)}$ by any fixed rotation $\mathbf{Q}^*$ in (3) without changing the value of the loss function. Therefore, $\widetilde{\mathbf{\Phi}}^*$ can be freely rotated post hoc for enhancing interpretation, as commonly done in FA. Aßmann, Boysen-Hogrefe and Pape (2016) prove that the optimization can be carried out by iterating the following two steps until convergence:

1. Given the current value of $\mathbf{\Phi}^*$, compute $\widetilde{\mathbf{Q}}^{(r)} = \mathbf{U}_r\mathbf{V}_r^\top$, where $\mathbf{U}_r$ and $\mathbf{V}_r$ are obtained from the singular value decomposition of $\mathbf{S}_r = (\mathbf{\Phi}^{(r)})^\top\mathbf{\Phi}^*$.
2. Compute $\widetilde{\mathbf{\Phi}}^* = \frac{1}{R}\sum_r \mathbf{\Phi}^{(r)}\widetilde{\mathbf{Q}}^{(r)}$; then, update $\mathbf{\Phi}^*$ by replacing it with $\widetilde{\mathbf{\Phi}}^*$.

The method requires an initial choice for $\mathbf{\Phi}^*$. Aßmann, Boysen-Hogrefe and Pape (2016) suggest that a natural choice is a value from the unconstrained sampler.

4. Simulation results.

In this section we use simulation experiments to assess BMSFA's ability to recover the shared covariance matrix $\mathbf{\Sigma}_\Phi$, the common factor loading matrix $\mathbf{\Phi}$ via both the OP and SD procedures, and the number of common latent dimensions K. We further provide comparisons to standard FA applied to the merged datasets.

We generated 50 collections of datasets for each of the scenarios specified in Table 2 and Figure 1. We fixed $\mathbf{\Phi}$, $\mathbf{\Lambda}_s$ and $\mathbf{\Psi}_s$, for $s = 1, \dots, S$. We consider four scenarios differing in the number of studies, sample sizes, and covariance structure (see Figure 1, column 1). Scenarios 1 and 2 are similar to Zhao et al. (2016): n_s is chosen to be smaller than P to mimic large P and small n_s conditions while operating with a manageable set of variables for visualization and summarization. In Scenario 3 only some of the studies have $P \gg n$. In this scenario, nonzero study-specific factor loadings are randomly set equal to 1 or -1. The motivation behind this scenario is to investigate if our method recovers large study-specific components of variability. In Scenario 4 we mimic the case study data, choosing $S = 7$ and matching the sample sizes to those of Table 1. In all the scenarios considered, we randomly allocate the zeros in each column of $\mathbf{\Phi}$ and $\mathbf{\Lambda}_s$ (Table 2).

We ran the Gibbs sampler (Supplementary Material, Section A) for 30,000 iterations with a burn-in of 20,000 iterations. These iterations are sufficient for convergence and good mixing (Supplementary Material, Section B). The hyperparameters are defined in Table 3. Our

TABLE 2
Distributions used to generate observations in study s, in simulation experiments

$\mathbf{X}_s \sim \mathrm{MVN}(\mathbf{0}, \mathbf{\Sigma}_s)$,
$\mathbf{\Sigma}_s = \mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Lambda}_s\mathbf{\Lambda}_s^\top + \mathbf{\Psi}_s$,
$\mathbf{\Phi}$, $\mathbf{\Lambda}_s$: matrices with ≈ 70 % of zeros with nonzeros elements generated using uniform densities U(0.6, 1) or U(−1, 0.6),
$\mathbf{\Psi}_s$: diagonal elements drawn once from U(0, 1)

TABLE 3
Prior distributions used in the simulation experiments and real data analysis

Common factor loadings: $\omega_{pk} \sim \Gamma(\frac{3}{2}, \frac{3}{2})$,
Study-Specific factor loadings: $\omega^s_{pj} \sim \Gamma(\frac{3}{2}, \frac{3}{2})$,
$\delta_1 \sim \Gamma(2.1, 1)$ and $\delta_l \sim \Gamma(3.1, 1)$ with $l \geq 2$,
$\delta^s_1 \sim \Gamma(2.1, 1)$ and $\delta^s_l \sim \Gamma(3.1, 1)$ with $l \geq 2$,
$\mathbf{\Psi}_s^{-1} \sim \Gamma(1, 0.3)$

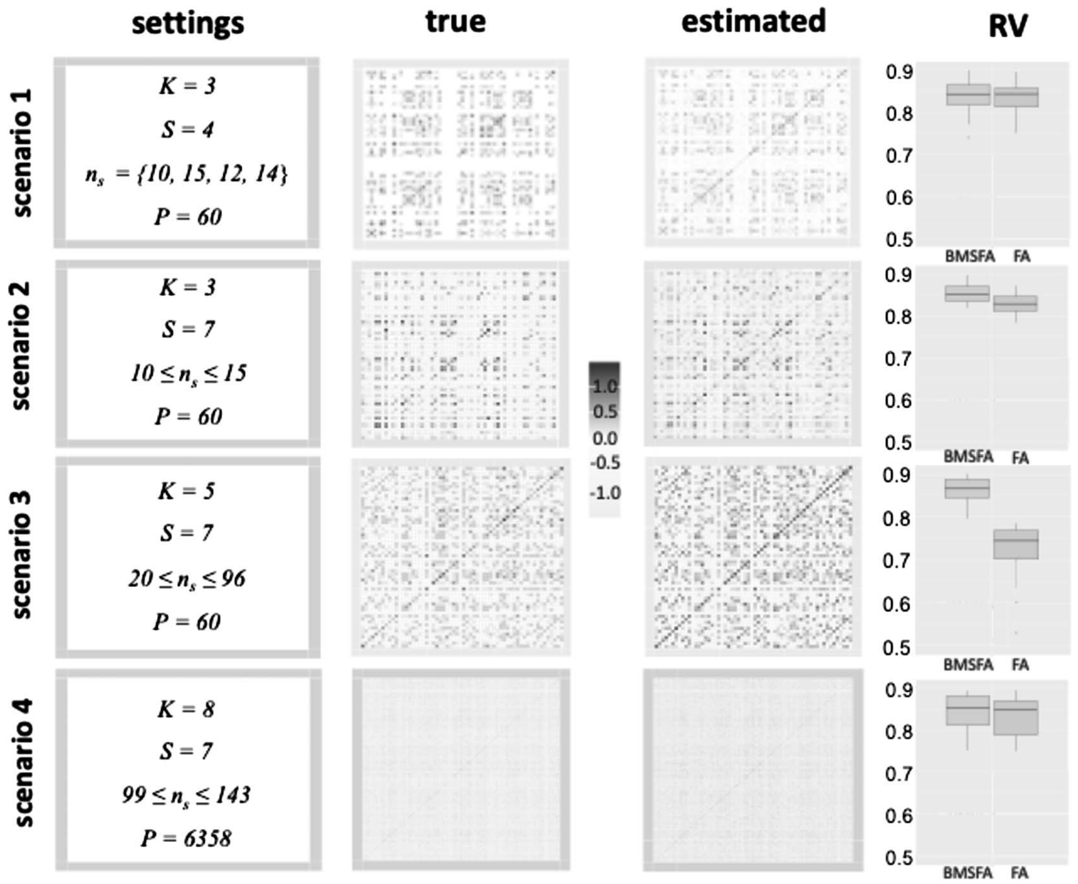

FIG. 1. *Covariance matrices $\mathbf{\Sigma}_\Phi$ and their Bayesian estimates in four simulation scenarios. The first column provides the scenarios' settings. The second column depicts the true $\mathbf{\Sigma}_\Phi = \mathbf{\Phi}\mathbf{\Phi}^\top$, with $\mathbf{\Phi}$ derived as described in Table 2. The third column represents the estimates of $\mathbf{\Sigma}_\Phi$ averaged over the 50 simulated datasets. The fourth column shows the boxplots of RV coefficient replicates between the true and the estimated $\mathbf{\Sigma}_\Phi$, obtained by both our BMSFA and the FA after stacking all the studies in one.*

choice is based on guidelines by Bhattacharya and Dunson (2011) and Durante (2017) as well as numerical experiments exploring the sensitivity of the algorithm for different choices (Supplementary Material, Section C).

We first evaluate BMSFA's ability to recover the covariance component $\mathbf{\Sigma}_{\Phi}$ determined by the common factors as well as the common factor loadings $\mathbf{\Phi}$. Figure 1 compares the true and estimated $\mathbf{\Sigma}_{\Phi}$. To quantify the similarity between $\mathbf{\Sigma}_{\Phi}$ and $\widehat{\mathbf{\Sigma}}_{\Phi}$, we use the RV coefficient (Robert and Escoufier (1976)) which varies in [0, 1]. The closer RV is to 1, the more similar the two matrices are.

Next, we compare BMSFA to Bayesian FA, adopting the same prior (Bhattacharya and Dunson (2011)) for each study. For Bayesian FA we combined all studies into a single dataset, ignoring study labels. The RV coefficients for BMSFA (Figure 1, fourth column) are systematically greater than for FA (Figure 1, fourth column), demonstrating that our BMSFA recovers common factors better than a merged analysis. In Scenario 3 the gap is more pronounced, as study-specific factor loadings are large, and BMSFA can isolate them. Also, the distribution of BMSFA's RV coefficient is less dispersed than FA's. This comparison illustrates that BMSFA identifies the shared signal across the studies and improves its estimation compared to standard Bayesian FA.

Figure 2 presents a similar analysis comparing the true to the estimated factor loadings when using the SD (second column) and the OP procedure (third column). The matrices depicted in Figure 2 are obtained averaging the results of 50 simulations after performing each procedure. For both the SD and OP methods, we applied the varimax rotation to obtain a better defined and more interpretable loading structure.

Both the OP and SD procedures capture the true common factor loading matrix. Figure 2 also shows boxplots of the RV coefficient to quantify the similarity between $\mathbf{\Phi}$ and its estimates, obtained by both SD and OP. The RV coefficient for factor matrices is invariant to rotations, resulting in identical RV coefficients whether we consider the factor loading matrices or their rotations. The RV coefficients are generally high, suggesting that our sampling approaches are appropriate. In our simulation scenarios the performance of SD is at least as good, and sometimes better, than that of OP in recovering the factor loadings. The boxplots of the RV coefficients for the estimates obtained by the SD approach are generally located at higher values than the boxplots of the RV coefficients for the estimates obtained with the OP approach. The SD is also straightforward to compute and interpretable. In our case study we will present results with the SD approach. The OP estimates would lead to the same biological conclusions.

So far, we took K, the number of common factors, to be known. We next focus on determining K via SD of the common covariance component $\widehat{\mathbf{\Sigma}}_{\Phi}$. We use a 5% threshold for the variance component explained by each reported factor relative to the total variability captured by $\widehat{\mathbf{\Sigma}}_{\Phi}$. A threshold of 10% led to the selection of the same number of factors in our simulation settings. As a general rule, we prefer to adopt lower thresholds, as we are more concerned with losing important shared biological factors than with including slightly redundant shared factors. Table 4 shows the results in the four scenarios considered previously. Our method consistently selects the correct dimension for the common factors. Results for the study-specific dimension choices are reported in the Supplementary Material (Section C.2).

5. Breast cancer case study. Our analysis includes five parts: (i) Exploration of the common covariance matrix $\mathbf{\Sigma}_{\Phi}$, (ii) Analysis of the common factor loadings $\mathbf{\Phi}$, (iii) Interpretation of $\mathbf{\Phi}$ in terms of breast cancer subtypes, (iv) Exploration of the study-specific components, $\mathbf{\Sigma}_{\Lambda_s}$; and (v) Illustration of regularization properties.

Covariance Structure. We start by identifying the common covariance structure across the $S = 7$ breast cancer studies of Table 1. We use the prior of Table 3. We visualize the shared

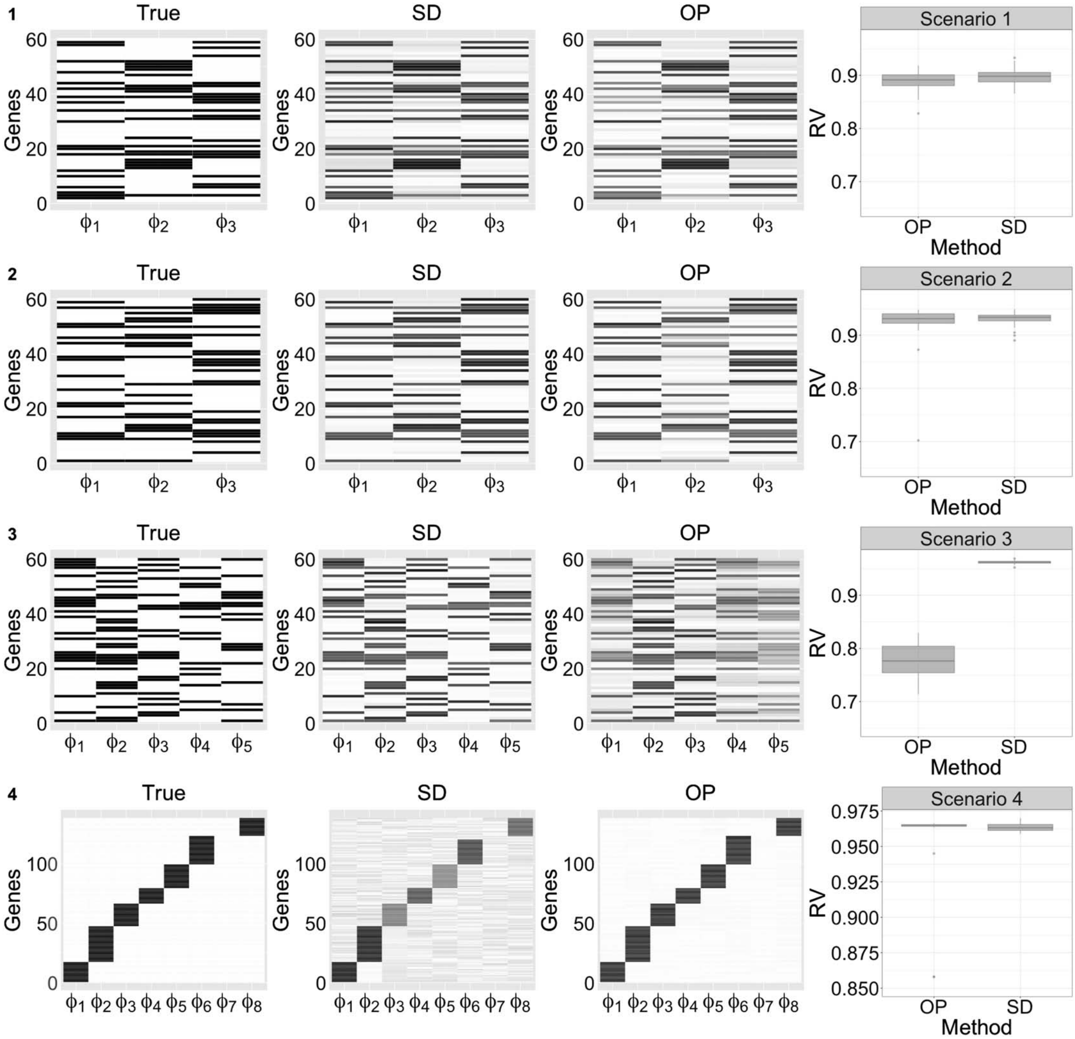

FIG. 2. *Heatmap of the true (first column) and average estimates, across* 50 *collections per scenario, both via SD (second column) and via OP (third column), of the common factor loadings* $\mathbf{\Phi}$. *In Scenario* 4 *we only show common factor loadings* ≥ 0.6, *in absolute value. The right column displays boxplots of the RV coefficients between the true and estimated* $\mathbf{\Phi}$ *via OP and via SD over* 50 *datasets for each scenario.*

TABLE 4

The number of reported common factors K. *We report on* 50 *independent datasets using the same simulation scenarios described in Figure* 1: $K = 3$ *for Scenario* 1, $K = 3$ *for Scenario* 2, $K = 5$ *for Scenario* 3, *and* $K = 8$ *for Scenario* 4. *We proceed by selecting the number of common factors as described in Section* 3.4

Simulation Scenario	Number of simulations in which the reported number of common factors and true value of K are identical
1	48
2	50
3	49
4	49

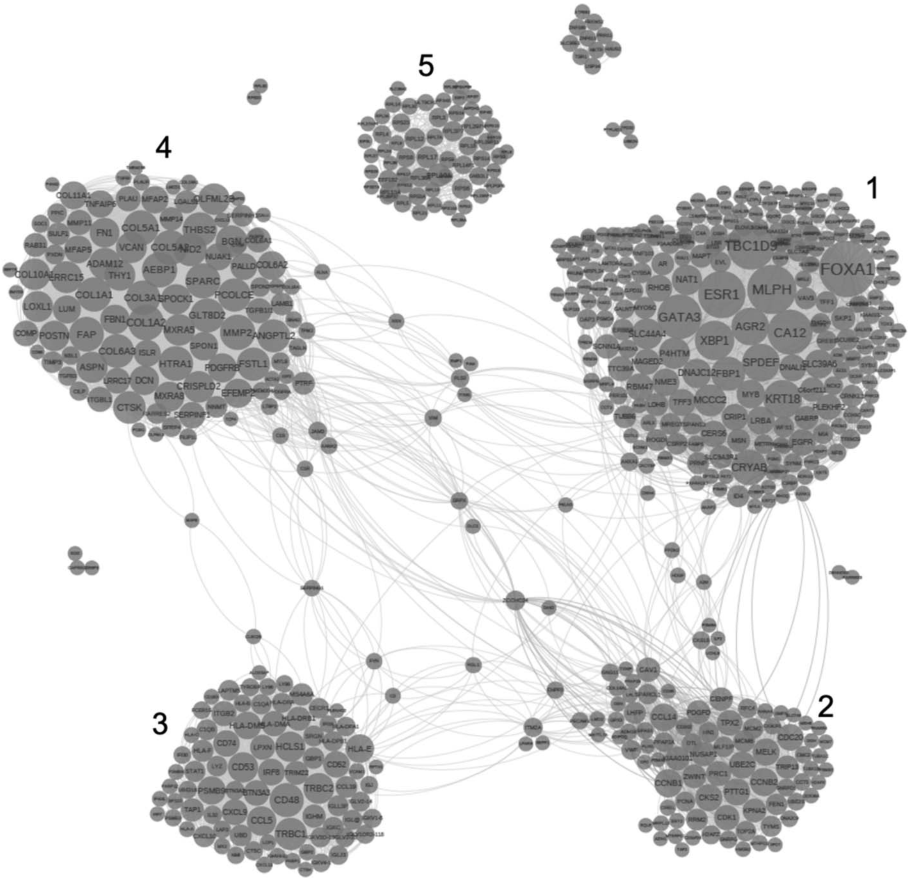

FIG. 3. *Shared gene coexpression network based on* $\widehat{\boldsymbol{\Sigma}}_\Phi$ *across the seven studies* (*Table* 1) *obtained using* Gephi (*Bastian, Heymann and Jacomy* (2009)). *We include edges between two genes if the corresponding element in the shared covariance matrix* $\widehat{\boldsymbol{\Sigma}}_\Phi$ *is greater than* 0.5 *in absolute value. We do not graph the* 15% *of the genes which have no surviving edges. Numbers refer to clusters identified by the ForceAtlas*2 *algorithm* (*Jacomy et al*., 2014).

coexpression relations via a network built using $\widehat{\boldsymbol{\Sigma}}_\Phi$ and thus representative of all studies (Figure 3). A gene coexpression network is an undirected graph. Nodes represent genes, and edges represent the degree of coexpression for pairs of genes. The size of each node is proportional to the number of connections between the corresponding gene and others within the same cluster.

The five clusters of genes displayed in Figure 3 are all associated with important biological processes. Cluster 1 includes the estrogen receptor (*ESR1*) which has an important role in the biology and treatment of breast cancer (Sørlie et al. (2001)). This cluster captures a large transcriptional program. Important transcription factors and activators in this cluster include *GATA3*, *XBP1*, and *FOXA1*, upon which we return later. In breast cancer a strong and positive coexpression of *GATA3*, *XBP1*, and *FOXA1* with *ESR1* is noted in luminal A breast tumors (Sørlie et al. (2003)). Cluster 2 appears to be related to the cell cycle. One of the most connected genes in this cluster is *CCNB1* which encodes cyclin *B*. Cyclins are prime cell cycle regulators. In particular, mitotic cyclins *A* and *B* and their dependent kinase are overexpressed in breast cancer (Basso et al. (2002)). Two other notable genes in this cluster are *CDK1*, a kinase which depends on cyclins, and *CDC20*, a gene related to the metaphase and anaphase of the cell cycle. Most genes in cluster 3 are related to the regulation of the immune response, as exemplified by the prominence of *CD* and *HLA* genes which are essential for the

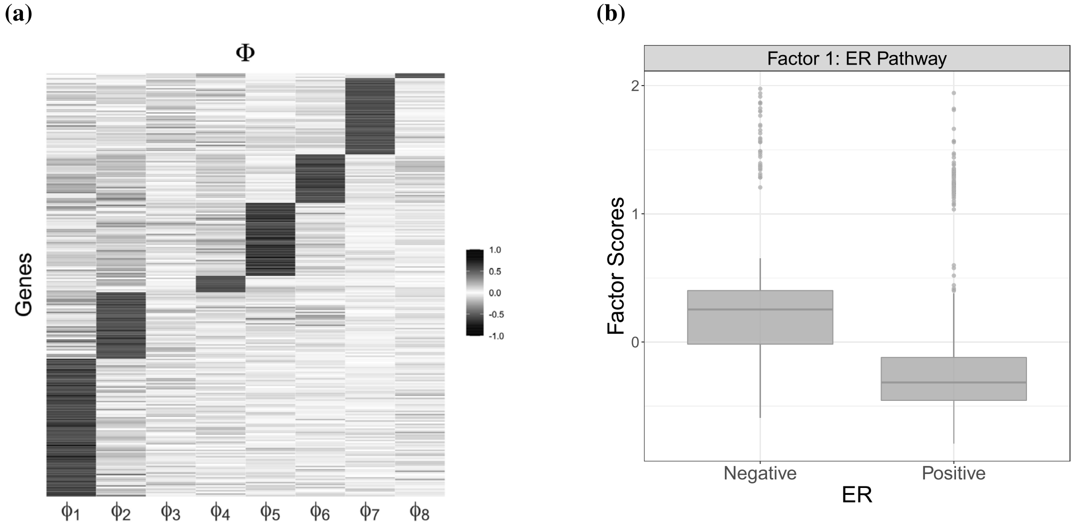

FIG. 4. (*a*) *Heatmap of* $\widehat{\boldsymbol{\Phi}}$ *obtained with BMSFA across all studies. We only show common factor loadings greater than* 0.6 *in absolute value. The factor loadings are ordered by the variance explained. Overall, the eight common factors explain* 83% *of the total common variance.* (*b*) *Boxplots of the estimated common factor score for Factor* 1, *associated to the ER pathway. The boxplot on the right* (*left*) *corresponds to ER negative* (*positive*) *samples.*

immune system function. Cluster 4 includes several genes expressed by the connective tissue, such as collagen genes. Finally, Cluster 5 is largely composed of the *RP* genes codifying the ribosome which synthesizes proteins. Dysregulation of ribosome function is related to tumor progression in breast cancer (Belin et al. (2009)).

Shared Factor Loadings. We then estimate the common factor loadings using the SD of $\widehat{\boldsymbol{\Sigma}}_\Phi$ (Figure 4(a)). Our method chooses $K = 8$ through the eigenvalue analysis of $\widehat{\boldsymbol{\Sigma}}_\Phi$. We further elucidate the relation of common factor loadings with biological pathways through a gene set enrichment analysis (GSEA) (Mootha et al. (2003)), investigating whether gene sets capturing pathway memberships are enriched among the loadings in the sense of having larger loadings than a random set of that size. The sets capturing biological pathways are from `reactome.org`. We use the package `RTopper` in `R` in `Bioconductor`, following the method of Tyekucheva et al. (2011). Five common factor loadings out of eight are significantly enriched in specific biological pathways, while the remaining three relate to multiple pathways. These five are more straightforward to interpret and map nicely to the clusters of Figure 3 though not in the same order. Factors 1, 2, 3, 5, and 7 are significantly enriched, respectively, for gene sets related to ER, cell cycle, immune system, collagen, and metabolic pathways (Supplementary Material E.1).

The first common factor, $\boldsymbol{\Phi}_1$, is enriched for the ER pathway, which is expressed in 70% of breast cancers (Reinert et al. (2017)), and it is essential for both the classification of tumors into subtypes and therapeutic strategies (Reinert et al. (2017)). We explore whether ER positive tumors (ER+) are projected in distinct regions of the latent space, compared to ER negative tumors (ER-). Figure 4(b) displays the estimated Factor 1 scores. As expected, the distributions of scores for the ER+ and ER- samples differ markedly (Figure 4(b)).

Subtypes. Following early gene expression clustering efforts (Sørlie et al. (2001, 2003)), breast cancer is often categorized molecularly into four subtypes: basal-like, luminal A (LumA), luminal B (LumB), and human epidermal growth factor receptor 2 (HER2)-enriched. These subtypes present differences in clinical outcome and prognosis. Next, we describe the association between the common factor scores we identified and these subtypes. We first classify patients into subtypes, following the hierarchical clustering approach Sørlie

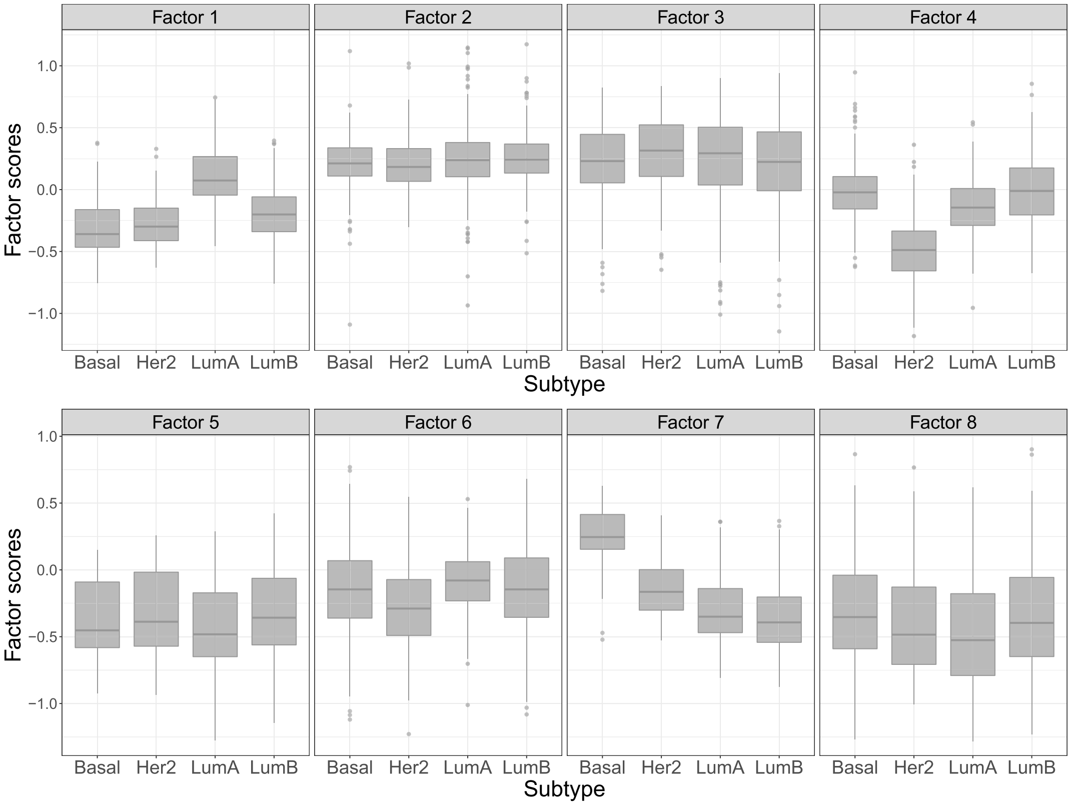

FIG. 5. *Distributions of common factor scores for breast cancer molecular subtypes*: *Basal*, *Her2*, *Luminal A*, *and Luminal B*.

et al. (2001). Then, we illustrate the distribution of the factor scores in these four subtypes (Figure 5). Sørlie et al. (2001) also suggest the presence of a so-called "normal-like" subtype to which 28 out of our 1445 patients are assigned. Unfortunately, 17 of these 28 patients are concentrating on study VDX. As a result, this subtype is highly confounded with study. Inclusion of this confounded subtype is likely to impair the interpretation of both common and study-specific effects. Therefore, these 28 patients were excluded.

For each common factor a boxplot in Figure 5 summarizes its distribution within a specific subtype. Distributions of Factors 1, 4, and 7 differ in exactly one subtype, while the boxplots of all the other factors are similar across subtypes. Factor 1 presents higher scores for patients classified in the breast cancer LumA subtype, compared to all the others. Factor 1 has the highest factor loadings (≥ 0.8) in the ESR1, GATA3, and XBP1 genes. Sørlie et al. (2003) shows that these three genes (among nine) are highly expressed in luminal subtype A tumors. Factor 4 presents lower scores for patients classified in the Her2 subtype. This factor is enriched for metabolic pathways, regulation of actin cytoskeleton, and MAPK signaling. These three pathways have been found to be associated with HER2 subtype breast cancer (Yang et al. (2016)). Factor 7 presents higher scores for patients classified in the basal subtype, compared to patients classified in the luminal subtypes (LumA, LumB). Factor 7 is enriched in metabolic pathways, such as glycolysis and oxidative phosphorylation. These two metabolic pathways reveal an opposite association with the basal and the luminal subtypes. In particular, glycolysis is lower in the luminal subtypes and higher in the basal subtype (Tyanova et al. (2016)). On the contrary, the oxidative phosphorylation is higher in the luminal subtypes and lower in the basal subtype (Tyanova et al. (2016)). Effects of the same direction are observed in our analysis.

The remaining five common factors (2, 3, 5, 6, 8) do not show any differences in scores across breast cancer subtypes. These are interpretable in terms of enrichment for pathways essential to all subtypes. In our opinion this reveals the power of our multi-study method to detect a biological signal that is important, but not captured with cluster-based breast cancer subtype analysis, because it is shared by multiple subtypes. In particular, Factor 2 is significantly enriched for the cell cycle, DNA replication, and cell proliferation pathways. The replication rate does not vary significantly, on average, across breast cancer subtypes but does across tumors. Factor 2 likely captures this variation. Factor 3 is associated with immune system activity. Similarly, immune regulation of the tumor microenvironment can be relevant within each of the subtypes. Factor 5 is related to collagen accumulation which is increased by insulin and insulin-like growth factor I. The insulin-like growth factor-I is present in all breast cancer subtypes (Law et al. (2008)). Factor 6 presents negative and large (≤ -0.7) loadings in the *IRF8*, *CD53*, and *CD48* genes. These three genes with other 19 genes compose the T-cell metagene and are involved in multiple biological functions. For example, T-cell metagene expression is associated with a reduced risk of distant metastases in all cancer subtypes, possibly explaining the absence of variability of this factor across breast cancer subtypes. Finally, Factor 8 has high (≥ 0.7) positive loading in *CLK2*. This gene is amplified and overexpressed in a significant fraction of breast tumor (Yoshida et al. (2015)).

These analyses indicate that the BMSFA is able to capture biological signal shared across all studies and also to reveal new insight into breast cancer subtype analysis, compared to other unsupervised analyses based on hierarchical clustering.

Study-specific factor loadings. Next, we focus on the study-specific loadings. We compare two studies from Table 1: the CAL study (Figure 6(a)), which includes patients whose cancer has reached the axillary lymph nodes (positive nodal status) and not (negative nodal status), and the MAINZ study (Figure 6(b)) which includes only patients with negative nodal status. Nodal status is associated with cancer ability to spread (Carter, Allen and Henson (1989)), is part of clinical staging, and is pivotal for treatment decisions in breast cancer. In the CAL study, 54% of the patients presents a high histological grade, while in the MAINZ study only 17% do.

We first perform GSEA for the study-specific factor loadings. The CAL study-specific factors are significantly enriched with "Downregulation of TGF-β receptor signaling", a pathway related to an increase of lymph node metastasis and to a more aggressive clinical behavior, including shorter survival. This pathway defines a dimension likely to discriminate between node-positive and node-negative tumors. The gene network (Figure 6(a)) supports these results: one crucial gene in this network is *PCBP1*, previously found to have a role in cancer progression in conjunction with TGF-β (Ooshima, Park and Kim (2019)).

In contrast, the TGF-β pathway is not significantly enriched in the MAINZ study's factors, in line with survival and histological differences between the two studies. In the MAINZ study the loadings are associated with two pathways: "Cytokine signaling in immune system" and "Cell communications". These pathways, previously found in lymph node-negative patients, are associated with antitumor response (Chatterjee et al. (2018)). The gene network for the MAINZ study (Figure 6(b)) also emphasizes these pathways, for example, via the role of *CD46* and *CAPRIN1*, which are involved in the immune system and the cell cycle pathway, repectively.

Regularization. An essential feature of BMSFA is the regularization of the common factor loadings. To illustrate this in more detail, we compare BMSFA to MSFA which uses MLE for parameter estimation. Because MSFA cannot deal with $p > n$, we focus on the 63 genes in the immune system pathway whose loadings are compared in Figure 7. BMSFA regularizes common factor loadings by shrinking small and moderate MLE loadings to zero (Figure 7, left panel). In this regard the action of the prior is similar to that of a factor rotation. To

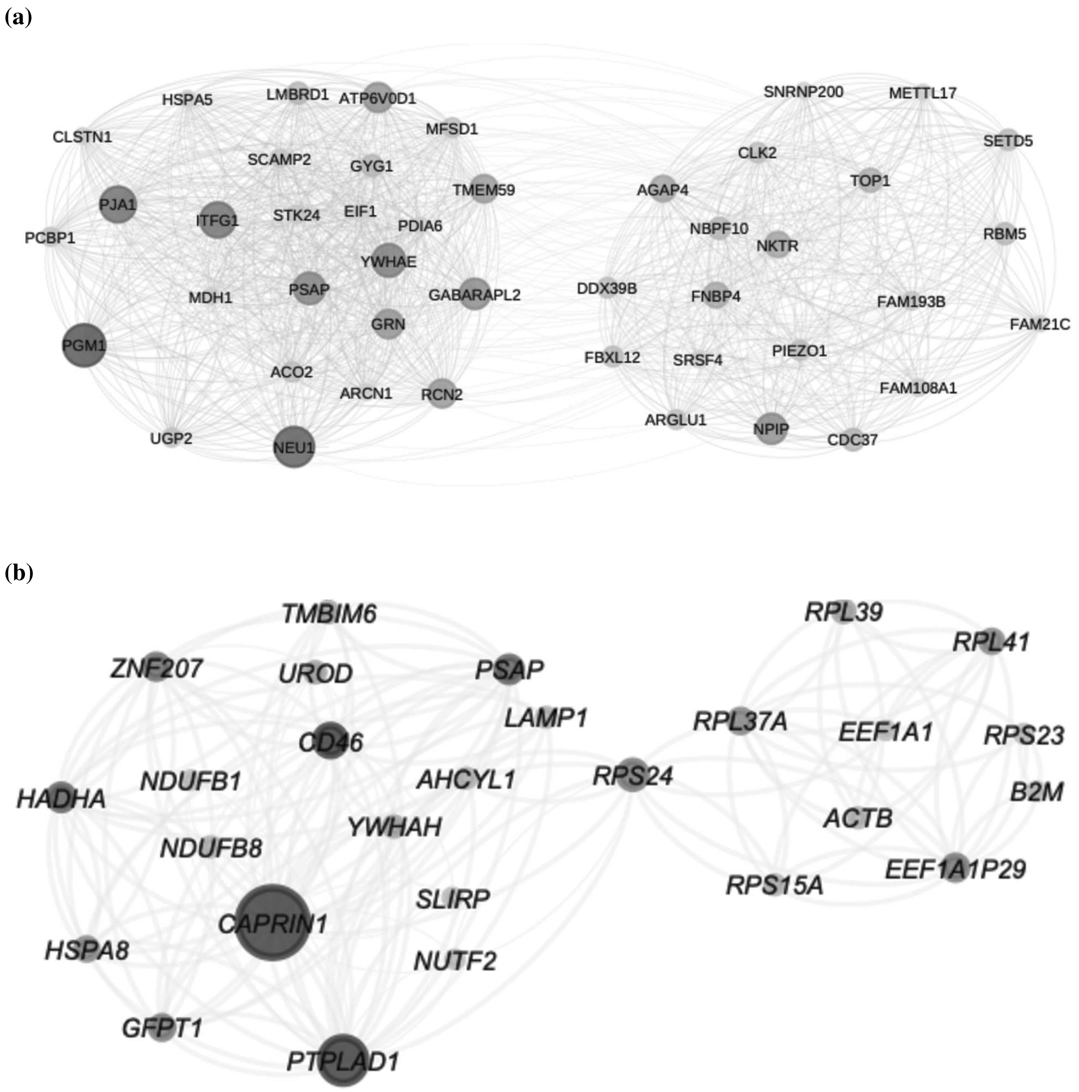

FIG. 6. *Study-specific gene coexpression network based on* $\widehat{\boldsymbol{\Sigma}}_{\Lambda_s}$ *of both the CAL* (*a*) *and MAINZ* (*b*) *study obtained using* Gephi. *We include edges between two genes if the corresponding element in the study-specific covariance matrix* $\widehat{\boldsymbol{\Sigma}}_{\Lambda_s}$ *is greater than* 0.5 *in absolute value.*

further illustrate this point, we rotate the loadings obtained with the MLE using the varimax rotation (Kaiser (1958)), and we compare it with the BMSFA (Figure 7 right panel). The BMSFA loadings in the low range are far more similar to the MLE estimates rotated by varimax, although the overall correlations only differ modestly as a result of in influence of large factors.

We performed additional simulations (Supplementary Material, Section D.2) that support these results, as does earlier research where penalty functions, based on the varimax criterion, were used for estimating rotated principal components (Park (2005)).

6. Discussion. In this article we seek to identify patterns of transcriptional variation in breast tumors that are replicable across multiple published studies. We are able to identify latent factors that recapitulate known breast cancer subtypes as well as novel latent factors. Some of these represent essential biological functions across tumors and may be important in identifying shared vulnerabilities across subtypes. Others are unique to individual studies,

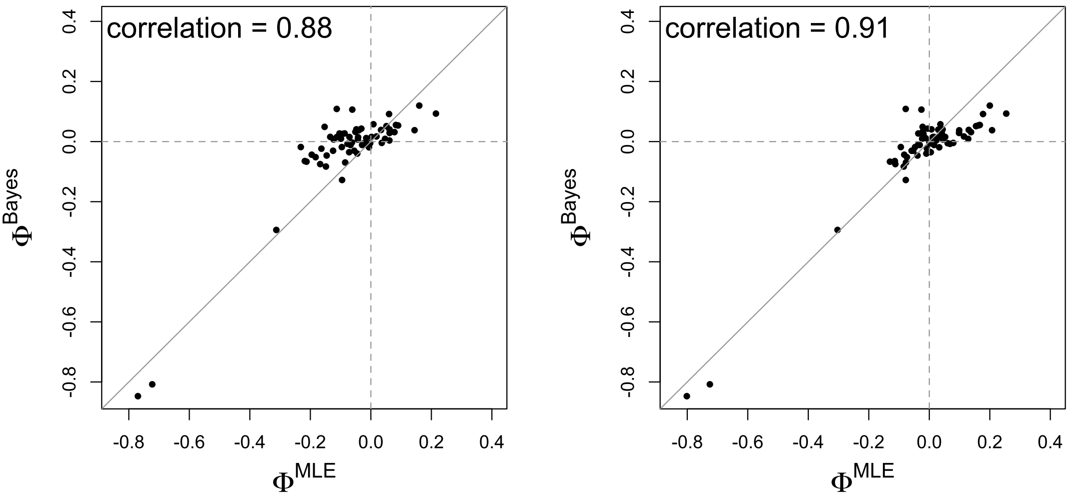

FIG. 7. *Comparisons between common factor* 1 *loadings obtained both via MLE* (*horizontal*) *and BMSFA* (*vertical*). *On the left are the original loadings while on the right the MLE loadings have been rotated using varimax.*

and in some cases can be interpreted based on sources of biological variation unique to the study's design.

To achieve these results, we proposed and implemented a novel and general Bayesian framework for the unsupervised analysis of high-dimensional biological data across multiple studies. We address the broader 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 shows considerable promise in modeling sparsity and enhancing interpretability via shrinkage. Systematic collections of gene expression data similar to those considered in our case study are widely available. These are obtained with technologies and study designs that evolved over time, emphasizing the importance of cross-study variation. Beyond gene expression the BMSFA method may have broad applicability in a wide variety of studies, including genome-wide association studies, metabolomics, electronic medical record (EMR), and high dimensional epidemiological data.

To characterize the common signal, we focus on both $\boldsymbol{\Sigma}_\Phi$ and $\boldsymbol{\Phi}$. MCMC samples of $\boldsymbol{\Sigma}_\Phi$ can be used to construct gene relevance networks (Butte et al. (2000)) as well as graphical model for gene association networks (Ni et al. (2018)). The matrix $\boldsymbol{\Phi}$ is particularly useful when one is interested in associating a phenotype label to the factors. For example, in dietary analyses we used the MSFA to derive common ($\boldsymbol{\Phi}$) and study-specific ($\boldsymbol{\Lambda}_s$) factors in seven studies and then investigated their association with cancer (De Vito et al. (2019)).

We also developed two ways to recover the factor loading matrix. The SD approach is simple and straightforward and, as noted by Darton (1980), only requires estimating the $\boldsymbol{\Psi}_s$ matrices, and fixing the latent factor dimensions. The OP approach, in contrast, requires the entire MCMC output. A potential limitation arises in scenarios with large factor loadings, present only in the first few loading matrix columns. We evaluated this limitation in simulations (Supplementary Material, Section D.1) and observe a moderate deterioration, despite which the method was still able to produce good estimates. Moreover, post hoc application of the varimax rotation improved the results, though those for the SD method remained more accurate. More generally, the SD method outperformed the OP method across the scenarios we considered. The computational time for SD was consistently less than for OP (see Section B, Table S2). In summary, both the theoretical properties and the simulation results support our preference for SD in the data analysis.

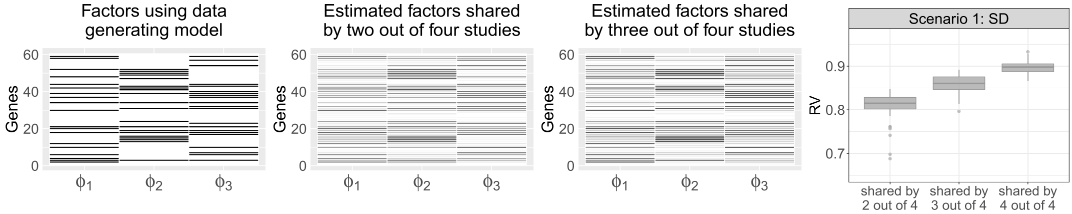

FIG. 8. *Heatmap of the true common loading matrix*, $\mathbf{\Phi}$ (*first column*). *Heatmap of the estimated* $\mathbf{\Phi}$ (*mean of* 50 *simulations*) *via BMSFA after generating the data with a factors partially shared by two studies over four* (*second column*) *and by three studies over four* (*third column*). *The right column displays three boxplots showing*, *respectively*, *the RV between the estimated* $\mathbf{\Phi}$ *and the true factors partially shared by two studies*, *the RV between the estimated* $\mathbf{\Phi}$ *and the true factors partially shared by three studies*, *and*, *finally*, *the RV between the estimated* $\mathbf{\Phi}$ *and the true common factors shared by all the studies over* 50 *datasets in Scenario* 1.

Recently, Roy et al. (2019) introduced perturbed factor analysis (PFA), a method that shares BMSFA's goal to improve replicability across studies. In summary, PFA introduces a matrix Q of perturbations across different groups/studies. PFA has the advantage of estimating a smaller number of parameters; hence, it can be particularly helpful when the aim is to directly estimate the common factor loading matrix. On the other hand, BMSFA relies on explicit estimation of the study-specific factor loadings, achieved directly in the Gibbs sampling algorithm. BMSFA estimates both the study-specific loadings and the study-specific covariance matrix while PFA quantifies the magnitude of the perturbation in each study with calculation and derivation of the Frobenius norm. Additionally, PFA controls the level of perturbation across groups via a hyperparameter. As a result, PFA requires proper tuning in order to precisely estimate the study-specific loading matrix, hence the requirement for a cross validation analysis. BMSFA and PFA are complementary methods addressing related, but not identical, challenges.

The BMSFA methodology shrinks the latent factor loadings and provides a strategy to deal with high-dimensional data. However, the performance may be sensitive to several choices of prior hyperparameters. We recommend sensitivity analyses, such as those presented here, which provide a practical strategy to study the shrinkage behavior in a multi-study setting.

We also explored the robustness of BMSFA, assuming that in the data generation some factors are not common to all the studies but instead are only common to a subset of studies. We evaluated the sensitivity of the estimates in two scenarios, the first with three common factors partially shared by two studies out of four and the second with three common factors partially shared by three studies out of four. The simulation settings of these two scenarios were the same as for Scenario 1. In Figure 8 we examine data generated by three factors partially shared by two studies out of four and by three factors partially shared by three studies out of four. The common signal is still captured, though it is also apparent that a higher number of shared factors leads to clearer signal. When the true factors are partially shared by three studies, the estimated common factors are more precise than when the true factors partially shared by two studies. To further corroborate this observation, we also computed the RV coefficients (Figure 8, column 4). The RV coefficient decreases from 0.9 to 0.8 on the average, although it remains high enough to suggest that BMSFA still captures the signal. This is an interesting setting that needs further explorations and analyses. Partially shared factors may be common, especially in the presence of different platforms, study design, and/or different treatments. Thus, it could be useful to develop an approach to estimate common, study-specific partially shared factors.

In summary, we found BMSFA to be effective in our application, and we hope that it 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 field and across data science.

Acknowledgments. We are grateful to Bianca Dumitrascu for help with Figure 3.

Funding. Research supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS_003 (RB), the U.S.A.'s National Science Foundation Grant DMS-1810829 (LT and GP), and the U.S.A.'s National Institutes of Health grant NCI-5P30CA006516-54 (GP).

SUPPLEMENTARY MATERIAL

Supplement 1 (DOI: 10.1214/21-AOAS1456SUPPA; .zip). Supplement 1 includes all the codes to generate the simulation and the data analyses presented in the manuscript. It includes the data files used in the paper, and the simulations for the different scenarios considered. It also includes the R folder with the functions used for the Gibbs sampling algorithm.

Supplement 2 (DOI: 10.1214/21-AOAS1456SUPPB; .pdf). Supplement 2 is a folder containing the pdf file with further simulations and analyses. In particular, the file includes (A) explicit expression for the steps of the Gibbs sampling algorithm; (B) additional details on computational efficiency and convergence of the MCMC; (C) an examination of sensitivity to the choice of hyperparameters; (D) additional simulation settings that could not be covered in the main text and (E) Further analysis of breast cancer case study.

REFERENCES

AACH, J., RINDONE, W. and CHURCH, G. M. (2000). Systematic management and analysis of yeast gene expression data. *Genome Res.* **10** 431–445.

ABDI, H., WILLIAMS, L. J. and VALENTIN, D. (2013). Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. *Wiley Interdiscip. Rev.: Comput. Stat.* **5** 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. *J. Econometrics* **192** 190–206. MR3463672 https://doi.org/10.1016/j.jeconom.2015.10.010

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** 1159–1166.

BASTIAN, M., HEYMANN, S. and JACOMY, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Third International AAAI Conference on Weblogs and Social Media*.

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** e7147.

BERNAU, C., RIESTER, M., BOULESTEIX, A.-L., PARMIGIANI, G., HUTTENHOWER, C., WALDRON, L. and TRIPPA, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* **30** i105–i112.

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

BLUM, Y., MIGNON, G. L., LAGARRIGUE, S. and CAUSEUR, D. (2010). A factor model to analyze heterogeneity in gene expression. *BMC Bioinform.* **11** 368. https://doi.org/10.1186/1471-2105-11-368

BUTTE, A. J., TAMAYO, P., SLONIM, D., GOLUB, T. R. and KOHANE, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA* **97** 12182–12186.

CARTER, C. L., ALLEN, C. and HENSON, D. E. (1989). Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer* **63** 181–187.

CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. MR2655722 https://doi.org/10.1198/016214508000000869

CHATTERJEE, G. et al. (2018). Molecular patterns of cancer colonisation in lymph nodes of breast cancer patients. *Breast Cancer Res.* **20** 143.

CIRIELLO, G., MILLER, M. L., AKSOY, B. A., SENBABAOGLU, Y., SCHULTZ, N. and SANDER, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45** 1127–1133.

DARTON, R. A. (1980). Rotation in factor analysis. *J. R. Stat. Soc., Ser. D, Stat.* **29** 167–194.

DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2019). Multi-study factor analysis. *Biometrics* **75** 337–346. MR3953734 https://doi.org/10.1111/biom.12974

DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2021). Supplement to "Bayesian multi-study factor analysis for high-throughput biological data." https://doi.org/10.1214/21-AOAS1456SUPPA, https://doi.org/10.1214/21-AOAS1456SUPPB

DE VITO, R. et al. (2019). Shared and study-specific dietary patterns and head and neck cancer risk in an international consortium. *Epidemiology* **30** 93–102.

DRAGHICI, S. et al. (2007). A systems biology approach for pathway level analysis. *Genome Res.* **17** 1537–1545.

DRAY, S., CHESSEL, D. and THIOULOUSE, J. (2003). Co-inertia analysis and the linking of ecological data tables. *Ecology* **84** 3078–3089.

DURANTE, D. (2017). A note on the multiplicative gamma process. *Statist. Probab. Lett.* **122** 198–204. MR3584158 https://doi.org/10.1016/j.spl.2016.11.014

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. *Ann. Oncol.* **23** 1869–1880.

GAO, J., CIRIELLO, G., SANDER, C. and SCHULTZ, N. (2014). Collection, integration and analysis of cancer genomic profiles: From data to insight. *Curr. Option Genet. Dev.* **24** 92–98.

HAIBE-KAINS, B. et al. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *J. Natl. Cancer Inst.* **104** 311–325.

HAYES, D. N. et al. (2006). Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *J. Clin. Oncol.* **24** 5079–5090.

HICKS, S. C., TENG, M. and IRIZARRY, R. A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *BioRxiv*.

HUO, Z., DING, Y., LIU, S., OESTERREICH, S. and TSENG, G. (2016). Meta-analytic framework for sparse K-means to identify disease subtypes in multiple transcriptomic studies. *J. Amer. Statist. Assoc.* **111** 27–42. MR3494636 https://doi.org/10.1080/01621459.2015.1086354

HUTTENHOWER, C. et al. (2006). A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics* **22** 2890–2897.

IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264.

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

KERR, K. F. (2007). Extended analysis of benchmark datasets for agilent two-color microarrays. *BMC Bioinform.* **8** 371. https://doi.org/10.1186/1471-2105-8-371

KIM, S., KANG, D., HUO, Z., PARK, Y. and TSENG, G. C. (2017). Meta-analytic principal component analysis in integrative omics application. *Bioinformatics* **34** 1321–1328.

LARSEN, M. J., THOMASSEN, M., TAN, Q., SØRENSEN, K. P. and KRUSE, T. A. (2014). Microarray-based RNA profiling of breast cancer: Batch effect removal improves cross-platform consistency. *BioMed Research International* **2014**.

LAW, J. H. et al. (2008). Phosphorylated insulin-like growth factor-i/insulin receptor is present in all breast cancer subtypes and is related to poor survival. *Cancer Res.* **68** 10238–10246.

LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14** 41–67. MR2036762

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

MENG, C., KUSTER, B., CULHANE, A. C. and GHOLAMI, A. M. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinform.* **15** 162. https://doi.org/10.1186/1471-2105-15-162

MOOTHA, V. K. et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34** 267–273.

NI, Y., MÜLLER, P., ZHU, Y. and JI, Y. (2018). Heterogeneous recipocal graphical models. *Biometrics* **74** 606–615. MR3825347 https://doi.org/10.1111/biom.12791

OOSHIMA, A., PARK, J. and KIM, S.-J. (2019). Phosphorylation status at Smad3 linker region modulates transforming growth factor-β-induced epithelial-mesenchymal transition and cancer progression. *Cancer Science* **110** 481.

PARK, T. (2005). A penalized likelihood approach to rotation of principal components. *J. Comput. Graph. Statist.* **14** 867–888. MR2211371 https://doi.org/10.1198/106186005X78134

PARKER, J. S. et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27** 1160–1167.

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

PHAROAH, P. D. P. et al. (2013). GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat. Genet.* **45** 362–370.

PLANEY, C. R. and GEVAERT, O. (2016). CoINcIDE: A framework for discovery of patient subtypes across multiple datasets. *Gen. Med.* **8** 27. https://doi.org/10.1186/s13073-016-0281-4

REINERT, T., SAAD, E. D., BARRIOS, C. H. and BINES, J. (2017). Clinical implications of ESR1 mutations in hormone receptor-positive advanced breast cancer. *Front Oncol* **7** 26. https://doi.org/10.3389/fonc.2017.00026

RIESTER, M. et al. (2014). Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *J. Natl. Cancer Inst.* **106**.

ROBERT, P. and ESCOUFIER, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **25** 257–265. MR0440801 https://doi.org/10.2307/2347233

ROČKOVÁ, V. and GEORGE, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *J. Amer. Statist. Assoc.* **111** 1608–1622. MR3601721 https://doi.org/10.1080/01621459.2015.1100620

ROY, A., LAVINE, I., HERRING, A. H. and DUNSON, D. B. (2019). Perturbed factor analysis: Improving generalizability across studies. Preprint. Available at arXiv:1910.03021.

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

SHI, L. et al. (2006). The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24** 1151–1161.

SØRLIE, T. et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* **98** 10869–10874.

SØRLIE, T. et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* **100** 8418–8423.

TYANOVA, S., ALBRECHTSEN, R., KRONQVIST, P., COX, J., MANN, M. and GEIGER, T. (2016). Proteomic maps of breast cancer subtypes. *Nat. Commun.* **7** 10259. https://doi.org/10.1038/ncomms10259

TYEKUCHEVA, S., MARCHIONNI, L., KARCHIN, R. and PARMIGIANI, G. (2011). Integrating diverse genomic data using gene sets. *Genome Biol.* **12** R105. https://doi.org/10.1186/gb-2011-12-10-r105

WANG, X. V., VERHAAK, R. G. W., PURDOM, E., SPELLMAN, P. T. and SPEED, T. P. (2011). Unifying gene expression measures from multiple platforms using factor analysis. *PLoS ONE* **6** e17691. https://doi.org/10.1371/journal.pone.0017691

YANG, F., LYU, S., DONG, S., LIU, Y., ZHANG, X. and WANG, O. (2016). Expression profile analysis of long noncoding RNA in HER-2-enriched subtype breast cancer by next-generation sequencing and bioinformatics. *OncoTargets and Therapy* **9** 761.

YOSHIDA, T. et al. (2015). CLK2 is an oncogenic kinase and splicing regulator in breast cancer. *Cancer Res.*.

ZHANG, Y., BERNAU, C., PARMIGIANI, G. and WALDRON, L. (2020). The impact of different sources of heterogeneity on loss of accuracy from genomic prediction models. *Biostatistics* **21** 253–268. MR4133359 https://doi.org/10.1093/biostatistics/kxy044

ZHAO, S., GAO, C., MUKHERJEE, S. and ENGELHARDT, B. E. (2016). Bayesian group factor analysis with structured sparsity. *J. Mach. Learn. Res.* **17** Paper No. 196, 47. MR3580349