Università degli studi di Udine

Multi-study factor analysis

Original

Availability:
This version is available http://hdl.handle.net/11390/1147470 since 2020-03-19T10:54:08Z

Publisher:

Published
DOI:10.1111/biom.12974

Terms of use:
The institutional repository of the University of Udine (http://air.uniud.it) is provided by ARIC services. The aim is to enable open access to all the world.

Publisher copyright

(Article begins on next page)

04 May 2024

Multi-study Factor Analysis

Roberta De Vito[1,*], Ruggero Bellio[2], Lorenzo Trippa[3,4], and Giovanni Parmigiani[3,4]

[1]Department of Computer Science, Princeton University, Princeton, NJ, USA

[2]Department of Economics and Statistics, University of Udine, Udine, Italy

[3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA

[4]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA

*email: rvito@princeton.edu

Summary:

We introduce a novel class of factor analysis methodologies for the joint analysis of multiple studies. The goal is to separately identify and estimate 1) common factors shared across multiple studies, and 2) study-specific factors. We develop an Expectation Conditional-Maximization algorithm for parameter estimates and we provide a procedure for choosing the numbers of common and specific factors. We present simulations for evaluating the performance of the method and we illustrate it by applying it to gene expression data in ovarian cancer. In both, we clarify the benefits of a joint analysis compared to the standard factor analysis. We have provided a tool to accelerate the pace at which we can combine unsupervised analysis across multiple studies, and understand the cross-study reproducibility of signal in multivariate data. An R package (*MSFA*), is implemented and is available on GitHub.

Key words: Cross-study analysis; Dimension reduction; ECM algorithm; Gene Expression; Meta-analysis; Reproducibility.

1. Introduction

Analyses that integrate multiple sources, studies, and data-collection technologies are common in current statistical research. When considering multiple studies, a fundamental challenge is learning common features shared among studies while isolating the variation specific to each study. Two important statistical questions remain largely unanswered in this context: i) To what extent is the common signal shared across studies? ii) How can this shared signal be extracted? In this paper we develop a methodology to address these two questions via multi-study factor analysis.

Joint factor analysis of multiple studies is used in several areas of science. For example Scaramella et al. (2002) researched adolescent delinquent behavior in two independent samples, analyzing the same variables, to identify shared patterns. Andreasen et al. (2005) studied remission in schizophrenia, applying factor analysis (FA) to each individual samples. Their work showed that replicable results are found across all these FAs, leading to similar components. In nutritional epidemiology Edefonti et al. (2012) analyzed the dietary habits in relation to risk of head and neck cancer. They measured the same variables (nutrients) in five different populations, merged these into a single dataset and applied FA to the merged data to determine common dietary patterns and their relation with head and neck cancer. Wang et al. (2011) used FA to obtain a unified gene expression measurement from distinct types of measurements on the same samples. These examples, however, lack the ability to jointly derive in a single analysis (1) factors that capture common information, shared across studies, and (2) study-specific factors.

Our motivating examples arise in the analysis of gene expression data (Irizarry et al., 2003; Shi et al., 2006; Kerr, 2007). In gene expression analysis, as well as in much of high-throughput biology analyses on human populations, variation can arise from the intrinsic biological heterogeneity of the populations being studied, or from technological differences

in data acquisition. In turn both these types of variation can be shared across studies or not. As noted by Garrett-Mayer et al. (2008), the fact that the determinants of both natural and technological variation differ across studies implies that study-specific effects occur in most datasets. Both common and study-specific effects can be strong, and both need to be identified and studied. Our interest in this issue is a natural development of our previous work on unsupervised identification of integrative correlation (Parmigiani et al., 2004; Garrett-Mayer et al., 2008; Cope et al., 2014), and multi-study supervised analyses including cross-study differential expression (Scharpf et al., 2009), multi-study gene set analysis (Tyekucheva et al., 2011), comparative meta-analysis (Riester et al., 2014; Waldron et al., 2014), and cross study validation (Bernau et al., 2014).

In high-throughput biology, as well as in a number of other areas of application, the ability to separately estimate common and study-specific factors can contribute significantly to two important questions: the cross-study validation question of whether factors are found repeatedly across multiple studies; and the meta-analytic question of more efficiently estimating the factors that are indeed common. With regard to interpretation, the shared signal is more likely to capture genuine biological information, while the study-specific signal can point to either artifactual or biological sources of variation. Thus, modeling both shared and unshared factors may enable a more reliable identification of artifacts, facilitate more efficient experimental designs, and inform further technological advances.

In this article we propose a dimension-reduction approach that allows for joint analysis of multiple studies, achieving the goal of capturing common factors. Specifically, we define a generalized version of FA, able to handle multiple studies simultaneously. Our model, termed Multi-study Factor Analysis (MSFA), learns the common features shared among studies, and identifies the unique variation present in each study.

While unsupervised multi-study analysis is not an adequately studied field, our work

draws from existing foundations from related problems. In the social science literature, there is extensive methodology to identify factor structures shared among different groups, forming the body of *multigroup factor analysis* methods (see, among many others, Thurstone (1931); Jöreskog (1971); Meredith (1993)). These methods focus mainly on investigating measurement invariance among different groups, which typically results in testing whether the data support the hypothesis of a common loading matrix across groups. A notable special case is given by partial measurement invariance (see for example Byrne et al., 1989), which inspired our mathematical formulation. In our MSFA we have extended the scope to detection of both study-specific factors and factors that are identical across multiple studies. Our MSFA has also an exploratory goal, different from the confirmative approach under which measurement invariance is usually investigated in the social sciences.

The plan of the paper is as follows. Section 2 introduces the MSFA, and describes the Maximum Likelihood Estimation (MLE) of model parameters, implemented via an Expectation Conditional Maximization (ECM) algorithm. Section 3 presents simulation studies, providing numerical evidence on the performance of the proposed estimation methods. Next it investigates choosing the dimension of the latent factor, via model selection. Section 4 applies the methodology to study the Immune System pathway in ovarian cancer. Section 5 is the discussion.

2. Methods

2.1 *The multi-study factor analysis (MSFA) model*

We consider S studies, each with the same P variables. Generic study s has n_s subjects and, for each subject, a P-dimensional centered data vector $\mathbf{x}_{is}$ with $i = 1, \ldots, n_s$. To motivate, we begin with the case where a standard FA is carried out separately in each study. The observed variables in study s are decomposed into T_s factors. In particular, let $\mathbf{l}_{is},\ i = 1, \ldots, n_s$ be

the values of the *study-specific* factors in individual i of study s and $\mathbf{\Lambda}_s$, $s = 1, \ldots, S$ be the $P \times T_s$ corresponding factor loading matrices. FA assumes that each $\mathbf{x}_{is}$ is decomposed as

$$\mathbf{x}_{is} \quad = \quad \mathbf{\Lambda}_s \mathbf{l}_{is} + \mathbf{e}_{is} \quad i = 1, \ldots, n_s \,, \tag{1}$$

where $\mathbf{e}_{is}$ is a normal error term with covariance matrix $\mathbf{\Psi}_s = \mathrm{diag}(\psi_{1\mathrm{s}}, \ldots, \psi_{\mathrm{ps}})$ (e.g. Jöreskog, 1967, 1971). FA aims at explaining the dependence structure among observations by decomposing the $P \times P$ covariance matrix $\mathbf{\Sigma}_s$ as $\mathbf{\Sigma}_s = \mathbf{\Lambda}_s \mathbf{\Lambda}_s^\top + \mathbf{\Psi}_s$.

[Figure 1 about here.]

Figure 1.a summarizes the studies analyzed in this paper. Additional information can be found in the Supplementary Materials §A. Figure 1.b displays the loading vectors, and suggests there may be common factors across studies, as further illustrated in Figure 1.c where three of the loading vectors of the GSE9891 study are strongly correlated with four loading vectors in the GSE20565 study. Highly correlated pairs of loading vectors are more likely to represent common factors. On the other hand, some loading vectors of GSE9891 (e.g. $\boldsymbol{\lambda}_{41}$) exhibit low correlation with all loading vectors of GSE20565. These loadings are likely to result from feature unique to this study. The RV coefficient (Robert and Escoufier, 1976), assessing the multivariate correlation between the two loading matrices, is 0.76, indicating a high similarity between the two matrices. When we restrict the calculation to the first four factor loadings the RV coefficient increases further to 0.86.

Next we introduce our MSFA model, designed to analyze multiple studies jointly, replacing the heuristic interpretation above with a principled statistical approach. MSFA explicitly models common biological features shared among the studies, as well as unique variation present in each study. Specifically, the observed variables in study s are decomposed into K factors shared with all the other studies, and J_s additional factors reflecting its unique sources of variation, for a total of $T_s = K + J_s$ factors. Let $\mathbf{f}_{is}$ be the *common* factor vector in subject i of study s, and $\mathbf{\Phi}$ be the $P \times K$ common factors loading matrix. Moreover, let

$\mathbf{l}_{is}$ be the *study-specific* factor and $\mathbf{\Lambda}_s$ be the $P \times J_s$ specific factors loading matrix. MSFA assumes that the P-dimensional centered response vector $\mathbf{x}_{is}$ can be written as

$$\mathbf{x}_{is} = \mathbf{\Phi}\mathbf{f}_{is} + \mathbf{\Lambda}_s\mathbf{l}_{is} + \mathbf{e}_{is}\,, \qquad i = 1, \ldots, n_s \qquad s = 1, \ldots, S. \tag{2}$$

where the $P \times 1$ random error vector $\mathbf{e}_{is}$ has a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Psi}_s$, with $\mathbf{\Psi}_s = \mathrm{diag}(\psi^2_{s_1}, \ldots, \psi^2_{s_\mathrm{p}})$. We also assume that the marginal distribution of $\mathbf{l}_{is}$ is multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}_{J_s}$, and the marginal distribution of $\mathbf{f}_{is}$ is multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}_k$, where $\mathbf{I}$ denotes the identity matrix.

As a result of the model assumptions, the marginal distribution of $\mathbf{x}_{is}$ is multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_s = \mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Lambda}_s\mathbf{\Lambda}_s^\top + \mathbf{\Psi}_s$, with the three terms reflecting the variance of the common factors, the variance of the study-specific factors, and the variance of the error, respectively.

2.2 *Identifiability*

To specify an identifiable MSFA model we need to address two separate concerns. First we need to avoid orthogonal rotation indeterminacy, similarly to the classic FA. To clarify, if we define $\mathbf{\Phi}^* = \mathbf{\Phi}\mathbf{Q}$ and $\mathbf{\Lambda}_s^* = \mathbf{\Lambda}_s\mathbf{Q}_s, s = 1, \ldots S$, where $\mathbf{Q}$ and each $\mathbf{Q}_s$ are square orthogonal matrices with K and J_s rows respectively, we have

$$\mathbf{\Sigma}_s = \mathbf{\Phi}^*(\mathbf{\Phi}^*)^\top + \mathbf{\Lambda}_s^*(\mathbf{\Lambda}_s^*)^\top + \mathbf{\Psi}_s = \mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Lambda}_s\mathbf{\Lambda}_s^\top + \mathbf{\Psi}_s.$$

Thus our decomposition of $\mathbf{\Sigma}_s$ is not uniquely identified. FA (1) identifies the parameters by imposing constraints on the factor loadings matrix. One possibility often used in practice is to take $\mathbf{\Lambda}_s$ in (1) to be a lower triangular (LT) matrix (Geweke and Zhou, 1996, pp. 565-566), (Lopes and West, 2004; Carvalho et al., 2008). Here we extend this approach to MSFA, by specifying $\mathbf{\Phi}$ and all the $\mathbf{\Lambda}_s$'s to be LT matrices. We refer to this condition as block LT.

Similarly to FA, this resolves the orthogonal rotation indeterminacy. However, in MSFA a

second concern arises, since the S equations

$$\mathbf{\Sigma}_s - \mathbf{\Psi}_s = \mathbf{\Phi}\,\mathbf{\Phi}^\top + \mathbf{\Lambda}_s\,\mathbf{\Lambda}_s^\top\,, \qquad s = 1, \ldots, S\,, \tag{3}$$

for fixed values of $\mathbf{\Sigma}_s$ and $\mathbf{\Psi}_s$, still involve the $S+1$ matrices $\mathbf{\Phi}, \mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_S$. Thus block LT does not guarantee the uniqueness of the solution.

To address this we require the further condition that the concatenated matrix $\mathbf{\Omega} = [\mathbf{\Phi}, \mathbf{\Lambda}_1, \ldots \mathbf{\Lambda}_s]$ has a full column rank $r(\mathbf{\Omega}) = K + \sum_{s=1}^S J_s$, with $K + \sum_{s=1}^S J_s \leqslant P$. Denote by $\mathrm{span}(A)$ the span of the column vectors of A. Then, since $\mathrm{span}(A) = \mathrm{span}(A\,A^\top)$

$$\mathrm{span}(\mathbf{\Sigma}_s - \mathbf{\Psi}_s) = \mathrm{span}(\mathbf{\Phi}) \oplus \mathrm{span}(\mathbf{\Lambda}_s)\,, \qquad s = 1, \ldots, S\,,$$

where $\oplus$ is the direct sum between the two linear spaces. Then it follows that $\mathrm{span}(\mathbf{\Phi})$ is uniquely determined, since it is the intersection of the S vector spaces $\mathrm{span}(\mathbf{\Sigma}_s - \mathbf{\Psi}_s)$, and similarly $\mathrm{span}(\mathbf{\Lambda}_s)$ are uniquely obtained as orthogonal complements of $\mathrm{span}(\mathbf{\Phi})$ in $\mathrm{span}(\mathbf{\Sigma}_s - \mathbf{\Psi}_s)$. Then the only indeterminacy left in (3) is due to the action of the orthogonal matrices $\mathbf{Q}$ and $\mathbf{Q}_s$, $s = 1, \ldots, S$. This latter point is solved by the block LT constraint.

This identification strategy requires that the number of latent factors be no larger than the number of variables, that is: $K + \sum_{s=1}^S J_s \leqslant P$. An important issue for MLE in the MSFA model concerns some constraints that ought to be considered. For the sth study, the number of elements in the sample covariance matrix must be greater than the number of free parameters in $\mathbf{\Sigma}_s$. This constraint implies that

$$PK - K(K-1)/2 + \sum_{s=1}^{S} \{P\,J_s - J_s\,(J_s - 1)/2\} + SP \;\leqslant\; SP(P+1)/2.$$

For $S > 2$, this condition is less restrictive than $K + \sum_{s=1}^S J_s \leqslant P$ as long as the total number of latent factors is larger than the number of studies. A further issue, largely inconsequential for MLE, is that we can simultaneously change the sign to all the elements of columns of the loading matrices and to all the corresponding latent factors, without changing the model. This could be fixed by constraining the sign of a subset of loadings.

An alternative and more restrictive constraint for model identification is to impose the LT condition directly on $\mathbf{\Omega}$, as discussed in the Supplementary Materials.

2.3 *Parameter estimation*

The parameters to be estimated in the MSFA are $\boldsymbol{\theta} = (\mathbf{\Phi}, \mathbf{\Lambda}_s, \mathbf{\Psi}_s)$. For notational simplicity in both (1) and (2) we assume that the observed variables in each study have been centered. In the following, the MLE will be obtained by the Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993), a class of generalized EM algorithms (Dempster et al., 1977). The details of the ECM algorithm for the MSFA model are reported in the Supplementary Materials.

2.3.1 *Estimable loading matrices.* When both the LT constraints for various loading matrices and the full rank condition for the $\mathbf{\Omega}$ matrix mentioned in §2.2 are applied within the ECM algorithm, the model is identified and the MLE can be obtained. Furthermore, under the assumption of sample information increasing across all the studies, standard asymptotic theory implies the consistency of the MLE. The asymptotic limit of the MLE of the loading matrices, i.e. the estimable parameters, depends on the constraints imposed, and it is useful to provide further details.

Let us assume that the data come from the MSFA model (2) with unconstrained loading matrices $\mathbf{\Phi}$ and $\mathbf{\Lambda}_s$. When the block LT is enforced, for study s there exist only two rotations $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{Q}}_s$, without considering the sign indeterminacy, such that $\mathbf{\Phi}\,\tilde{\mathbf{Q}}$ and $\mathbf{\Lambda}_s\,\tilde{\mathbf{Q}}_s$ are LT. The MLEs then converges to the LT products $\mathbf{\Phi}\,\tilde{\mathbf{Q}}$ and $\mathbf{\Lambda}_s\,\tilde{\mathbf{Q}}_s$. With simple algebra we derive that $\tilde{\mathbf{Q}}$ equals the transpose matrix of the Q matrix obtained from the QR-decomposition of $\mathbf{\Phi}^\top$, and that $\tilde{\mathbf{Q}}_s$ equals the transpose of the QR-decomposition of $\mathbf{\Lambda}_s^\top$. These results are useful in simulation studies, since they provide a benchmark for measuring the performance of the MLE for finite samples.

2.3.2 *Dimension Selection.* Selecting the dimension of the model can be challenging. We found the following two-step procedure to be effective in our applications. First, we determine the total latent dimension $T_s = K + J_s$ for each of the S studies using standard techniques for FA, such as Horn's parallel analysis (Horn, 1965), Cattell's scree test (Cattell, 1966) or the use of indexes, such as the RMSEA (Steiger and Lind, 1980). Next, we apply model selection techniques to the overall MSFA model to select the number K of latent factors sharing a common loading matrix $\mathbf{\Phi}$, as described in §3.2. Lastly, we derive the dimensions J_s residually as $T_s - K$, with the restrictions that $T_s - K \geqslant 0$ for all $s = 1, \ldots, S$.

3. Simulation studies

We performed simulation experiments to evaluate the effectiveness of the ECM in estimating the MSFA model parameters, as well as our strategy for selecting the latent dimensionality. Our simulation studies are designed to closely mimic the data of Figure 1. We consider $S = 4$ studies, with latent factor dimension $T_s = \{6, 7, 10, 9\}$, and generate $\mathbf{x}_{is}$ from P-dimensional normal distributions, with sample size equal to $n_s = \{285, 140, 195, 578\}$. We assume that samples from the four studies are drawn from different population, each with zero mean and covariance matrix $\mathbf{\Sigma}_s = \mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Lambda}_s\mathbf{\Lambda}_s^\top + \mathbf{\Psi}_s$. Factor loadings matrices are not constrained. We investigate three simulation scenarios, with $K = 0$, $K = 1$, and $K = 3$. To produce more realistic results, in each scenario we generate data from parameter values close to those estimated with the data of Figure 1.

3.1 *Parameter estimation via the ECM algorithm*

We first analyze the performance of the ECM algorithm for a given selection of K and J_s, $s = 1, \ldots, S$. Irrespective of the optimization method adopted, the choice of the starting point is crucial for achieving a good performance. Details of the strategy proposed here are reported in the Supplementary Materials.

Figure 2 shows that MSFA is able to recover the estimable parameters of the MLE, as described in Section 2.3.1, for the common factor loadings. Moreover, MSFA performs better than FA in estimating the true shared factor loadings. FA is computed after stacking the studies into a single dataset. Different analysis for checking if the MSFA recovers the true factors, and results for the other scenarios are reported in Supplementary Materials.

[Figure 2 about here.]

3.2 *Selection of the latent factor dimensions*

Next we consider selecting the dimension of the latent space, by simulation experiments as above. For each data set, we first choose T_s by standard FA techniques, and then choose K. We compare three model selection techniques for selecting K: BIC, for which there is an extensive literature (Burnham and Anderson, 2002; Preacher and Merkle, 2012), AIC, whose properties in FA are actively investigated (Chen and Chen, 2008; Hirose and Yamamoto, 2014), and the likelihood ratio test (LRT) for choosing between nested models with different K's. Table 1 shows the results for 100 different data sets generated independently.

[Table 1 about here.]

BIC emerges as the best criterion, always leading to the selection of the true model. Generally, unlike AIC, BIC selects the true model with probability 1 as the sample size increases. However, the AIC may be better than the BIC in term of mean squared error (MSE) of prediction. Therefore, the strategy employed in our applied example will use both AIC and BIC.

4. Expression of Genes in Immune System Pathways in Ovarian Cancer

To illustrate MSFA in an important biological example, we analyzed the four studies described in Figure 1.a, with $n_s (s = 1, \ldots, 4)$. We focus on transcription of genes involved in

immune system activity by considering genes included in the pathways "Adaptive Immune System" (AI), "Innate Immune System" (II) and "Cytokine Signaling in Immune System" (CSI) from `reactome.org`. As defined, these pathways do not have overlapping genes. In addition, we restricted attention to genes which are common across all studies.

We conducted preliminary analyses to asses the total latent factor dimensions, the number of common factors across studies and the number of specific factors for each study. AIC estimates the number of common factors as one, while BIC yields five.

We then compare the cross-validation prediction errors computed by the MSFA to those computed by FA, the latter applied in two different ways: merging the 4 studies into a single data set; and separately computing FA in each study. For each cross-validation iteration, we train on a random 80% of the data, and evaluate the prediction error on the remaining 20%. Predictions are obtained as

$$\text{MSFA:} \quad \hat{\mathbf{x}}_{is} = \hat{\mathbf{\Phi}}\hat{\mathbf{f}}_i + \hat{\mathbf{\Lambda}}_s^{MSFA}\hat{\mathbf{l}}_{is} \qquad\qquad \text{FA:} \quad \hat{\mathbf{x}}_{is} = \hat{\mathbf{\Lambda}}_s^{FA}\hat{\mathbf{l}}_{is}.$$

where $\hat{\mathbf{\Lambda}}_s^{MSFA}$ are the specific factor loadings estimated with MSFA and $\hat{\mathbf{\Lambda}}_s^{FA}$ are the factor loadings estimated with FA. For BIC, the MSE is 0.7% smaller for MSFA than for FA after merging the data and is 5.75% smaller for MSFA than for FA applied separately to each study. For AIC the MSE is 5.20% smaller for MSFA than for FA after merging the data and is 11.50% smaller for MSFA than for FA applied separately to each study. This analysis illustrates how MSFA borrows strength across studies in the estimation of the factor loadings, in such a way that the predictive ability in independent observation is not only preserved but even improved. The model selected by AIC has smaller MSE than the one selected by BIC. Therefore we continue our discussion focusing on the model chosen by AIC.

[Figure 3 about here.]

Next, we focus on the analysis of the factors themselves. The heatmap in Figure 3 depicts the estimates of the factor loadings, both common (in the black rectangle) and specific

(this figure appears in color in the electronic version of this article, and color refers to that version). To help interpreting the biological meaning of the common factor, we apply Gene Set Enrichment Analysis (GSEA) for determining whether one of the three gene sets is significantly enriched among loadings that are high in absolute value (Subramanian et al., 2005). We used the package `RTopper` in `R` in `Bioconductor`, following the method illustrated in Tyekucheva et al. (2011). The resulting analysis shows that the common factor is significantly enriched for genes in the Innate Immune System pathway, suggesting that genuine biological signal may have been identified. Further, Figure 3.b shows that three of the specific factors of the GSE9891 study are strongly correlated with three corresponding factors in the GSE20565 study.

To further investigate this observation we analyze studies GSE9891 and GSE20565 separately from the other two using MSFA (figure reported in Supplementary materials). The AIC chooses a model with $K = 3$. Studies GSE9891 and GSE20565 use the same microarray platform, Affy U133 Plus2.0, unlike the other two. This prompts the conjecture that the four stronger correlations observed may be related to technological rather than biological variation. Naturally it is also possible that there may be specific technical features of this platform that enable it to identify additional biological factors, although this is less likely in view of the fact that our analysis is restricted to a common set of genes.

We next performed GSEA on the estimated factor loadings for the two-study analysis. The results show that the first common factor is still related to the II system pathway, as was the case for the single common factor shared between the four studies in the earlier analysis. The two remaining common factors are not related to any of the remaining pathways, further corroborating the hypothesis that they may represent the results of spurious variation unique to the specific platform used.

We also checked the impact of the choice of a gene order, because of the dependence induced

by the block LT structure assumed for $\mathbf{\Phi}$ and $\mathbf{\Lambda}_s$, to address identifiability. In particular, we repeated the same analysis after permuting the variables. Despite minor discrepancies, the final conclusion is still that the single common factor is significantly enriched only with genes in the innate immune system pathway.

Overall, this analysis illustrates important features of this method, including its ability to capture biological signal common to multiple studies and technological platforms, and at the same time to isolate the source of variation coming, for example, from the different platform by which gene expression is measured. It also illustrates in real data how pooling important factors across studies leads to increased stability, resulting in improved predictive ability.

5. Discussion

In this article we introduced and studied a novel class of factor analysis methodologies for the joint analysis of multiple studies. We hope that our work will provide a valuable tool to accelerate the pace at which we can combine unsupervised analysis across multiple studies, and understand the cross-study reproducibility of signal in multivariate data.

The main concept is to separately identify and estimate 1) common factors shared across multiple studies, and 2) study-specific factors. This is intended to help address one of the most critical steps in cross-study analysis, namely to identify factors that are reproducible across studies and to remove idiosyncratic variation that lacks cross-study reproducibility. The method is simple and is based on a generalized version of FA able to handle multiple studies simultaneously and to capture the two types of information.

Several methods have been proposed to analyze diverse data sets and to capture the correlation between different studies. CPCA was introduced by Flury (1984) to investigate the hypothesis that the covariance matrices for different populations are simultaneously diagonalizable. This method estimates a common principal axes across the different population and the deviation of the data from the model of common principal axes. Co-inertia analysis

(CIA) emerged in ecology to explore the common structure of two distinct sets of variables (such as species' abundances of flora and fauna) measured at the same sites (Dolédec and Chessel, 1994; Dray et al., 2003). It proceeds by separately performing dimension reduction on each set of variables, to derive factor scores for the sites. In a second, independent, stage the correlation between these factors is investigated. The Multiple Co-inertia analysis (MCIA) (Dray et al., 2003) is a generalization of CIA to consider more than two data sets. MCIA finds a hyperspace, where variables showing similar trends are projected close to each other (Meng et al., 2014).

A related method is the Multiple Factor analysis (MFA) (Abdi et al., 2013), an extension of component analysis (PCA) which consists of three steps. The first is a PCA for each study. In the second step each data set is normalized by dividing by its first singular value. In the third step, a single data set is created by stacking the normalized data from different studies by row, and a final PCA is done.

Two differences can be emphasized between these approaches and MSFA. First they are focused on analyzing only the common structure after having excluded the noise. Instead our method estimates both common and study-specific components. Second they operate stage-wise, decomposing each matrix separately, while our study analyzed the data jointly. This is critical in a meta-analytic context because the presence of a recognizable factor in one study can assist with the identification of the same factor in other studies even when it is more difficult to recognize it.

The MSFA needs to be constrained to be identifiable. The constraints used here is the block LT matrix. Although this condition is often used in classical FA, it induces an order dependence among the variables (Frühwirth-Schnatter and Lopes, 2010). As noted in Carvalho et al. (2008), the choice of the first variables in the order is an important modeling decision, to be made with some care. In our application, it is somewhat reassuring that the

checks made on the impact of the variable order on the final conclusion leads to the same conclusions. This is likely to hold broadly in gene expression where much biology operates on modules of several correlated genes rather than single genes. However, general conclusions cannot be drawn.

We also present an alternative approach for addressing identifiability, based on applying the LT assumption directly to $\mathbf{\Omega}$, as described in the Supplementary Materials. This approach does not require that $K + \sum_s J_s \leqslant P$, but it is based on stronger assumptions as it implies a larger number of zeros in $\mathbf{\Lambda}_s$ and eliminates some variables altogether from the study specific factors. Unless the condition $K + \sum_s J_s \leqslant P$ fails, we suggest using the constraint proposed in the main paper. This has the cost of imposing a bound of the total latent dimensions, but it does not prevent the productive application of the methodology in practical settings. Indeed, in the applications we considered for the present work as well as the work reported in De Vito et al. (2018), the condition $K + \sum_s J_s \leqslant P$ was always met for sensible model specifications. Other constraints or rotation methods, such as the varimax criterion (Kaiser, 1958), could be considered, though their extension to the MSFA setting would require further investigation.

In settings characterized by high-dimensional data where there are more variables than observations, the estimation of the MSFA model requires a different approach. To this end, our ongoing research is focusing on the extension of the Bayesian infinite factor model proposed by Bhattacharya and Dunson (2011) to the MSFA setting.

The MSFA model can be applied to many settings when the aim is to isolate commonalities and differences across different groups, population or studies. In our gene expression application the goal is estimating the biological signal shared among studies, while removing study-specific features less likely to be reproducible across populations, and potentially arising from technological issues. Elsewhere the goal may be to capture study-specific features of interest

after removing common factors. Finally, other applications may focus on both common and specific factors. Examples where study-specific patterns are germane arise in nutritional epidemiology (Carrera et al., 2007; Ryman et al., 2015) where population-specific diets may have a lower or higher impact on specific diseases, such as obesity or cancer. MSFA may have broad applicability in a wide variety of genomic platforms (e.g. microarrays, RNA-seq, SNPs, epigenomics), as well as datasets in other fields of biomedical research, such as those generated by exposome studies. Beyond, the concept is straightforward, universal and of general interest across all applications of multivariate analysis.

Supplementary Materials

Web appendices and Figures referenced in Section 2, 3, and 4 are available with this paper at the *Biometrics* website on Wiley Online Library. An R package (*MSFA*), is implemented and is available on GitHub at `https://github.com/rdevito/MSFA`.

Acknowledgements

We thank the editors and referees for insightful comments and suggestions. One referee specifically suggested the direction that ultimately led to our current identifiability constraint. We are also grateful to Francesco Lin, Luigi Pace, Cinzia Viroli for their helpful comments.

References

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,** 149–179.

Andreasen, N. C. et al. (2005). Remission in schizophrenia: proposed criteria and rationale for consensus. *American Journal of Psychiatry* **162,** 441–449.

Bernau, C. et al. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* **30,** 105–112.

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

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: a Practical Information-theoretic Approach.* Springer, New York, second edition.

Byrne, B. M., Shavelson, R. J., and Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin* **105,** 456–466.

Carrera, P. M., Gao, X., and Tucker, K. L. (2007). A study of dietary patterns in the mexican-american population and their association with obesity. *Journal of the American Dietetic Association* **107,** 1735–1742.

Carvalho, C. M. et al. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association* **103,** 1438–1456.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research* **1,** 245–276.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95,** 759–771.

Cope, L., Naiman, D. Q., and Parmigiani, G. (2014). Integrative correlation: Properties and

relation to canonical correlations. *Journal of Multivariate Analysis* **123,** 270–280.

De Vito, R. et al. (2018). Shared and study-specific dietary patterns. *Epidemiology*, in press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

Dolédec, S. and Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species–environment relationships. *Freshwater Biology* **31,** 277–294.

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

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,** 1869–1880.

Flury, B. N. (1984). Common principal components in k groups. *Journal of the American Statistical Association* **79,** 892–898.

Frühwirth-Schnatter, S. and Lopes, H. F. (2010). Parsimonious Bayesian factor analysis when the number of factors is unknown. *Unpublished Working Paper, Booth Business* .

Garrett-Mayer, E. et al. (2008). Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics* **9,** 333–354.

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

Hirose, K. and Yamamoto, M. (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics & Data Analysis* **79,** 120–132.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* **30,** 179–185.

Irizarry, R. A. et al. (2003). Exploration, normalization, and summaries of high density

oligonucleotide array probe level data. *Biostatistics* **4,** 249–264.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32,** 443–482.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika* **36,** 409–426.

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 Bioinformatics* **8,** 371–377.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14,** 41–68.

Meng, C. et al. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* **15,** 162–175.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80,** 267–278.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* **58,** 525–543.

Parmigiani, G. et al. (2004). A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clinical Cancer Research* **10,** 2922–2927.

Preacher, K. J. and Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods* **17,** 1–14.

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,** 1–12.

Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Applied Statistics* **25,** 257–265.

Ryman, T. K. et al. (2015). Characterising the reproducibility and reliability of dietary patterns among Yup'ik Alaska native people. *British Journal of Nutrition* **113,** 634–643.

Scaramella, L. V. et al. (2002). Evaluation of a social contextual model of delinquency: a cross-study replication. *Child Development* **73,** 175–195.

Scharpf, R. et al. (2009). A Bayesian model for cross-study differential gene expression. *Journal of the American Statistical Association* **104,** 1295–1310.

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

Steiger, J. H. and Lind, J. M. (1980). Statistically based tests for the number of common factors. *Paper presented at Psychometric Society Meeting, Iowa City, May* .

Subramanian, A. et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 15545–15550.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review* **38,** 406–427.

Tyekucheva, S. et al. (2011). Integrating diverse genomic data using gene sets. *Genome Biology* **12,** 105–129.

Waldron, L. et al. (2014). Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *Journal of the National Cancer Institute* **106,** 49–61.

Wang, X. V. et al. (2011). Unifying gene expression measures from multiple platforms using factor analysis. *PloS One* **6,** 1932–1943.

a.

Study	Samples	Platform	Late Stage (%)	Reference	T_s
GSE9891	285	Affy U133Plus 2.0	85	Tothill et al. (2008)	6
GSE20565	140	Affy U133Plus 2.0	48	Meyniel et al. (2010)	7
GSE26712	195	Affy U133a	96	Bonome et al. (2008)	10
TCGA	578	Affy HT U133a	90	Cancer Genome Atlas Research Network (2011)	9

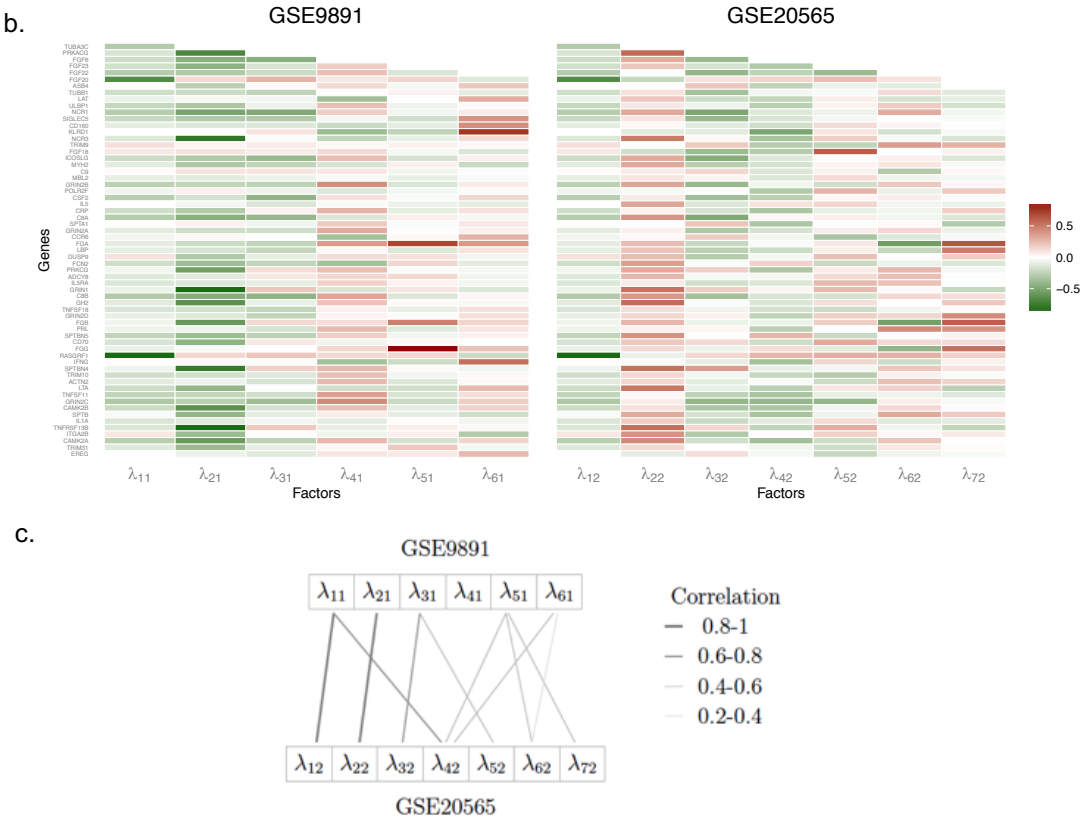

Figure 1: **Part a**: Summary of the four data sets analyzed, including: GEO labels if available, sample size, microarray platform used, proportion of patients diagnosed with late state (Stage III or Stage IV) ovarian cancer, references, and the number of latent dimensions T_s estimated by the `nFactors` package in `R`. **Part b**: Heatmap of the estimated factor loadings, $\boldsymbol{\Lambda}_1 = \{\boldsymbol{\lambda}_{11}, \boldsymbol{\lambda}_{21}, \boldsymbol{\lambda}_{31}, \boldsymbol{\lambda}_{41}, \boldsymbol{\lambda}_{51}, \boldsymbol{\lambda}_{61}\}$ and $\boldsymbol{\Lambda}_2 = \{\boldsymbol{\lambda}_{12}, \boldsymbol{\lambda}_{22}, \boldsymbol{\lambda}_{32}, \boldsymbol{\lambda}_{42}, \boldsymbol{\lambda}_{52}, \boldsymbol{\lambda}_{62}, \boldsymbol{\lambda}_{72}\}$, obtained by performing separate factor analyses as in equation (1) in studies GSE9891 and GSE20565. Each column $\boldsymbol{\lambda}_{is}$ is thus the ith loading vector of the sth study. We estimated parameters using the identifying constraint that the loading matrix is lower triangular (LT). This figure appears in color in the electronic version of this article. **Part c**: Bipartite graph representing the absolute value of the correlations between pairs of study-specific factor loadings. Correlations smaller than .2 are not shown. Darker lines denote larger correlations in absolute value. In computing these correlations, the same variables are considered in each study, and their order is preserved.

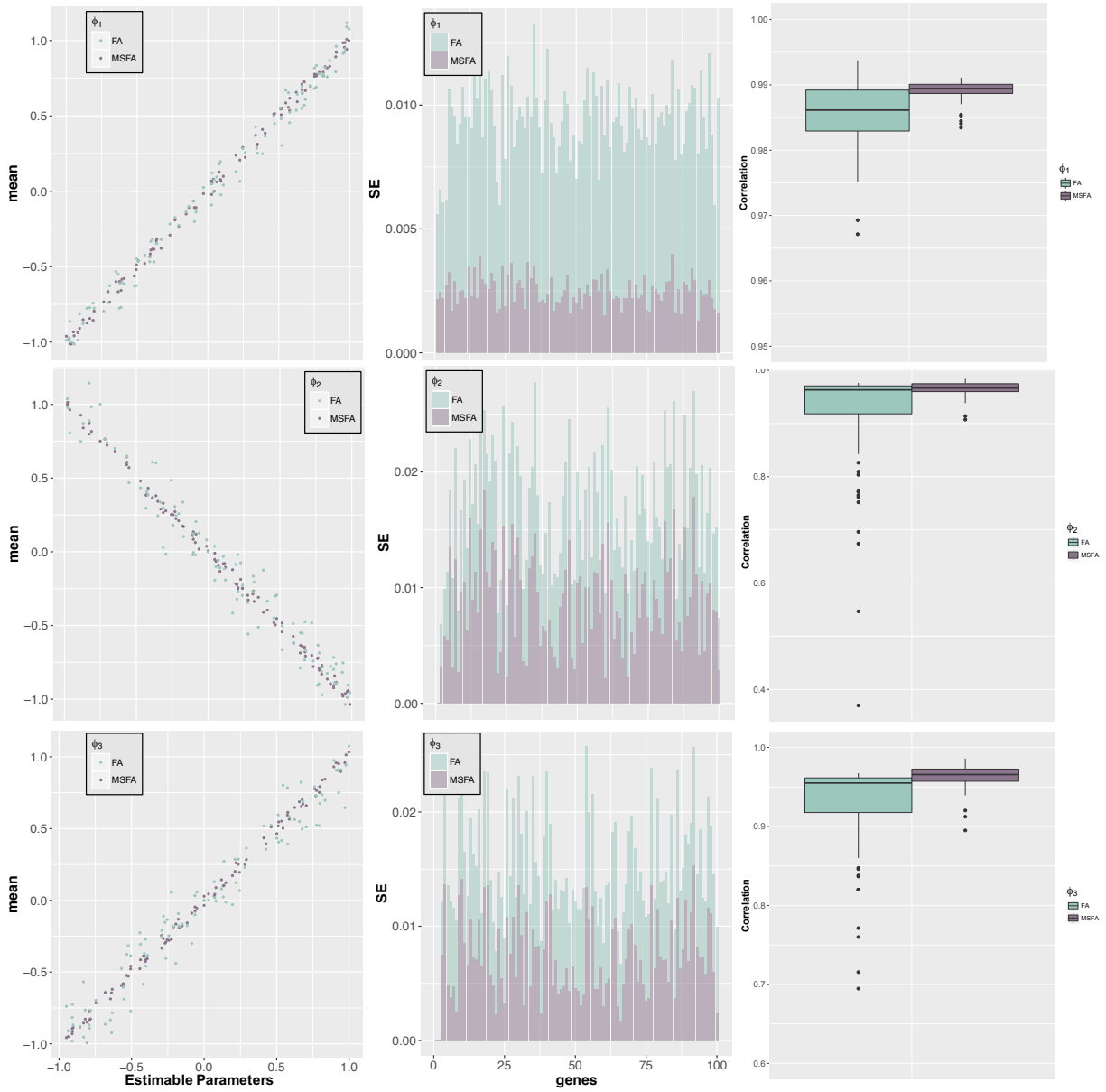

Figure 2: Comparison of common factor loadings estimated by MSFA (in purple) and FA after stacking the datasets into one (in green). We display the results of 100 simulations from Scenario 3, where $K = 3$. Each row corresponds to a common factor loading. The left column shows the mean of the estimated factor loadings ("mean"), versus the true common factor loadings ("Estimable Parameters") for each simulation. The center column shows the standard errors of the distances between true and estimated parameters, by gene. Consistently, standard errors from MSFA are smaller than those from FA for all three common factors. The right column shows boxplot of the correlations between estimated factor loadings and true common factor loadings across simulation. This figure appears in color in the electronic version of this article, and color refers to that version.

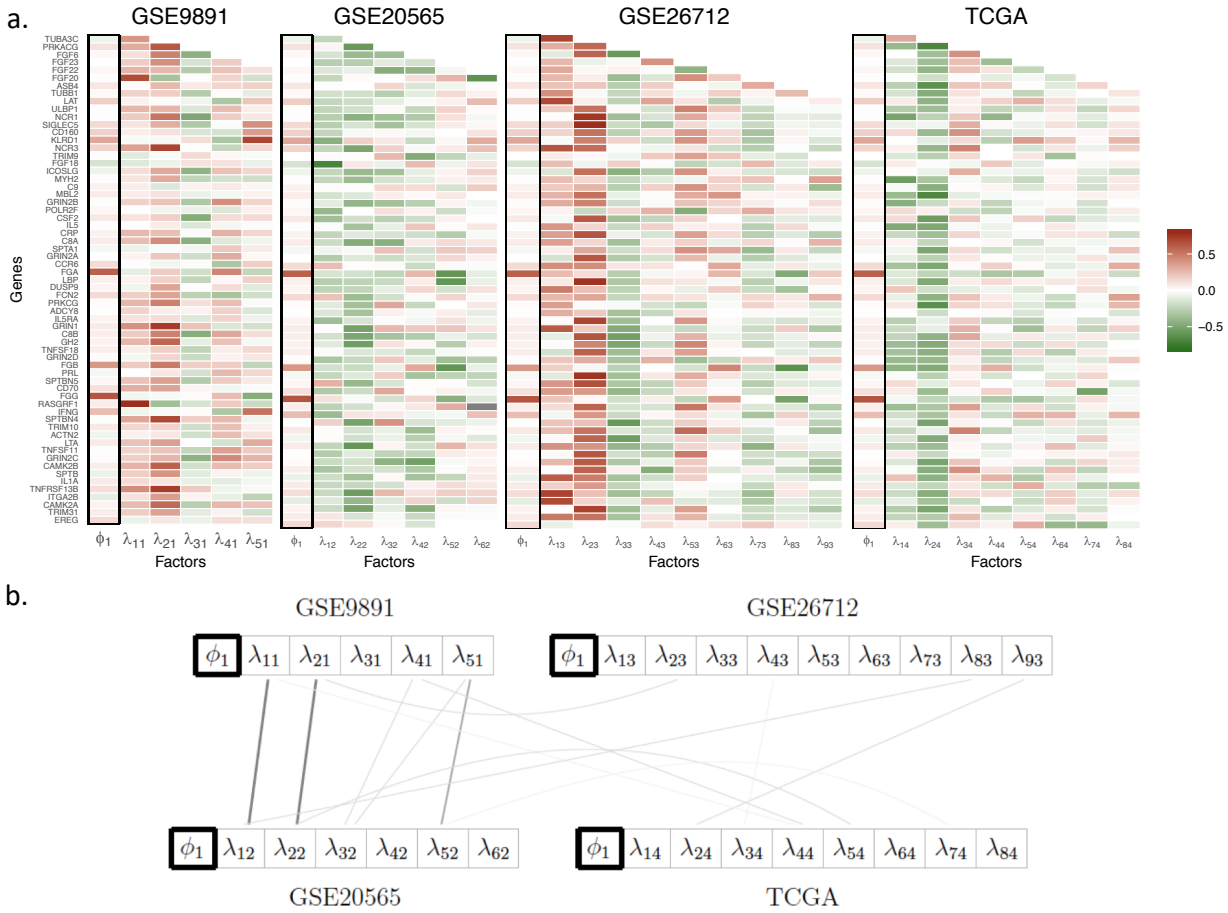

Figure 3: **Part a.** Heatmap of the estimated factor loadings, both common (black rectangle) and specific, obtained with MSFA in the data sets of Figure 1.a. This figure appears in color in the electronic version of this article, and color refers to that version. **Part b.** Graphical representation of the cross-study pairwise correlation between study-specific factor loadings. Darker grey lines correspond to higher correlations. Correlations smaller than .25 are not shown. Absolute correlations range from 0.66 to 0.81.

Table 1: Comparison of methods for choosing K. We report on 100 independent data sets generated from either $K = 0$, $K = 1$ and $K = 3$. We proceed as in 2.3.2 considering AIC, BIC and LRT in turn. Colums correspond to estimated values. If $K = 0$ or $K = 1$ all three methods choose the true K in all or almost all cases. If $K = 3$, BIC and LRT choose the true K more often than AIC.

	Method	$\boldsymbol{K = 0}$	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
True $K = 0$	AIC	100	0	0	0	0	0
	BIC	100	0	0	0	0	0
	LRT	100	0	0	0	0	0
	Method	$K = 0$	$\boldsymbol{K = 1}$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
True $K = 1$	AIC	0	100	0	0	0	0
	BIC	0	100	0	0	0	0
	LRT	2	98	0	0	0	0
	Method	$K = 0$	$K = 1$	$K = 2$	$\boldsymbol{K = 3}$	$K = 4$	$K = 5$
True $K = 3$	AIC	0	34	0	76	0	0
	BIC	0	0	0	100	0	0
	LRT	0	0	0	91	9	0