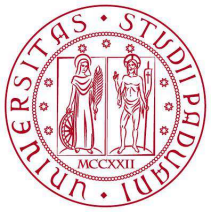

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXVIII

Multi-study factor models for high-dimensional biological data

Direttore della Scuola: Prof.ssa Monica Chiogna

Supervisore: Prof. Ruggero Bellio

Co-supervisori: Prof. Giovanni Parmigiani e Prof. Lorenzo Trippa

Dottoranda: Roberta de Vito

31/01/2016

Ad Andrea

Acknowledgements

In primo luogo vorrei ringraziare Giovanni Parmigiani perché senza di lui questi progetti non avrebbero preso forma, perché senza i suoi rilevanti insegnamenti professionali e soprattutto umani questo lavoro non esisterebbe. Allo stesso modo, desidero ringraziare Ruggero Bellio per il forte supporto, il contributo, la disponibilità e la pazienza nel seguirmi in questa tesi di dottorato. Ringrazio Lorenzo per gli stimolanti suggerimenti di lavoro. Un ringraziamento ad Emanuele che ha reso il percorso intrapreso a Boston piacevole e che mi ha dato sempre la sua assistenza nei momenti più difficili. Ringrazio tutti i dottorandi: specialmente Daniele, Giovanni, Giulio e Ronaldo per lo scambio di idee, le pizze in dipartimento alle 10 di sera, i caffè, i momenti di pausa e gli "esterrefanti". Un ringraziamento particolare va alla mia famiglia che mi ha supportato e sopportato nei momenti critici e alle amiche di sempre, Francy, Spu e Polly.

Abstract

High-throughput assays are transforming the study of biology, and are generating a rich, complex and diverse collection of high-dimensional data sets. Building systematic knowledge from this data is a cumulative process, which requires analyses that integrate multiple sources, studies, and technologies. The increased availability of ensembles of studies on related clinical populations, assaying technologies, and genomic features poses two categories of very important multi-study statistical components: 1) common factors shared across multiple studies; 2) study-specific factors. To capture these two different quantities, in this thesis we propose a novel class of factor analysis models, both under a frequentist and Bayesian approach.

In the frequentist approach an ECM algorithm is provided to obtain the maximum likelihood estimates. Moreover, we propose a Bayesian approach to apply the method to settings with more variables than subjects. In modeling dependencies among many variables, a sparse structure underlying the associations among genes is assumed.

Both methods allow to perform joint analysis of multiple high-throughput studies. The results are helpful for combining multiple studies, identifying reproducible biology across studies and interesting study-specific components, and removing idiosyncratic variation that lacks cross-study reproducibility.

Abstract

Le analisi scientifiche su un alto numero di campioni (*high-throughput assays*) stanno trasformando gli studi biologici. In particolare gli *high-throughput assays* generano una ricca, complessa e varia collezione di dati a più dimensioni.

Estrarre informazioni significative in maniera sistematica da questo tipo di dati richiede un processo progressivo che si basa sull'analisi simultanea di risorse, studi e tecnologie differenti.

La crescente disponibilità di numerosi studi clinici su rilevanti gruppi, popolazioni e diversi studi genetici genera due categorie: la prima, una categoria relativa ai fattori condivisi da tutti gli studi ed una seconda, relativa a fattori specifici di ogni studio.

Per catturare queste due differenti categorie abbiamo proposto, nell'ambito di tale tesi, una nuova classe di modellizzazione di analisi fattoriale che abbiamo sviluppato in un approccio sia frequentista che Bayesiano.

Nell'approccio frequentista, è stato proposto un algoritmo ECM per la stima di massima verosimiglianza dei parametri. Inoltre, in questa tesi, si è proposto un approccio Bayesiano per adattare questo modello ad un contesto di più variabili che soggetti, $p > n$. Nel modellizzare la dipendenza tra variabili, si è assunta una struttura sparsa per sottolineare le associazioni tra i geni.

Entrambi i metodi hanno consentito di modellizzare i diversi studi. Inoltre, i risultati hanno permesso di poter identificare un segnale biologico riproducibile e comune in tutti gli studi, nonché ad eliminare quella parte di varianza che oscura questo segnale.

Contents

Chapter 1

Introduction

1.1 Overview

This thesis deals with the issue of parameter estimation when a generalization of factor model for high-dimensional biological data is entailed.

High-throughput assays are transforming the study of biology, and are generating a rich, complex and diverse collection of high-dimensional data sets. Building systematic knowledge from this data is a cumulative process, which requires analyses that integrate multiple sources, studies, and technologies.

As a result of multiple studies, most components from high-throughput biology experiments show variation arising both from biological and artifactual sources. Indeed, biological differences include natural variations in different samples. Artifactual differences are related to the different measurements or platforms in which gene expression data are collected (Irizarry et al., 2003; Kerr, 2007; Shi et al., 2006). As noted in Garrett-Mayer et al. (2007), the fact that the determinants of both technological and biological variation differ across studies and laboratories implies that study-specific and laboratory-specific effects occur in most biological data sets. Study-specific effects can be so large to cover the biological signal under investigation for many genes (Aach et al., 2000).

While some biological features reappear across studies, genuine biological signal is more likely than spurious signal to be reproducibly present in multiple studies.

The increased availability of ensembles of studies on related clinical populations, assaying technologies, and genomic features poses two categories of very important multi-study statistical questions: i) To what extent is biological signal reproducibly shared across multiple studies? ii) How can this common signal be extracted? Furthermore, these questions need to

be answered by considering the challenges of learning common biological features shared among the studies and isolating the variation specific to each study.

1.2 Main contributions of the thesis

This work builds on extensive experience with multi-study analysis, including work on Bayesian multi-study analysis (Dominici et al., 1997, 1999; Muller et al., 1999), integrative correlation (Cope et al., 2014; Garrett-Mayer et al., 2007; Parmigiani et al., 2004), cross-study differential expression (Scharpf et al., 2009a), multi-study gene set analysis (Tyekucheva et al., 2011) and comparative meta-analysis (Riester et al., 2014; Waldron et al., 2014a).

This thesis aims at answering to both questions by extending previous literature approaches in two directions: frequentist analysis and Bayesian analysis. The focus is in particular on gene expression measurements.

The structure of the thesis is as follows. Chapter 2 describes the type of data, the main model frameworks introduced in literature to describe the data considered and the crucial problem of distinguish the reproducible biological signal from the study-specific component. Particular attention is on dimension reduction techniques, and factor analysis applied in this field. Factor Analysis (FA) have been used in several gene expression studies. Among others, Wang et al. (2011) used FA in gene expression, and Blum et al. (2010) adopted FA for multiple testing developed by Friguet et al. (2009). Finally, Runcie and Mukherjee (2013) used a sparse factor model by Bhattacharya and Dunson (2011).

In research applications, joint analyses of multiple genomic data sets have begun more than a decade ago, they are now increasingly common and can be highly successful (Ciriello et al. (2013); Gao et al. (2014); Hayes et al. (2006); Huttenhower et al. (2006); Pharoah et al. (2013); Riester et al. (2014)). Yet, the formal investigation of the questions listed above in high dimensions from a statistical standpoint is relatively new, and has been identified as a critical need in several venues, including a recent NAS workshop on *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results* (National Academy of Sciences, 2015).

In Chapter 3 we propose a dimension-reduction approach that allows for joint analysis of multiple studies, achieving the goal of capturing biological factors. To this end, we propose a a generalized version of FA, able to handle multiple studies simultaneously. The proposed approach allows to learn the common biological features shared among the studies, identifying the unique variation present in each study. Chapter 3 is focused on the frequentist approach, and an Expectation-Conditional Maximization (ECM) algorithm is proposed for

parameter estimation. Several simulation studies are performed in order to assess properties of such method.

Chapter 4 describes some applications of the method. The published data sets employed are curated data collection for gene expression of patients with ovarian cancer (Ganzfried et al., 2013). Two different applications are provided and it is shown how the proposed method can identify the stable signal across multiple biology studies.

In Chapter 5 we propose a Bayesian approach to apply the method described in Chapter 3 to settings with more variables than subjects for which a frequentist approach is not feasible. In modeling dependencies among many variables, a sparse, parsimonious structure underlying the associations among genes is assumed. Indeed, the sparsity assumption implies that within a set of genes, only a few are interacting (Tegner et al., 2003). Bayesian sparsity model have been considered in genomic application and with different approaches: lasso prior (Hans, 2009; Park and Casella, 2008), shrinkage prior (Bhattacharya and Dunson, 2011) and spike and slab prior (Carvalho et al., 2008; George and McCulloch, 1993, 1997). There we focus on the latter one. By means of simulated data we evaluate the performance of the proposal method, an application to the ovarian cancer data sets is also included, illustrating the usefulness of the Bayesian approach.

The most original contribution of the thesis is the attention given to the two distinct components of variation in data, namely the part of communality across the studies related to the genuine biological signal, and the study-specific part associated to the artifactual and biological sources of variation. The key idea of the model is that the signal which shows such stability is more likely to capture genuine biology. At the same time, the proposal enables a more reliable identification of artifacts and thus facilitate more efficient experimental designs, driving technological advances.

Chapter 2

Some statistical methods in genomic applications

This chapter provides an introduction to the analysis of biological data and, in particular, gene expression data. The methods described in this thesis can have many applications, but we apply them to microarray gene expression. The chapter has been divided into three sections. The first section provides basic concepts on microarrays and describes the basic principles behind a microarray experiment. The second section deals with the representation and extraction of information from microarray experiments in the context of data integration. The third section addresses different methods for dimension reduction techniques in genomic applications. Particular attention will be given to factor analysis (FA).

2.1 Microarray experiments

The analysis of high-dimensional biological data sets is related to functional genomics (Do et al., 2006). A typical experiment of this kind amounts to observing the expression levels of an extensive quantity of genes simultaneously, named gene expression microarray analysis (Alberts, 2008).

Measurements of gene expression are crucial for biological understanding. Indeed, they are relevant to understanding common and complex diseases, such as cancer, detecting strategies to treat and prevent such diseases (Schulze and Downward, 2001). However, at the same time, they need particular attention since the quantity of data generated from each experiment is enormous and the biological signals of interest can be dominated by some sort of errors (Speed, 2003).

Most measurements from high-throughput experiments display variation arising from both biological and artifactual sources, due for example to technological and laboratory-based differences between studies.

For instance, biological differences include natural variations in different samples. On the other side, technological and laboratory-based differences are associated with different measurement laboratories or platforms in which gene expression data are collected (Irizarry et al., 2003; Kerr, 2007; Shi et al., 2006). Indeed, several techniques are available for measuring gene expression, and the data are currently collected on a diverse platform.

Microarray platforms are different from one another in some features that can influence their precision and efficiency (Deshwar and Morris, 2014). Although the most modern platforms (e.g., Affymetrix, Agilent, ABI, and Illumina) provide data of similar quality, they need to be normalized in the correct way (Shi et al., 2006).

However, some variations and study-specific effects could remain and sometimes could influence, if not included in analysis, the biological signal.

As noted in Garrett-Mayer et al. (2007), the fact that the determinants of both technological and biological variation differ across studies and laboratories implies that study-specific and laboratory-specific effects occur in most biological data sets. Study-specific effects can be so large to cover the biological signal for many genes (Aach et al., 2000).

2.1.1 Ovarian cancer data

In this context, we analyze data related to ovarian cancer. As reported in Waldron et al. (2014b), ovarian cancer is one of the most lethal cancer causing a large number of deaths among women. It has been studied in numerous clinical investigations (Siegel et al., 2012).

Multiple databases of gene expression data offer the potential to identify sets of genes predictive of cancer survival and of patient resistance to chemotherapy, using thousands of samples from multiple laboratories. In order to develop reproducible biomarker discovery, a data resource must be accurate and retain clinical variables of known importance as much as possible.

We have chosen some publicly available ovarian cancer (OC) microarray gene expression studies (Ganzfried et al., 2013) included in the `curatedOvarianData` package in `Bioconductor` (Gentleman et al., 2006). The package focuses on the study of the tumour and provide information regards patient survival and clinical annotation. Other main factors of interest included drug resistance, and the stage of OC.

The `curatedOvarianData` is a database of different studies and provide an optimal microarray data resource for genomic analysis. The package has standardized gene expression

studies and clinical data for 2970 ovarian cancer patients from 23 studies computed on 11 different platforms.

In the `curatedOvarianData` package, the problem of the multiple probe sets was solved by setting the genes with multiple probe sets at the highest mean across all data sets. The resulting data sets are then normalized.

Gene expression data are from public databases and provide information about platforms, stage of patients and percents of patients diagnosed with Stage III or Stage IV OC.

In this context we have chosen four data sets. Table 2.1 provides an overview of the studies with corresponding references.

Study	Samples	Platform	Late Stage[a] (%)	Reference
GSE9891	285	Affy U133Plus 2.0	85	Tothill et al. (2008)
GSE20565	140	Affy U133Plus 2.0	48	Meyniel et al. (2010)
GSE26712	195	Affy U133a	96	Bonome et al. (2008)
TCGA	578	Affy HT U133a	90	Network et al. (2011)

Table 2.1 Data sets: [a] only FIGO Stages III and IV

The total samples is $n = 1198$. For each data set the genes of interest in this research are the ones in common across the studies.

2.2 Data integration

To increase the reliability and efficiency of biological investigations, it is critical to combine data from several studies. As a result of multiple studies, some components can be found to be related across studies, revealing interesting common features across different populations, such as parameters that capture the relationship between genes and phenotypes.

However, as reported above, when considering multiple studies, most measurements from high-throughput experiments display variation arising from both biological and artifactual sources. Due to these challenges, the most critical step in cross-study analysis of gene expression is to identify a subset of genes that is biologically reproducible across studies and to more reliably remove idiosyncratic variation that lacks cross-study reproducibility.

The increased availability of ensembles of studies on related clinical populations, assaying technologies, and genomic features gives rise to two important statistical questions: i) To what extent is biological signal reproducibly shared across multiple studies? ii) How can this common signal be extracted? Furthermore, these questions need to be answered by

considering the challenges of learning common biological features shared among the studies and isolating the variation specific to each study.

There are several natural approaches for combining information from microarray studies. We will give a brief description of the three main ones.

2.2.1 Notation

Let us introduce some notation. We consider S studies, each with the same P genomic variables. The generic study s has n_s subjects and, for each subject, P-dimensional centered data vector $\mathbf{x}_{is}$ with $i = 1, \ldots, n_s$.

2.2.2 Individual study analyses

The first approach is to compute, separately for each study, statistics that summarize the relationship between each gene and the phenotype of interest. For example, one procedure is to combine the studies using methodologies such as combination of p-values (Rhodes et al., 2002). In this paper one-sided statistical tests are applied for the two null hypotheses, i.e. no genes are overexpressed and no genes are underexpressed in the context of prostate cancer.

Instead, Li and Ghosh (2012) propose 'assumption weighting', a weighted hypothesis testing was used for between-study variation.

Wang et al. (2004) introduce a Bayesian approach in order to integrate microarray data on matched genes from three different studies.

Garrett-Mayer et al. (2008) provide simple expression measures in order to compare different studies or platforms. They evaluate the "reliability" of gene expression across studies. They explain that when it has been considered only one genomic study, it can be hard to determine whether or not gene levels are "reliably" measured. There is the need to compare the same gene expression levels across two studies measured on different platforms. They are able to assess whether there is consistency by comparing the genes. If a gene varies in all the platforms, there is consistency. Gene reliability is defined through a correlation measure across studies. The integrative correlation (IC) is denoted by r_p^{ss+1}

$$r_p^{ss+1} = \frac{\sum_{p=1,p\neq p+1}^{P}(\rho_{sp(p+1)} - \bar{\rho}_{sp})(\rho_{(s+1)p(p+1)} - \bar{\rho}_{(s+1)p})}{\sqrt{\sum_{p=1,p\neq p+1}^{P}(\rho_{sp(p+1)} - \bar{\rho}_{sp})^2 \sum_{p=1,p\neq p+1}^{P}(\rho_{(s+1)p(p+1)} - \bar{\rho}_{(s+1)p})^2}}, \qquad (2.1)$$

where $\rho_{sp(p+1)}$ is the correlation between genes p and $p+1$ in study s, and $\bar{\rho}_{sp}$ is the average correlation between gene p and all the other P genes. This gene-specific measure can be used to detect which genes tend to be measured consistently across studies. Some genes showed

negative or absent trends due to different gene signal across studies. This fact can be related to, for example, artifacts of the experimental conditions. In this way Garrett-Mayer et al. (2008) choose genes to be included in the analysis based on their reliability, and they could compared the association between gene and phenotype.

Garrett-Mayer et al. (2008) amount to meta-analytic approach to analyze gene expression data. They not combine the data of different studies, but instead perform comparative analyses.

These methods and approaches define simple tools for the comparison and the combination of measures across two or more studies. In this approach, simple statistics are computed separately for each study.

2.2.3 Cross-study normalization methods

At the extreme of study combination, there are cross-study normalization methods that combine the sample measurements of each study into a single data set, in order to apply a single-study analysis.

Shen et al. (2004) use a Bayesian mixture formulation, to merge and analyze four different microarray studies to develop an inter-study "signature" in breast cancer. They combine multiple studies on a common probability scale and determine a meta-signature associated with breast cancer survival.

Hayes et al. (2006) analyze three diverse cohorts of patients with lung cancer using the IC approach. Through the IC a subset of genes is selected, and clusters of genes define tumor subtypes.

Johnson et al. (2007) propose a method for adjusting “batch effect”, i.e. differences due to experimental conditions. Many components can cause batch variations, such as the different platforms. The method proposed by Johnson et al. (2007) tries to solve the problem of batch effect by adding a parameter, related to batch effect for gene p, in a regression model. Then, they compare all the studies after adjusting for the batch effect.

Shabalin et al. (2008) develop another cross-platform method. Let define the platform $a = 1,\ldots,A$ of each study. In their approach, the observed value x_{psa} is a scaled block mean plus noise. The block mean is constant for a determined set of genes and sample values, and it is the same in each platform a. The slope and the variance of the noise depend on the gene p and the platform a. More precisely, Shabalin et al. (2008) assume that

$$x_{psa} = A_{\alpha^\star(p),\beta_a^\star(s),a} b_{pa} + c_{pa} + \sigma_{pa}\varepsilon_{psa}. \tag{2.2}$$

The functions $\alpha^\star$ and $\beta_a^\star$ with $a = 1,\ldots,A$, define the determined sets of genes and samples, respectively. The numbers $A_{\alpha^\star(p),\beta_a^\star(s),a}$ are the block means, while b_{pa} and c_{pa} represent sensitivity and offset parameters, respectively, that are specific to each gene p and platform a. In short, their method is based on the concept that there exist a group structure across studies defined by a common and constant gene profile.

2.2.4 Joint modeling

A third approach, intermediate between the two above, is to integrate biological information from different studies using a joint model. In this approach, only selected common features can be selected across studies. For example, these features could be parameters that capture the relationship between genes and phenotypes. Several methods have followed this approach.

Conlon et al. (2007) perform a Bayesian meta-analysis model. In their approach, the standardized expression means are not the same in each study, but there is a study-specific mean from a common population distribution. Inter-study variability is studied as a parameter in the model.

A hierarchical Bayesian model is performed by Scharpf et al. (2009b) to identify genes that show differential expression between two phenotypes.

A Bayesian mixture model is assumed in Xie et al. (2010).

In this thesis we adopt the latter, intermediate approach to integrate the diverse genomic studies, i.e. joint models, and we propose dimension-reduction tools that allow for joint analysis of multiple studies.

In the next section we show some different dimension reduction techniques and their practical consequences.

2.3 Factor models for biological data

In biological applications with large amount of data, dimension reduction techniques are required in order to summarize the information of interest and to capture the intrinsic biological characteristic or signal of the studies.

In the dimension reduction methods, the “best” way of reducing the dimension of data matrix is crucial. Indeed, if they are not used carefully there can be an enormous loss of information (Härdle and Simar, 2003).

Many methods are considered in the genomic field. Principal Component Analysis (PCA) were applied in genome-wide association studies (Hirschhorn and Daly, 2005; Price et al.,

2006). A very widely approach, a generalization of (PCA), is the Co-Inertia Analysis (CIA) (Culhane et al., 2003; Meng et al., 2014). CIA finds, as the PCA, the "best" axis, i.e. the axis that maximizes the variance of the observed variables through its spectral decomposition after transforming the data into a table of chi-square values. The chi-square values give the associations between each gene or variables and each array.

We develop methods that are generalized version of Factor Analysis (FA), able to handle multiple studies simultaneously. In the next section we will describe the FA in more details and its properties in biological data.

In order to describe the application of FA in the biological context, we give a brief description of FA (Mulaik, 2009, Section 8), its distributional properties and some modeling choices.

2.3.1 Factor analysis

We consider S studies each with the same P genomic variables. The generic study s has n_s subjects and centered data matrix $\mathbf{x}_{ps},\ p = 1,\ldots,P$. In the standard FA, the observed variables in study s are decomposed into J_s factors related to the source of study-specific variation.

Factor loadings relate linearly the observed variables to the latent factors. In particular, let l_{js} be a *study-specific* factor and $\lambda_{pjs},\ j = 1,\ldots,J_s$ be its loadings. The FA assumes that the vector $\mathbf{x}_{ps}$ for a generic subject i and for centered variable p is decomposed as:

$$\mathbf{x}_{ps} = \sum_{j=1}^{J_s} \lambda_{pjs} l_{js} + \mathbf{e}_{ps}, \tag{2.3}$$

where $\mathbf{e}_{ps}$ is a Gaussian error term with covariance matrix $\Psi_s = \mathrm{diag}(\psi_{1\mathrm{s}},\ldots,\psi_{\mathrm{ps}})$.

In order to extend the notation to the matrix form, let $\mathbf{x}_{is}$ the $p \times 1$ observation vector, and let $\mathbf{l}_{is},\ i = 1,\ldots,n_s$ the $J_s \times 1$ latent random specific vector, and $\Omega_s,\ s = 1,\ldots,S$ be the $P \times J_s$ corresponding factor loading matrix, where $J_s < p$.

Equation (2.3) can be rewritten as:

$$\mathbf{x}_{is} = \Omega_s \mathbf{l}_{is} + \mathbf{e}_{is} \quad i = 1,\ldots,n_s\,, \tag{2.4}$$

where $\mathbf{e}_{is}$ is a Gaussian error term. In particular, we assume that the marginal distribution of $\mathbf{l}_{is}$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_{j_s}$, where $\mathbf{I}$ indicates the identity matrix.

The $p \times 1$ random error vector $\mathbf{e}_s$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Ψ_s with $\Psi_s = \text{diag}(\psi_{s_1}^2, \ldots, \psi_{s_p}^2)$.

FA aims at explaining the dependence structure among high-dimensional observations through a decomposition of a $P \times P$ covariance matrix Σ_s given by

$$\Sigma_s = \Omega_s \Omega_s^\top + \Psi_s \,. \tag{2.5}$$

Given Ω_s and Ψ_s, the expected value of the factors for the study s can be computed through the linear projection

$$E[\mathbf{l}_{is}|\mathbf{x}_{is}] = \Omega_s^\top \Sigma_s^{-1} \mathbf{x}_{is},$$

with $i = 1, \ldots, n_s$. These properties imply the so called *local independence*, namely:

$$\text{Cov}(\mathbf{X}_s|\mathbf{l}_s) = \Psi_s.$$

This means that, given the latent variables, the manifest variables are conditionally independent. The conditional distribution of $\mathbf{X}_s$ is equal to:

$$\mathbf{X}_s|\mathbf{l}_s \sim \mathbf{N}_p(\Omega_s \mathbf{l}_s, \Psi_s)$$

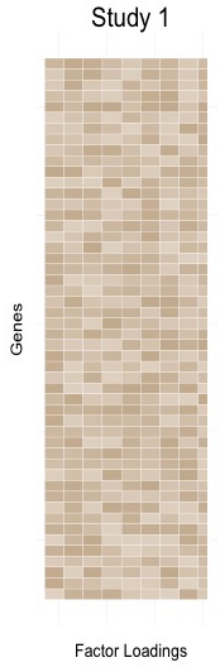

Figure 2.1 A graphical representation of the factor analysis performed in one study

If we conduct only one FA in one study, s, a possible schematic schematic representation is that provided in Figure 2.1. The heatmap 2.1 provides a graphical representation of the factor loadings, each rectangle is a factor loadings estimation. The darker rectangle corresponds to the the higher influence of latent factor on the genes or observed variables. Consequently, in this graphical representation we could underline pathways highly expressed genes.

To obtain a model free from identification problems, FA must be further constrained.

A specification of Ω_s and Ψ_s generates one and only one Σ_s; conversely different Ω_s can generate the same Σ_s. Indeed, if Ω_s is replaced by $\Omega_s^* = \Omega_s \mathbf{T}_s$, where $\mathbf{T}_s$ is a square orthogonal matrix ($\mathbf{T}_s\mathbf{T}_s^\top = \mathbf{T}_s^\top\mathbf{T}_s = \mathbf{I}$), we obtain

$$\Omega_s^*(\Omega_s^*)^\top = \Omega_s\mathbf{T}_s\mathbf{T}_s^\top\Omega_s^\top. \tag{2.6}$$

From these properties follow that

$$\Sigma_s = \Omega_s^*(\Omega_s^*)^\top + \Psi_s = \Omega_s\Omega_s^\top + \Psi_s.$$

Therefore, Σ_s is not uniquely identified. There are many ways of identifying the model by imposing constraints on Ω_s, including constraints to orthogonal Ω_s matrices and constraints such that $\Omega_s^\top\Sigma_s^{-1}\Omega_s$ is diagonal. The alternative preferred here is to constrain so that Ω_s is a block lower triangular matrix (Geweke and Zhou, 1996; Lopes and West, 2004), showed in Figure 2.2.

Notwithstanding this condition is largely used in classical FA settings, it should be notice that it induces an order-dependence among the variables (Frühwirth-Schnatter and Lopes, 2010). As stressed in Carvalho et al. (2008), the choice of the first J_s variables is an important modeling decision, to be made with some care.

2.3.2 Choice of the number of factors

A crucial step in the explorative FA is the choice of the appropriate number of factors (Guttman, 1954). There are many approaches for that choice. The latent factors explain the correlation between the manifest variables, so that it is important to carefully select their number. There are some approaches using the spectral decomposition of Σ_s and other approaches based on suitable tests.

Guttman (1954) made several proposals. In particular, he showed that, in a situation of perfect fit of the model to the observed data, the minimum number of factors is equal to the number of the eigenvalues of the correlation matrix greater than one. This quantity is called

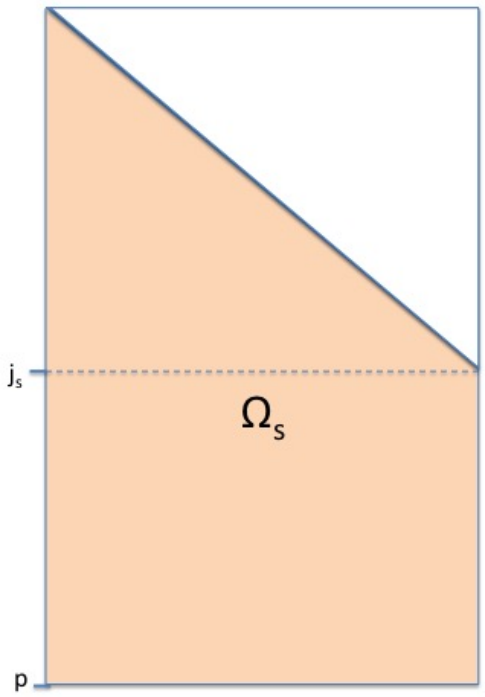

Figure 2.2 A schematic representation of the block lower triangular matrix.

"inferior boundary of Guttman". It is a very widely used rule, but often it is inadequate, since it is efficient only with small samples.

Another procedures is the Scree plot, or Cattell's scree test (Cattell, 1966). It is a procedure based on the eigenvalues represented in a plot. When the eingenvalue decreases, we stop to include further factors in the model.

Horn (1965) proposed parallel analysis (PA), a method based on simulation to determine the number of factors to retain. PA compares the observed eigenvalues extracted from the correlation matrix to be analyzed with those obtained from uncorrelated normal variables. From a computational point of view, PA uses a Monte Carlo simulation, since "expected" eigenvalues are obtained by simulating normal random samples that parallel the observed data in terms of sample size and number of variables. A latent factor is considered significant if the associated eigenvalue is larger than the mean of those obtained from the random uncorrelated data. Various studies indicate that PA is an appropriate method to determine the number of factors (Humphreys and Montanelli, 1975). Zwick and Velicer (1986) found that, among the methods analyzed, PA is the most accurate.

These methods are explorative in nature but further analyses, in some context and especially with the presence of large amount of data, need to be done.

More formal analyses may be based on some tests on the model fit, such as the likelihood ratio test, see Mulaik (2009, Section 8.3.5) for some details.

2.3.3 Factor analysis in the genomic field

Let consider now $s = 1, \ldots, S$ studies as in Figure 2.3.

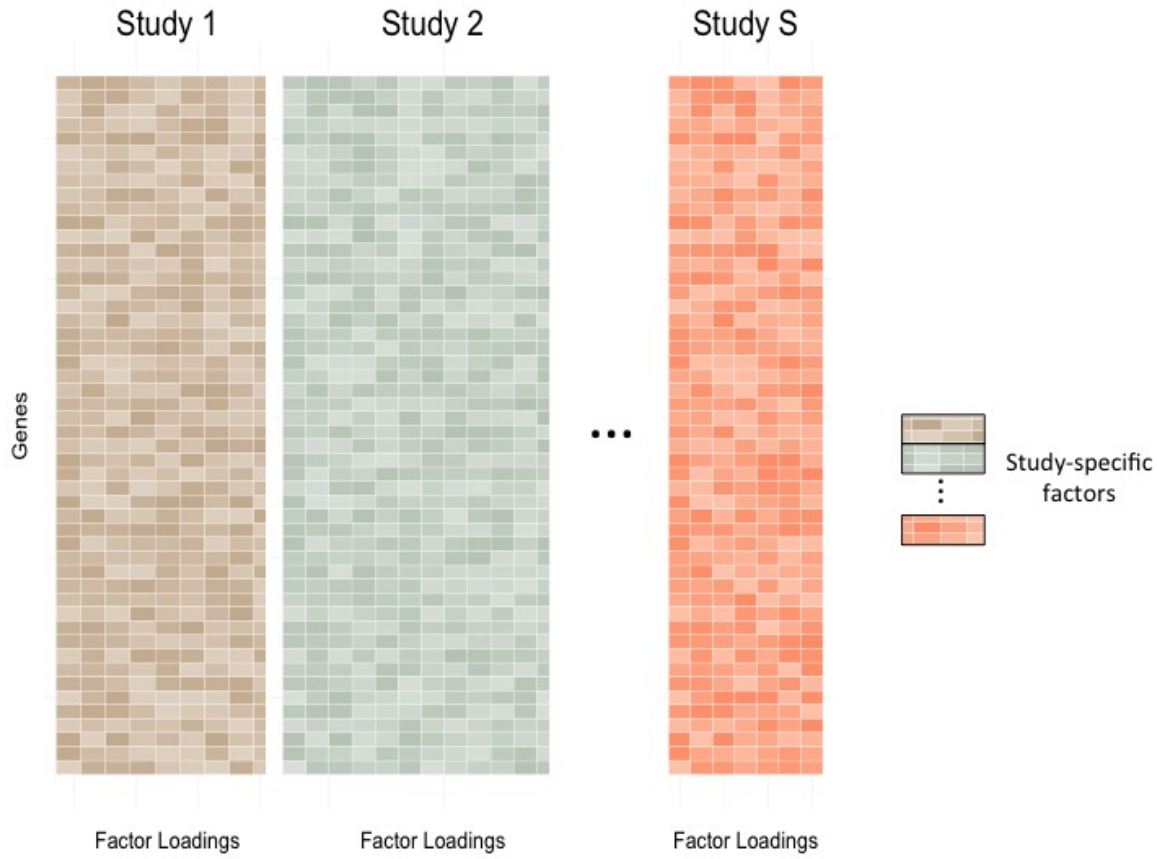

Figure 2.3 A schematic representation of the factor analysis performed in different studies

Factor analysis is largely used in multiple gene expression studies. Wang et al. (2011) use FA to obtain a unified gene expression measure from multiple platforms. Blum et al. (2010) use FA for multiple testing by Friguet et al. (2009) to characterize simple patterns of heterogeneity in gene expression data sets. And, finally, Runcie and Mukherjee (2013) use the sparse factor model by Bhattacharya and Dunson (2011) to capture subsets of important biological factors that control the variation in high-dimensional phenotypes.

The standard FA, given in equation (2.3), is performed in the first two studies in Table 2.1, Tothill et al. (2008) and Meyniel et al. (2010).

For each data set, the Immune System pathway is analyzed. The Immune System is of particular interest, as knowledge gained drives the development of targeted therapy (such as new antibody therapies) and tumor marker-based diagnostic tests (Méhes et al., 2001). For each study, the Immune System genes of interest are only those in common across the various studies.

Figure 2.4 Heatmap of the factor loadings obtained performing a separated factor analysis in the studies GSE9891 and GSE20565 as in Table 2.1

The heatmap in Figure 2.4 shows the estimated factor loadings, $\Omega_1 = \{\lambda_{11}, \lambda_{21}, \lambda_{31}, \lambda_{41}, \lambda_{51}, \lambda_{61}\}$ for the GSE9891 study and $\Omega_2 = \{\lambda_{12}, \lambda_{22}, \lambda_{32}, \lambda_{42}, \lambda_{52}, \lambda_{62}, \lambda_{72}\}$ for the GSE20565 study, obtained by performing a separated FA in each study. Each column λ_{is} is thus the i^{th} loading vector of the s^{th} study.

Figure 2.4 suggests that there are loading vectors in each study that exhibit a common pattern. This point is further explored in Figure 2.5, obtained starting from the cross-study pairwise correlation of some loading vectors, $\Omega_1 = \{\lambda_{11}, \lambda_{21}, \lambda_{31}, \lambda_{41}, \lambda_{51}, \lambda_{61}\}$ for the GSE9891 study and $\Omega_2 = \{\lambda_{12}, \lambda_{22}, \lambda_{32}, \lambda_{42}, \lambda_{52}, \lambda_{62}, \lambda_{72}\}$ for the GSE20565 study. We note that taking the cross-study pairwise correlation is meaningful since the same variables or genes are considered in each study. In Figure 2.5 darker lines denote larger correlations (in absolute value) compared to lighter ones, so that three of the loading vectors of the GSE9891 study are strongly correlated with four corresponding factors in the GSE20565 study.

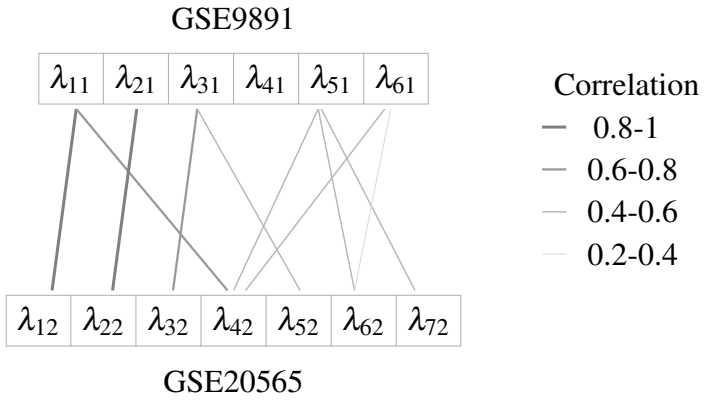

Figure 2.5 Graphical representation of the absolute value of correlations of the specific factor loadings obtained with factor analysis. Correlations smaller than .25 are not shown.

Highly correlated loading vectors are more likely to represent common factors, as they assign factor loadings that have similar pattern and interpretation across studies. On the other hand, some loading vector of GSE9891 exhibit no correlation with any loading vector of GSE20565 (e.g. λ_{41}). These loadings are then related to the uniqueness of the specific study. From a statistical perspective, it is important to focus on the common features shared among the studies, and on the connection with the biological signal. Then, a secondary important goal is the interpretation of study-specific variations that are only present in a single data set.

In order to meet the challenge of integrating multiple studies, we develop a generalized version of FA. The rest of this thesis will be devoted to such generalization.

Chapter 3

Multi-study factor model

In this chapter, we develop and study the multi-study factor model. In Section 3.1 we give a general description of the method, and why such analysis may be useful for genomic data. Furthermore, we introduce the model and the main assumptions in Section 3.2. Then we develop an estimation algorithm in Section 3.3 and we study the features and behavior of the proposed algorithm by means of some simulation study in Section 3.4.

3.1 Methods

As already surveyed in Chapter 2, gaining knowledge from high-dimensional studies is a cumulative process that requires integration of multiple, somewhat diverse studies, and relies critically on methods of analysis.

The methodology proposed here has three main goals. First, we combine multiple studies to identify common factors that are consistent across the studies. Second, we identify an additional variability component, specific of single studies, that is captured by study-specific latent factors. Third, the analysis of study-specific latent factors allow to identify possible idiosyncratic variation that lacks cross-study reproducibility.

Indeed, the model allows for a residual component, defined for each study and for each variable.

While the thesis is focused on genomic applications, the methods developed here can be applied to other situations where similarities and differences are warranted across multiple data sets.

In biological context, the approach proposed in this work can be applied to a large and different scale of studies. These studies may be microarray/ gene expression RNA-seq,

genome-wide association study (GWAS), or Electronic Medical Record (EMR) from different sources or systems.

Moreover, our tools enable a more reliable identification of artifacts and thus facilitate more efficient experimental designs and guide technological advances.

3.2 The multi-study factor model

In the social science literature, there is an extensive amount of methods developed for factor structures shared among different groups, forming the body of *multigroup factor analysis* methods; see, among many others, Jöreskog (1971); Meredith (1993); Thurstone (1931). Such large class of methods are generally focused on exploring measurement invariance among different groups, that typically results in testing whether the data support the hypothesis of a common loading matrix across groups. Here the emphasis is different, and even though there are some mathematical features in common with the models of multigroup FA, the existence of some study-specific factor loadings is always assumed, with equality assumed across studies only for a further set of loadings. The details are given as follows.

The Multi-study Factor Analysis (MFA) model proposed here can handle multiple studies, and allows to identify the common biological features shared among the studies, isolating the unique variation present in each study. The observed variables in study s are decomposed into K factors shared with the other studies, and J_s factors reflecting its unique sources of variation. Factor loadings relate linearly the observed variables to the latent factors.

Let f_k be a *common* factor, and $\phi_{kp},\ k = 1,\dots,K$ be its loadings; also let l_{js} be a *study-specific* factor and $\lambda_{pjs},\ j = 1,\dots,J_s$ be its loadings. The MFA assumes that the response $\mathbf{x}_{ps}$ for a generic subject i for variable p is decomposed as:

$$\mathbf{x}_{ps} = \sum_{k=1}^{K} \phi_{pk} f_k + \sum_{j=1}^{J_s} \lambda_{pjs} l_{js} + \mathbf{e}_{ps}, \tag{3.1}$$

where $\mathbf{e}_{ps}$ is the Gaussian error term.

We extend the notation in the matrix form included also the index i for the subject. Let $\mathbf{f}_i$ be the *common* factor, and Φ be its $P \times K$ common factor loading matrix and, moreover, let $\mathbf{l}_{is}$ be the *study-specific* factor with $i = 1,\dots,n_s$ and $\Lambda_s,\ s = 1,\dots,S$ be its $P \times J_s$ specific factor loading matrix. MFA assumes that the P-dimensional centered response $\mathbf{x}_{is}$ can be written as

$$\mathbf{x}_{is} = \Phi \mathbf{f}_i + \Lambda_s \mathbf{l}_{is} + \mathbf{e}_{is}\,, \tag{3.2}$$

where $\mathbf{e}_{is}$ is the Gaussian error term.

In MFA, the study s presents n_s subjects, P genomic variables and $K + J_s$ factors. We further 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 $\mathbf{f}_i$ is multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}_k$, where $\mathbf{I}$ indicates the identity matrix. Furthermore, the $p \times 1$ random error vector $\mathbf{e}_{is}$ has a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix Ψ_s with $\Psi_s = \text{diag}(\psi_{s_1}^2, \ldots, \psi_{s_p}^2)$. As a result, the marginal distribution of $\mathbf{x}_{is}$ is multivariate normal with mean vector $\mathbf{0}$ and covariance matrix

$$\Sigma_s = \Phi\Phi^\top + \Lambda_s\Lambda_s^\top + \Psi_s. \tag{3.3}$$

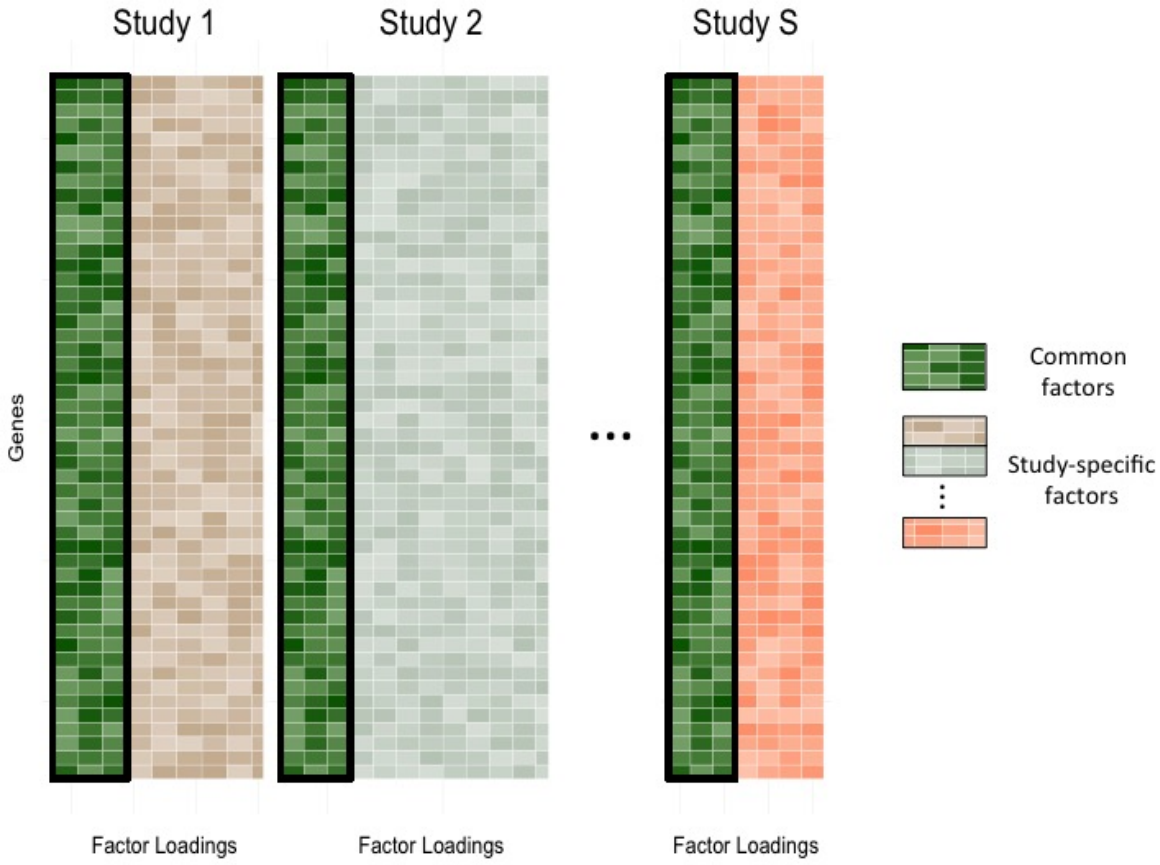

Figure 3.1 An illustrative example

A graphical representation of the results of model (3.2) is given in Figure 3.1. Our model is able to capture the genuine common biological features observed in multiple studies, identical for each study, and isolating the artifactual and biological sources of variation unique for each study.

A key consequence of the assumptions of model (3.2) is that the covariance matrix Σ_s decomposes as 3.3, with the three terms reflecting the variance of the common factors, the variance of the study-specific factors, and the variance of the errors.

3.2.1 Some distributional properties

The fundamental assumptions of the MFA have already been given but here more details are provided. We focus here on a given subject, dropping the *i* index.

The first relevant implication implies a structure for the variance of a manifest variables, namely

$$\begin{aligned}
\mathrm{Var}(x_{ps}) &= \mathrm{E}(x_{ps}-\mu_{ps})^2 = \phi_{p1}^2\mathrm{E}(f_1^2)+\cdots+\phi_{pK}^2\mathrm{E}(f_K^2)+ \\
&+ \lambda_{p1s}^2\mathrm{E}(l_{1s}^2)+\cdots+\lambda_{p1s}^2\mathrm{E}(l_{1s}^2)+\mathrm{E}(e_{ps}^2)= \\
&= \phi_{p1}^2+\cdots+\phi_{pK}^2+\lambda_{p1s}^2+\cdots+\lambda_{p1s}^2\mathrm{E}(l_{1s}^2)+\psi_{ps}= \\
&= \sum_{k=1}^{K}\phi_{pk}^2+\sum_{j=1}^{J_s}\lambda_{pjs}^2+\psi_{ps}.
\end{aligned} \tag{3.4}$$

The variance of x_{ps} is thus composed of three parts. The first, $\sum_{k=1}^{K}\phi_{pk}^2$, arises from what is common to all the studies. We called it *study commonality*. The second component, $\sum_{j=1}^{J_s}\lambda_{pjs}^2$, arises from what is common to all x_ss and called it *study specific*. The complementary part, ψ_{ps}, is the variance specific to that particular x_{ps}.

The covariance between x_{ps} and x_{ts}, due to the previous assumptions, is equal to

$$\begin{aligned}
\mathrm{Cov}(x_{ps},x_{ts}) &= \mathrm{E}(x_{ps}-\mu_{ps})(x_{ts}-\mu_{ts}) = \sum_{k=1}^{K}\sum_{h=1}^{K}\phi_{pk}\phi_{th}\mathrm{E}(f_kf_h)+ \\
&+ \sum_{j=1}^{J_s}\sum_{i=1}^{J_s}\lambda_{pjs}\lambda_{tis}\mathrm{E}(l_{js}l_{is})+\mathrm{E}(e_{ps}e_{ts})+\sum_{k=1}^{K}\sum_{i=1}^{J_s}\phi_{pk}\lambda_{tis}\mathrm{E}(f_kl_{is})+ \\
&+ \sum_{k=1}^{K}\phi_{pk}\mathrm{E}(f_ke_t)+\sum_{j=1}^{J_s}\sum_{h=1}^{K}\lambda_{pjs}\phi_{th}\mathrm{E}(l_{js}f_h)+\sum_{j=1}^{J_s}\lambda_{pjs}\mathrm{E}(l_{js}e_t)+ \\
&+ \sum_{h=1}^{K}\phi_{th}\mathrm{E}(e_{ps}f_h)+\sum_{i=1}^{J_s}\lambda_{tis}\mathrm{E}(e_{ps}l_{is})= \\
&= \sum_{k=1}^{K}\phi_{pk}\phi_{tk}+\sum_{j=1}^{J_s}\lambda_{pjs}\lambda_{tjs}.
\end{aligned} \tag{3.5}$$

The covariance between the p^{th} variable and the k^{th} *common* latent factors is equal to

$$\text{Cov}(x_{ps}, f_k) = \sum_{h=1}^{K} \phi_{ph}\text{E}(f_h f_k) = \phi_{pk}, \tag{3.6}$$

and the covariance between the p^{th} variable and the j^{th} *specific* latent factors is equal to

$$\text{Cov}(x_{ps}, l_{js}) = \sum_{i=1}^{J_s} \lambda_{pi}\text{E}(l_{is} l_{js}) = \lambda_{pj}. \tag{3.7}$$

The (3.4-3.7) equations are particularly important. The (3.4) equation highlights that the covariance between two manifest variables, x_{ps} and x_{ts}, $p,t = 1,\ldots,P$, is due to only the common and specific factors, i.e. the common factors shared between the studies and the specific factors shared between the variables in the study s. Moreover, the factor loadings, common and specific, can be explained as the covariance between the manifest variables and the latent factors, common and specific, in the (3.5), (3.6) and (3.7) equations.

Given Φ, Λ_s and Ψ_s, the expected value of the specific factor for the study s is given by the linear projection

$$E[\mathbf{l}_{is}|\mathbf{x}_{is}] = \Lambda_s^t \Sigma_s^{-1} \mathbf{x}_{is},$$

with $i = 1,\ldots,n_s$. In the same way, the conditional expected value of the common factors given $\mathbf{x}_{1s},\ldots,\mathbf{x}_{n_s s}$ is equal to

$$E[\mathbf{f}_i|\mathbf{x}_{is}] = \Phi^t \Sigma_s^{-1} \mathbf{x}_{is}.$$

The conditional expected values result from the joint normality of data, common and specific factors

$$\begin{bmatrix} \mathbf{X}_s \\ \mathbf{f} \\ \mathbf{l}_s \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} \underline{\mathbf{0}} \\ \underline{\mathbf{0}} \\ \underline{\mathbf{0}} \end{bmatrix}, \begin{bmatrix} \Sigma_s & \Phi & \Lambda_s \\ \Phi^t & \mathbf{I}_K & \mathbf{0} \\ \Lambda_s^t & \mathbf{0} & \mathbf{I}_{J_s} \end{bmatrix} \right).$$

The above distribution implies the *local independence* property, namely

$$\text{Cov}(\mathbf{X}_s|\mathbf{f}, \mathbf{l}_s) = \Psi_s.$$

This means that, for fixed the common and specific latent variables, the manifest variables are conditionally independent. The conditional distribution of $\mathbf{x}_{is}$ is equal to

$$\mathbf{x}_{is}|\mathbf{f}_i, \mathbf{l}_{is} \sim \mathbf{N}_p(\Phi\mathbf{f}_i + \Lambda_s\mathbf{l}_{is}, \Psi_s).$$

This result is required for fitting the model by the Expectation-Maximization (EM) algorithm, as exposed in what follows.

Notice that, given the observed variable $\mathbf{x}_{is}$, the common factor and the specific factor are not independent so that

$$\mathrm{Cov}[\mathbf{l}_{\mathrm{is}}, \mathbf{f}_{\mathrm{i}} | \mathbf{x}_{\mathrm{is}}] = \Phi^{\mathrm{T}} \Sigma_{\mathrm{s}}^{-1} \Lambda_{\mathrm{s}}.$$

The method proposed can be applied to many settings when the aim is to isolate commonalities and differences across different groups, population or studies. Here the focus is on the biological signal shared among the studies, removing study-specific features related to idiosyncratic error. This is only one of the applications of the MFA model. There might be other applications where the goal is to capture some study-specific features of interest and, instead, remove some common factors shared among the studies. Other applications may focus on capturing both common and specific factors, without removing any of them.

3.2.2 Set of identifiability conditions

To obtain an identifiable model, the MFA model must be further constrained to avoid orthogonal rotation indeterminacy, similarly to the classic FA model. Lack of identifiability can be simply assessed by considering the model for a subject in the s^{th} study. Let $\Omega_s = [\Phi, \Lambda_s]$ be the $P \times (K + J_s)$ loading matrix for the s^{th} study. If we define $\Omega_s^* = \Omega_s \mathbf{Q}_s$, where $\mathbf{Q}_s$ is a square orthogonal matrix with $(K + J_s)$ rows, it readily follows that $\Omega_s^* (\Omega_s^*)^\top = \Omega_s \mathbf{Q}_s \mathbf{Q}_s^\top \Omega_s^\top = \Omega_s \Omega_s^\top$, so that

$$\Sigma_s = \Omega_s^* (\Omega_s^*)^\top + \Psi_s = \Omega_s \Omega_s^\top + \Psi_s \,,$$

and Σ_s is not uniquely identified.

The standard factor model (2.4) identifies the parameters by imposing constraints on the matrix of factor loadings. One possibility often used in practice is to take Ω_s in (2.4) as being a block lower triangular matrix (Geweke and Zhou, 1996; Lopes and West, 2004).

Here we adapt this approach to the MFA model, and specify $\Omega_s = [\Phi, \Lambda_s]$ to be block lower triangular, as illustrated in Figure 3.2. With note that in such choice the matrices Φ and Λ_s are not interchangeable, and the number of elements of Λ_s set to zero is larger than the corresponding number for Φ.

The matrix representation of Ω_s is then

$$\Omega_s = \begin{bmatrix} \phi_{11} & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \phi_{21} & \phi_{22} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \phi_{31} & \phi_{32} & \phi_{33} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \phi_{k1} & \phi_{k2} & \phi_{k3} & \cdots & \phi_{kk} & 0 & \cdots & 0 \\ \phi_{(k+1)1} & \phi_{(k+1)2} & \phi_{(k+1)3} & \cdots & \phi_{(k+1)k} & \lambda_{11s} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{j1} & \phi_{j2} & \phi_{j3} & \cdots & \phi_{jk} & \lambda_{j1s} & \cdots & \lambda_{jjs} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi_{p1} & \phi_{p2} & \phi_{p3} & \cdots & \phi_{pk} & \lambda_{p1s} & \cdots & \lambda_{pjs} \end{bmatrix}.$$

Like in the standard FA models, assuming a block lower triangular form for Ω_s resolves the orthogonal rotation indeterminacy, for the same reason exposed in Geweke and Zhou (1996, pp. 565-566). There is residual labeling issue, as we can change the sign simultaneously to all the elements of the loading matrices and to all the latent factors without changing the model. This could be fixed by constraining the sign of a subset of loadings, but for maximum likelihood estimation of the model parameters this issue is largely inconsequential, as will be commented on later.

After assuming a suitable choice for Ω_s, it is important to note that the total number of elements in the sample covariance matrix $\mathbf{C}_{x_s x_s}$ for the s^{th} study must be less or equal to the free parameters in the covariance matrix Σ_s, implying that the following set of S conditions must hold

$$P(K+J_s)+P-\frac{(K+J_s)(K+J_s-1)}{2} \leq \frac{1}{2}P(P+1), \qquad s=1,\ldots,S.$$

This provides an upper bound on the number of the total latent dimension, common plus specific $(K+J_s)$

$$\begin{aligned} P &= 6 \longrightarrow K+J_s \leq 3 \\ P &= 12 \longrightarrow K+J_s \leq 7 \\ P &= 20 \longrightarrow K+J_s \leq 14. \end{aligned}$$

In MFA, we estimate the common and study-specific factor loadings by considering the constraints described above.

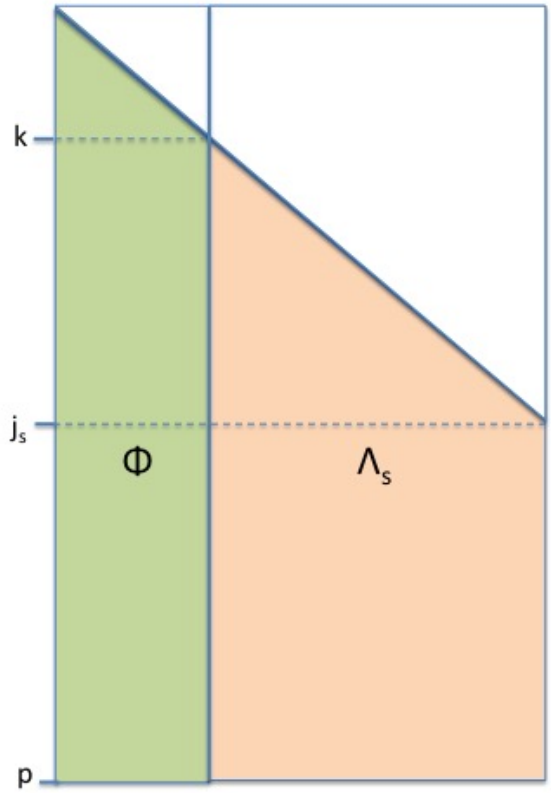

Figure 3.2 Construction of Ω_s.

3.3 Maximum likelihood estimation

The parameters to be estimated in the MFA consist of $\theta = (\Phi, \Lambda_s, \Psi_s)^\top$, as for notational simplicity in both (2.4) and (3.2) we assume that the observed variables in each study have been centered at the sample means.

Consequently, the marginal distribution of $\mathbf{x}_{is}$ given Φ, Λ_s, Ψ_s is multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\Sigma_s = \Phi\Phi^\top + \Lambda_s\Lambda_s^\top + \Psi_s$. The log-likelihood function corresponding to the MFA assumptions is given by

$$\ell(\theta) \;=\; \log \prod_{s=1}^{S} \prod_{i=1}^{n_s} p(\mathbf{x}_{is}|\theta) = \sum_{s=1}^{S} \left\{ -\frac{n_s}{2} \log|\Sigma_s| - \frac{n_s}{2} \mathrm{tr}(\Sigma_s^{-1} \mathbf{C}_{x_s x_s}) \right\},$$

where $\mathbf{C}_{x_s x_s}$ is the sample covariance matrix for the s^{th} study.

In order to maximize $\ell(\theta)$, we equate to zero its partial derivatives with respect to Φ, Λ_s, and Ψ_s. With a certain amount of algebra it is found that the score function for the specific factor loadings is

$$\frac{\partial}{\partial \Lambda_s}\ell(\theta) = -n_s\left(\Lambda_s^\top \Sigma_s^{-1} - \Lambda_s^\top \Sigma_s^{-1}\mathbf{C}_{x_s x_s}\Sigma_s^{-1}\right) , \tag{3.8}$$

whereas, for the common factor loadings is

$$\frac{\partial}{\partial \Phi}\ell(\theta) = \sum_{s=1}^{S}\left\{-n_s\left(\Phi^\top \Sigma_s^{-1} - \Phi^\top \Sigma_s^{-1}\mathbf{C}_{x_s x_s}\Sigma_s^{-1}\right)\right\} , \tag{3.9}$$

and for the covariance matrix of the error term is

$$\frac{\partial}{\partial \Psi_s}\ell(\theta) = -\frac{n_s}{2}\left(\Sigma_s - \Sigma_s \mathbf{C}_{x_s x_s}\Sigma_s\right) . \tag{3.10}$$

In the following, we compute the Maximum Likelihood Estimate (MLE) by the Expectation Conditional Maximization (ECM) algorithm.

3.3.1 Computation of MLE using the ECM algorithm

The Expectation-Maximization (EM) algorithm is a standard technique to compute maximum likelihood (ML) estimates, especially suitable for the context of missing data problems.

Maximum likelihood of the MFA can be conceptualized as MLE in a multivariate normal model with missing data. Indeed, in MFA the observed variables are influenced by two different latent components, namely the common latent component, through Φ, and the specific latent component, through Λ_s.

The EM algorithm (Dempster et al., 1977; Rubin and Thayer, 1982) compute the maximum likelihood estimations treating the latent variables, i.e. $\mathbf{f}$ and $\mathbf{l}_s$, as observed variables.

Choosing some starting values for all the parameters, we need to write down the complete log-likelihood of $(\mathbf{x}_{is}, \mathbf{f}_i, \mathbf{l}_{is})$ for $i = 1, \dots n_s$ given the parameters $\theta = (\Phi, \Lambda_s, \Psi_s)$. If $\mathbf{f}$ and $\mathbf{l}_s$ were observed, the log-likelihood would be

$$\begin{aligned} l_c(\theta) = \sum_{s=1}^{S}\Bigg\{ & -\frac{n_s}{2}\log|\Psi_s| - \frac{1}{2}\sum_{i=1}^{n_s}\left(\mathbf{x}_{is} - \Phi\mathbf{f}_i - \Lambda_s\mathbf{l}_{is}\right)^\top \Psi_s^{-1} \\ & \left(\mathbf{x}_{is} - \Phi\mathbf{f}_i - \Lambda_s\mathbf{l}_{is}\right) - \frac{1}{2}\mathbf{f}_i\mathbf{f}_i^\top - \frac{1}{2}\mathbf{l}_{is}\mathbf{l}_{is}^\top\Bigg\} . \end{aligned} \tag{3.11}$$

More in details, the complete log-likelihood is

$$
\begin{aligned}
l_c(\theta) &= \sum_{s=1}^{S}\left\{-\frac{n_s}{2}\log|\Psi_s| - \frac{1}{2}\sum_{i=1}^{n_s}\left(\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}\right)^\top\Psi_s^{-1}\right. \\
&\quad \left.\left(\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}\right)-\frac{1}{2}\mathbf{f}_i\mathbf{f}_i^\top-\frac{1}{2}\mathbf{l}_{is}\mathbf{l}_{is}^\top\right\} \\
&= \sum_{s=1}^{S}\left\{-\frac{n_s}{2}\log|\Psi_s| - \frac{1}{2}\sum_{i=1}^{n_s}\left(\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}\right)^\top\Psi_s^{-1}\left(\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}\right)\right\} \\
&= \sum_{s=1}^{S}\left\{-\frac{n_s}{2}\log|\Psi_s| - \frac{n_s}{2}\mathrm{tr}\left(\Psi_s^{-1}\frac{\sum_{i=1}^{n_s}\left(\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}\right)^\top}{n_s}\right.\right. \\
&\quad \left.\left.\frac{\left(\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}\right)}{n_s}\right)\right\} \\
&= \sum_{s=1}^{S}\left\{-\frac{n_s}{2}\log|\Psi_s| - \frac{n_s}{2}\mathrm{tr}\left(\Psi_s^{-1}\frac{\sum_{i}^{n_s}\left(\mathbf{x}_{is}\mathbf{x}_{is}^\top+\Phi\mathbf{f}_i\mathbf{f}_i^\top\Phi^\top+\Lambda_s\mathbf{l}_{is}\mathbf{l}_{is}^\top\Lambda_s^\top\right.\right.\right. \\
&\quad \left.\left.\left.\frac{-2\mathbf{x}_{is}\mathbf{f}_i^\top\Phi^\top-2\mathbf{x}_{is}\mathbf{l}_{is}^\top\Lambda_s^\top+2\Phi\mathbf{f}_i\mathbf{l}_{is}^\top\Lambda_s^\top)}{n_s}\right)\right\}.
\end{aligned}
\tag{3.12}
$$

To find the maximum likelihood estimations, the EM algorithm uses the complete log-likelihood (3.12).

There are two steps in each cycle of the EM algorithm. First, at the E step of the t^{th} iteration, we find the expectation of $l_c(\theta)$ by integrating over the latent variables $\mathbf{f}_i$ and $\mathbf{l}_{is}$ conditional distributions, with the parameter θ held fixed at the value θ_{t-1}.

$$
E\left\{l_c(\theta)|\mathbf{x}_{is},\theta_{t-1}\right\} = E\left[\sum_{s=1}^{S}\sum_{i=1}^{n_s}\log\left\{p(\mathbf{x}_{is}\mathbf{f}_i\mathbf{l}_{is}|\theta)p(\mathbf{f}_i|\theta)p(\mathbf{l}_{is}|\theta)\right\})|\mathbf{x}_{is},\theta_{t-1}\right].
$$

The second step of the EM algorithm, the M-step, requires to maximize the expected log-likelihood, and such maximization yields the next value θ_t of the parameter

$$
\frac{\partial E\left[l_c(\theta)|\mathbf{x}_{is},\theta_{t-1}\right]}{\partial\theta} = 0.
$$

Using the new value we can then proceed to the next $t+1$ iteration.

EM algorithms have two main properties (Dempster et al., 1977). First, each EM iteration increases the log-likelihood. Second, it converges to a local maximum of the log-likelihood.

In some contexts, handling the complete data could be complex. Indeed, finding a maximum of the log-likelihood especially with many parameters can be hard. A trivial

approach, generally, consists in dividing a large problem into several smaller ones (McLachlan and Krishnan, 2007, Section 5.2, p. 160). With many parameters, we could consider some parameters as known, and estimate the remaining ones.

This is the main idea of the ECM algorithm where the M-step of the EM algorithm is replaced by some conditional M-steps in which each step maximizes one parameter conditional to the remaining ones held fixed at their previous values.

The ECM algorithm has the same important properties of convergence of the EM algorithm (McLachlan and Krishnan, 2007, Section 1.7, p. 28).

E-step

As we saw from equation (3.12), finding the expectation of $l_c(\theta)$ given $\mathbf{x}_{is}$ for each $s = 1, \ldots, S$ and θ_{t-1} requires finding the conditional expectation of the following quantities

$$\begin{aligned}
\mathbf{C}_{x_s x_s} &= \frac{\sum_{i=1}^{n_s} \mathbf{x}_{is}\mathbf{x}_{is}^\top}{n_s}, \qquad & \mathbf{C}_{l_s l_s} &= \frac{\sum_{i=1}^{n_s} \mathbf{l}_{is}\mathbf{l}_{is}^\top}{n_s}, \\
\mathbf{C}_{x_s l_s} &= \frac{\sum_{i=1}^{n_s} \mathbf{x}_{is}\mathbf{l}_{is}^\top}{n_s}, \qquad & \mathbf{C}_{ff} &= \frac{\sum_{i=1}^{n_s} \mathbf{f}_i\mathbf{f}_i^\top}{n_s}, \\
\mathbf{C}_{x_s f} &= \frac{\sum_{i=1}^{n_s} \mathbf{x}_{is}\mathbf{f}^\top}{n_s}, \qquad & \mathbf{C}_{f l_s} &= \frac{\sum_{i=1}^{n_s} \mathbf{f}_i\mathbf{l}_{si}^\top}{n_s}.
\end{aligned}$$

For each quantity defined above we compute the conditional expectation given the parameters and the observed data:

$$\begin{aligned}
\mathbf{T}_{x_s x_s} &= \mathrm{E}[\mathbf{C}_{\mathrm{x_s x_s}}|\mathbf{x}_{\mathrm{is}}, \theta_{\mathrm{t-1}}] = \mathbf{C}_{\mathrm{x_s x_s}}, \\
\mathbf{T}_{x_s l_s} &= \mathrm{E}[\mathbf{C}_{\mathrm{x_s l_s}}|\mathbf{x}_{\mathrm{is}}, \theta_{\mathrm{t-1}}] = \mathbf{C}_{\mathrm{x_s x_s}}\delta_{\mathrm{s}}^\top, \\
\mathbf{T}_{x_s f} &= \mathrm{E}[\mathbf{C}_{\mathrm{x_s f}}|\mathbf{x}_{\mathrm{is}}, \theta_{\mathrm{t-1}}] = \mathbf{C}_{\mathrm{x_s x_s}}\delta^\top, \\
\mathbf{T}_{l_s l_s} &= \mathrm{E}[\mathbf{C}_{\mathrm{l_s l_s}}|\mathbf{x}_{\mathrm{is}}, \theta_{\mathrm{t-1}}] = \delta_{\mathrm{s}}\mathbf{C}_{\mathrm{x_s x_s}}\delta_{\mathrm{s}}^\top + \Delta_{\mathrm{s}}, \\
\mathbf{T}_{f f} &= \mathrm{E}[\mathbf{C}_{\mathrm{ff}}|\mathbf{x}_{\mathrm{is}}, \theta_{\mathrm{t-1}}] = \delta\mathbf{C}_{\mathrm{x_s x_s}}\delta^\top + \Delta, \\
\mathbf{T}_{f l_s} &= \mathrm{E}[\mathbf{C}_{\mathrm{f l_s}}|\mathbf{x}_{\mathrm{is}}, \theta_{\mathrm{t-1}}] = \delta\mathbf{C}_{\mathrm{x_s x_s}}\delta_{\mathrm{s}}^\top + \mathrm{Cov}[\mathrm{l_{is}}, \mathrm{f_i}|\mathbf{x}_{\mathrm{is}}, \theta_{\mathrm{t-1}}],
\end{aligned} \tag{3.13}$$

where

$$\begin{aligned}
\delta &= \Phi^\top \Sigma_s^{-1}, \\
\delta_s &= \Lambda_s^\top \Sigma_s^{-1}, \\
\Delta &= \mathrm{Var}[\mathbf{f}_{\mathrm{i}}|\mathbf{x}_{\mathrm{is}}] = \mathbf{I}_{\mathrm{k}} - \Phi^\top \Sigma_{\mathrm{s}}^{-1}\Phi, \\
\Delta_s &= \mathrm{Var}[\mathbf{l}_{\mathrm{is}}|\mathbf{x}_{\mathrm{is}}] = \mathbf{I}_{\mathrm{j_s}} - \Lambda_{\mathrm{s}}^\top \Sigma_{\mathrm{s}}^{-1}\Lambda_{\mathrm{s}}.
\end{aligned} \tag{3.14}$$

CM-step

The CM-step of the ECM algorithm maximizes the expected log-likelihood to find the estimate of a parameter conditional to the other parameters found at the previous steps, $l_c(\theta^{(t+1)}|\theta^{(t)})$. In our case $\theta = \{\Phi, \Lambda_s, \Psi_s\}$, so the CM update is then developed in three steps:

1. CM1 starts with the update of the error covariance matrix Ψ_s keeping the other parameters at their initial values

$$\begin{aligned} \Psi_s^{new} &= \text{diag}\Big(\mathbf{C}_{\mathrm{x_s x_s}} + \Phi\mathbf{T}_{\mathrm{ff}}\Phi^\top + \Lambda_{\mathrm{s}}\mathbf{T}_{\mathrm{l_s l_s}}\Lambda_{\mathrm{s}}^\top - 2\mathbf{T}_{\mathrm{x_s f}}\Lambda_{\mathrm{s}}^\top \\ &- 2\mathbf{T}_{x_s l_s}\Lambda_s^\top + 2\Phi\mathbf{T}_{f l_s}\Lambda_s^\top\Big) . \end{aligned} \tag{3.15}$$

2. CM2 updates Φ keeping Λ_s at the initial value and Ψ_s at the CM1 value

$$\begin{aligned} \text{vec}(\Phi^{new}) &= \sum_{s=1}^{S}\Big(\mathbf{T}_{ff}^\top \otimes n_s\Psi_s^{-1new}\Big) \\ &\quad \text{vec}\Big(n_s\Psi_s^{-1new}\mathbf{T}_{x_s f} - n_s\Psi_s^{-1new}\Lambda_s\mathbf{T}_{f l_s}^\top\Big) , \end{aligned} \tag{3.16}$$

   where $\otimes$ is the Kronecker product, vec is the vec operator and the linear equation is solved by the Lyapunov Equation, see for example Van Loan (2000).

3. CM3 updates Λ_s keeping Φ at the CM2 value and Ψ_s at their CM1 value

$$\Lambda_s^{new} = \Big(\mathbf{T}_{x_s l_s} - \Phi^{new}\mathbf{T}_{f l_s}\Big)\Big(\mathbf{T}_{l_s l_s}\Big)^{-1} . \tag{3.17}$$

Stopping Rule

It is really important to stop the algorithm when it arrives at convergence. The stopping criterion usually adopted with the EM algorithm is in terms of either the size of the relative change in the parameter estimates or the log-likelihood. One example is provided by Zhao et al. (2008).

For notation simplicity, let $l_c(\theta^{(t+1)}) = l^{(t+1)}$ be the complete log-likelihood obtained at $(t+1)$ iteration. Zhao et al. (2008) stop the EM algorithm if

$$l^{(t+1)} - l^{(t)} < \text{tol} ,$$

or $t > K_{max}$ with tol $= 10^{-6}$ and the maximal number of iteration, $K_{max} = 5000$.

Instead, Böhning et al. (1994) exploit Aitken's acceleration procedure, (McLachlan and Krishnan, 2007, Section 4.9, p. 142). This is the rule adopted here. It is applicable in the case where the sequence of log likelihood values $\left\{l^{(t)}\right\}$ linearly converges to a value $l^{\star}$. Under this assumption,

$$l_A^{(t+1)} = l^{(t)} + \frac{1}{1-c^{(t)}}(l^{(t+1)} - l^{(t)}),$$

where $c^{(t)} = (l^{(t+1)} - l^{(t)})/(l^{(t)} - l^{(t-1)})$. The ECM can be stopped if

$$|l_A^{(t+1)} - l_A^{(t)}| < \text{tol}.$$

3.4 Simulation studies

Some simulation experiments are designed to evaluate the ECM algorithm performances in estimating the MFA model parameters, and to assess the strategy for selecting the dimension of the latent factors. For the latter task, the following procedure is proposed. Given some data sets composed by S studies, the first step is to determine the total latent dimension for each study. For the MFA model, the latter is defined as

$$T_s = K + J_s. \tag{3.18}$$

The total latent dimension T_s for each study can be determined through techniques used in the standard FA, such as Horn's parallel analysis (Horn, 1965), Cattell's scree test (Cattell, 1966) and the use of indexes, such as the RMSEA (Steiger and Lind, 1980). Afterwards, some model selection techniques can be employed to select the value of the number K of latent factors sharing a common loading matrix Φ. The dimension J_s are then obtained as $T_s - K$.

This methodology has been tested by designing a simulation study mirroring the results obtained for the data of Table 2.1, considering the same $p = 100$ variables in each case. Therefore, $S = 4$ studies are considered, with the dimension of the latent factors reported in Table 3.1.

Three simulation scenarios are considered. In Scenario 1 there are no common factors, i.e. $K = 0$, in Scenario 2 we set $K = 1$, and finally in Scenario 3 we set $K = 3$. In each case, the data are generated by parameter values akin to those estimated with the data.

Table 3.1 Settings for the simulation studies.

S	n_s	$K+J_s$
1	285	6
2	140	7
3	195	11
4	578	10

3.4.1 Parameter estimation via the ECM algorithm

We first analyze the performances of the ECM algorithm for a given selection of K and J_s, $s = 1,\ldots,S$. In particular the results obtained with the ECM algorithm and by a standard optimizer are compared. The standard optimizer employed is the box-constrained Limited-memory Broyden-Fletcher-Goldfarb-Shannon (L-BFGS) method (e.g. Byrd et al., 1995), as implemented in the `R` function `optim`. Such method are denoted as *direct optimizer*, to stress that it was taken for benchmarking the ECM algorithm, without any particular effort to tailor the optimization to the model at hand.

Regardless of the optimization method adopted, the choice of the starting point is crucial for achieving good performances. The following strategy, given the factor dimensions K and J_s, $s = 1,\ldots,S$, has been used.

1. A single data set is created by stacking the data of the four studies by row, obtaining a single data set with $n = n_1+n_2+n_3+n_4 = 578+285+195+140 = 1198$ observations and $p = 100$ variables.

2. A Principal Components Analysis (PCA) is performed on the data set obtained at the first step. The first K principal components are taken as the starting point of the common factor loadings.

3. The variance of the K principal components is removed from each study.

4. A standard FA model is fitted to the data of each study separately. The factor loadings and uniquenesses obtained from the standard FA are used as the initial values for Λ_s and Ψ_s.

Figure 3.3 reports the log-likelihood function along the algorithm iterations for three illustrative data sets, after starting the two algorithms from the same point. In particular, the red line denotes the value of the log-likelihood function at the true parameter value, the blue line denotes the value of the maximized log-likelihood for the ECM algorithm, and the green line the corresponding value for the L-BFGS. Figure 3.3 shows the progress of the two

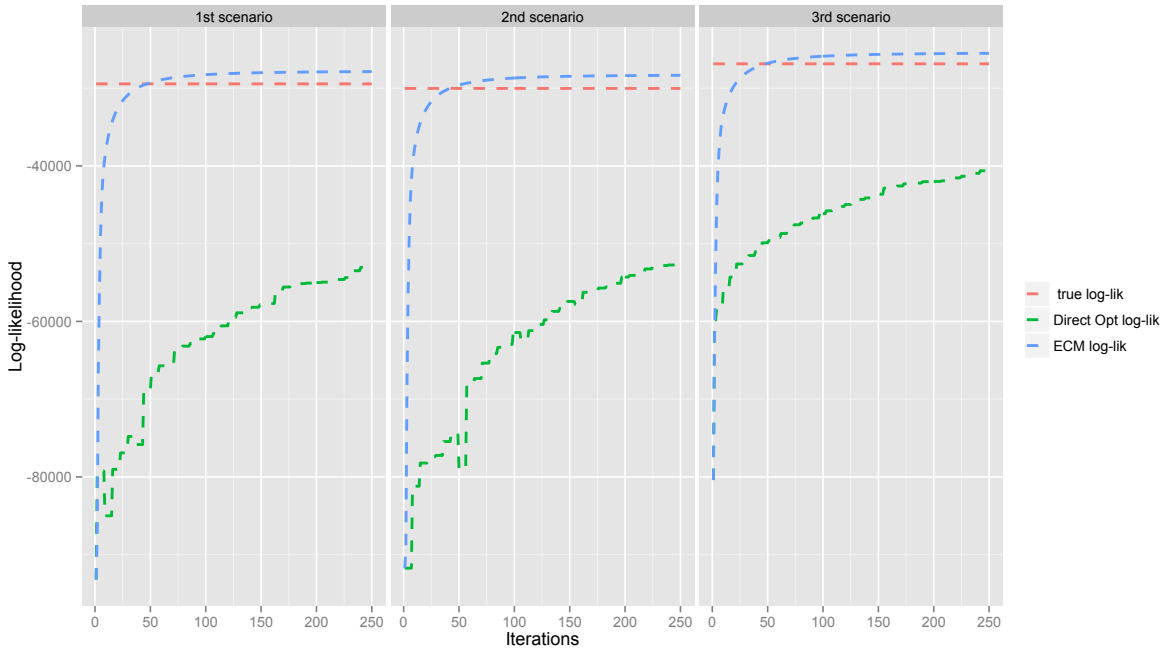

Figure 3.3 Comparison of the log-likelihood function obtained with the ECM algorithm (blue line) and with the direct optimizer (green line), whereas the red line represents the log-likelihood calculated at the true parameter value. The functions are computed based on three simulated data sets, one for each of the three scenarios.

algorithms for the first 250 iterations, with a clear suggestion of much faster convergence for the ECM algorithm. Indeed, convergence is attained for both methods, but for the *direct optimizer* the required number of iterations is much higher, around 20,000.

Table 3.2 Time of convergence (s) for each scenario.

Method	Scenario 1	Scenario 2	Scenario 3
ECM	14.25	12.66	10.74
L-BFGS	89.79	83.41	75.62

The speed of convergence is reflected not only in the number of iterations, but also in the processing time shown in Table 3.2. Moreover, the L-BFGS was plagued by severe convergence problems when used for larger number of variables, i.e. $P > 100$.

In Figure 3.4 the parameter estimates obtained with the ECM algorithm are represented, for 100 simulated samples in the Scenario 2. In particular, the empirical distribution of the difference between the estimated $\widehat{\Phi}$ and the true Φ are summarized by a boxplot for each element of the matrix. The overall impression is that the estimation bias is generally negligible. Here the sign indeterminacy alluded to in §3.2.2 is resolved by considering the

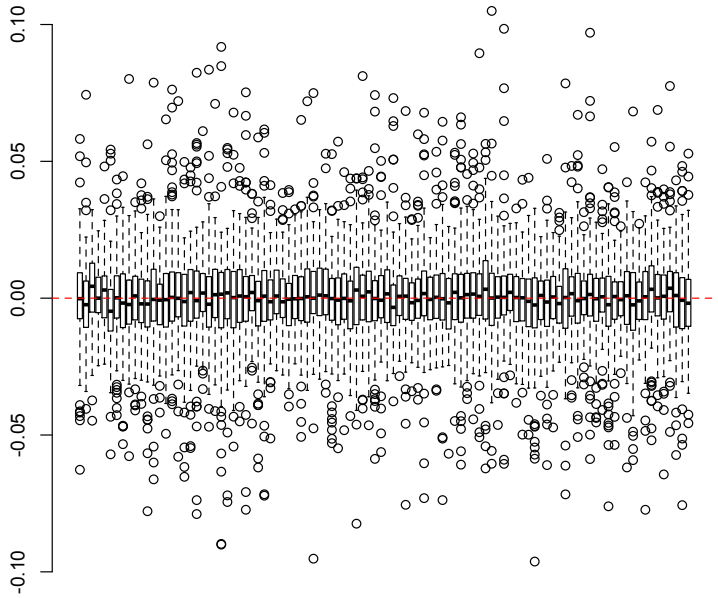

Figure 3.4 Distribution of the differences between each elements of Φ estimated by the ECM algorithm and the corresponding elements of Φ used to generate the data.

absolute value of the estimated loadings. Indeed, the local convergence property of the ECM algorithm resolves this issue in any given data sets, but for comparing the results of several simulated data sets some sort of post-processing is required. We note in passing that Adachi (2012) reported that the EM algorithm in FA models always gives proper solutions when the sample covariance and initial parameter matrices are proper. Likewise, the same result is empirically found in all the simulated data sets performed for the MFA model.

Finally, the Likelihood Ratio Test (LRT) for choosing between $K = 0$ and $K = 1$ is studied in order to confirm that standard likelihood asymptotics hold for the problem at hand. This is actually the case, as shown in Figure 3.5, that represents the estimated distribution of the LRT, obtained with 1,000 simulated data sets under Scenario 2, which is intermediate between the other two scenarios. The empirical distribution of the LRT seems close to the asymptotic distribution, and this also provides some further (albeit indirect) suggestion that the ECM is actually able to correctly locate the MLE.

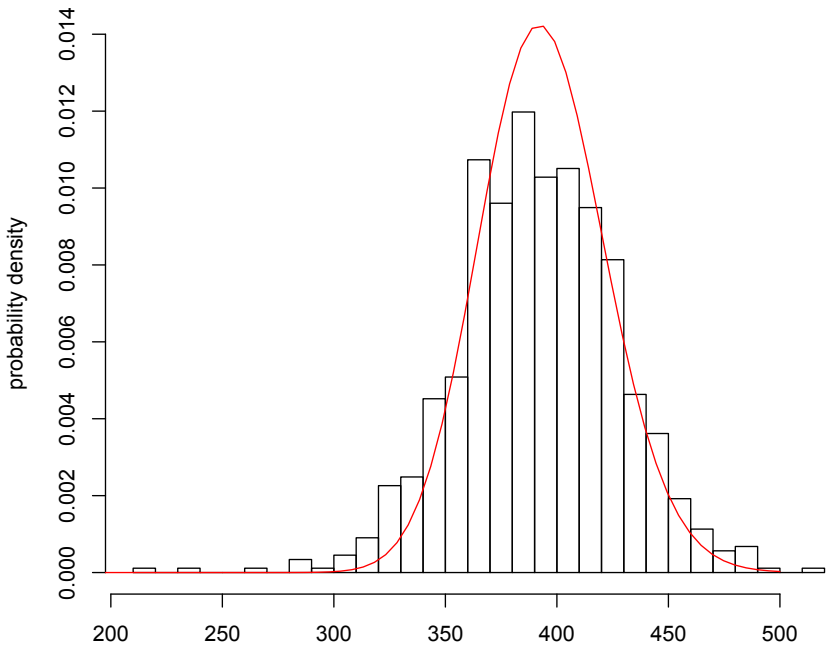

Figure 3.5 Distribution of the LRT for choosing between $K = 0$ and $K = 1$, obtained by 1,000 simulated data sets under Scenario 2. The red line represents the asymptotic distribution, a chi-square distribution with 192 degrees of freedom.

3.4.2 Selection of the latent factor dimensions

We consider here the problem of selecting the dimension of the latent space, again by means of some simulation studies performed under the same three scenarios considered above. For each data set, the same strategy aforementioned is followed, namely first T_s was chosen in each study by means of standard FA techniques, and then K was selected.

We focus in particular on the problem of selecting K, tackled by applying standard model selection techniques, such as the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz et al., 1978), for which there is an extensive literature (Burnham and Anderson, 2002; Preacher and Merkle, 2012). The study of model selection based on information criteria is still in progress in FA settings (Chen and Chen, 2008; Hirose and Yamamoto, 2014), so it seems useful to evaluate the behavior of both AIC and BIC for choosing K. Along the two information criteria, the likelihood ratio test (LRT) for choosing between nested models with different values of K was also considered.

Table 3.3 shows the results obtained by model fitting simulations for 100 different data sets generated independently from the MFA with $K = 0$, i.e. Scenario 1. The overall impression is that all the three methods tend to choose the model with $K = 0$, but both AIC and LRT outperform BIC.

Table 3.3 Comparison of model assessment methods in Scenario 1.

Method	$K = 0$	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
AIC	100	0	0	0	0	0
BIC	91	1	2	6	0	0
LRT	100	0	0	0	0	0

Table 3.4 reports the results for 100 different data sets generated independently from the MFA with $K = 1$, i.e. Scenario 2. Here AIC outperforms both BIC and LRT, though the latter is not much off the mark. The poor performance of the BIC is striking, as it often leads to a model with $K = 5$, thus over-simplifying the selected model. Indeed, BIC penalizes model complexity more strongly than AIC, so it is not surprising that BIC tends to prefer models with more common factors and thus less parameters. What it is worrisome is the intensity of such tendency for the problem at hand.

Table 3.4 Comparison of model assessment methods under Scenario 2.

Method	$K = 0$	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
AIC	0	100	0	0	0	0
BIC	0	0	0	2	6	92
LRT	3	97	0	0	0	0

Finally, Table 3.5 reports the results based on 100 different data sets are generated independently from the MFA with $K = 3$, i.e. Scenario 3. Again, AIC seems the best criterion, always leading to the selection of the true model.

Table 3.5 Comparison of model assessment methods under the Scenario 3.

Method	$K = 0$	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
AIC	0	0	0	100	0	0
BIC	0	0	0	1	23	76
Lik Ratio Test	0	0	0	91	9	0

The results of these three simulation studies point strongly towards the usage of AIC to select the value of K. This will be the strategy employed in the following chapter.

Chapter 4

Ovarian Cancer application

In order to validate the proposed methods, we have analyzed the four studies described in Table 2.1.

We have focused on common genes across studies and included in the Immune System pathway and the DNA-repair pathway. The Immune System pathway is really important in the field for the development of therapy (for example antibody therapies) and tumor diagnostic tests (Méhes et al., 2001). The DNA-repair pathway is of particular interest for the understanding of the insurgence of OC in the presence of reduced DNA repair capacity. For each study, the Immune System and DNA repair genes of interest are the ones in common across the studies. The analyses of this two pathways are developed in Section 4.1 and 4.2. Section 4.3 provides a discussion.

4.1 Immune system

In the Immune system, three different sub-pathways are included "Adaptive Immune System" (AI), "Innate Immune System" (II) and "Cytokine Signaling in Immune System" (CSI) obtained from `reactome.org` and belonging to the Immune System pathway. These sub-pathways do not have overlapping genes.

Initially, we have done some preliminary analyses in order to asses the total latent factor dimensions, the number of common factors across studies and the number of specific factors for each study. Using the AIC, the number of common factors is set to one, and the number of specific factors for each study results as showed in Table 4.1.

We compare the prediction errors computed by MFA to those computed by standard factor analysis (FA) applied separately to each study. We fit the MFA model and the standard FA models leaving out 15% of the sample for independent validation. Predictions are obtained

Table 4.1 Number of total and specific factors for each study, Table 2.1, in Immune System pathway.

Study	total latent factors	specific latent factors
GSE9891	6	5
GSE20565	7	6
GSE26712	10	9
TCGA	9	8

as

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

where $\hat{\Lambda}_s^{\text{MFA}}$ are the specific factor loadings estimated by MFA and $\hat{\Lambda}_s^{\text{FA}}$ are the factor loadings estimated by FA. We evaluate the mean squared error of prediction, MSE $= \frac{1}{n}\sum_{s=1}^{S}\sum_{i=1}^{n}(x_{is}-\hat{x}_{is})^2$, using 15% of the sample in each study, set aside for test-set to avoid overly optimistic conclusions. The MSE is 1.5% smaller for MFA than for FA. The MFA is imposing a strong equality constraint on the first factor. This simple check illustrates how this constraint allows to borrow strength across studies in the estimation of the factor loadings, in such a way that the predictive ability in independent observations is slightly improved.

Next, we focus on the analysis of the estimated factor loadings themselves. The heatmap in Figure 4.1 depicts the estimates of the factor loadings, both common (highlighted in the black rectangle) and specific ones.

To interpret the biological meaning of the common factor, we apply Gene Set Enrichment Analysis (GSEA) for determining whether a given gene set is significantly enriched in a list of gene markers (or significant pathways) ranked by their correlation with a phenotype of interest (Mootha et al., 2003; Subramanian et al., 2005). We consider all the three sub-pathways in the Immune System pathway In order to do that, the package `Rtopper` in `R` in `Bioconductor` is used, following the method illustrated in Tyekucheva et al. (2011).

The resulting analysis shows that the common factor is significantly enriched by the AI sub-pathway, suggesting that genuine biological signal may have been identified.

Further, we consider the cross-study pairwise correlations of the loadings involving the study-specific factors shown in Figure 4.2. Three of the specific factors of the GSE9891 study are strongly correlated with three corresponding factors in the GSE20565 study. Absolute correlations range from 0.66 to 0.81.

Studies GSE9891 and GSE20565 use the same platform, Affy U133 Plus2.0, prompting the conjecture that the three stronger correlations observed may be related to technological rather than biological variation. To further probe this possibility, at least within the Immune

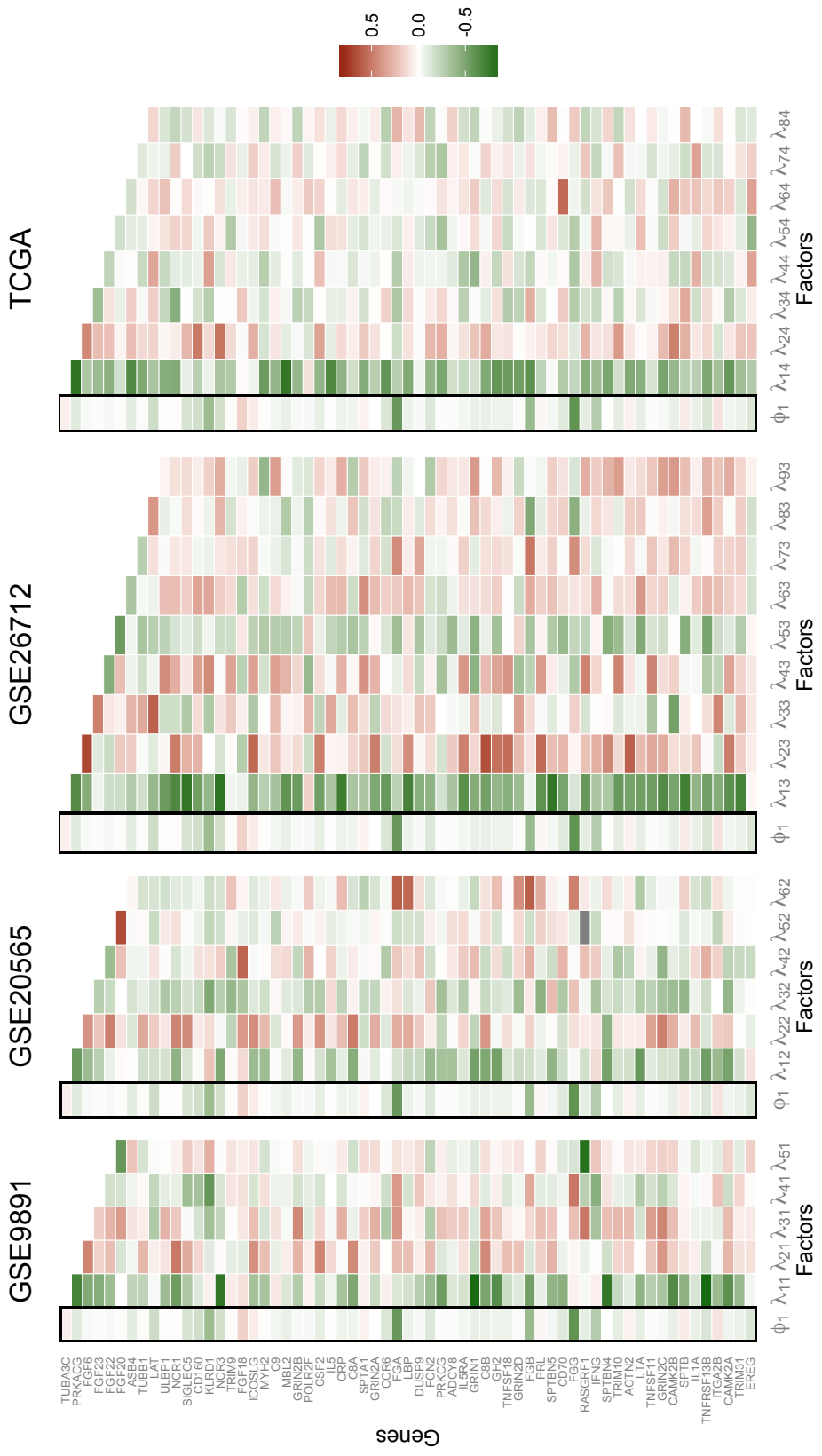

Figure 4.1 Immune System pathway: Heatmap of the estimated factor loadings obtained with MFA, both common (black rectangle) and specific ones.

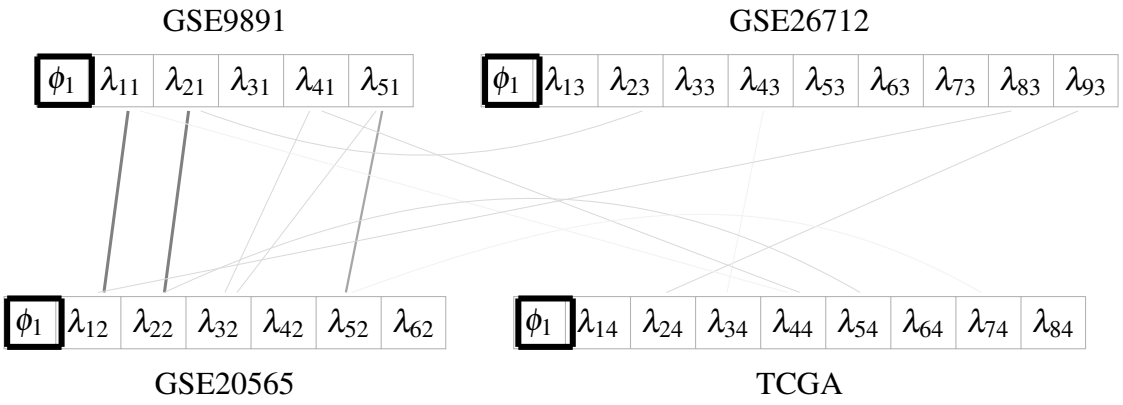

Figure 4.2 Graphical representation of the correlation of the specific factor loadings obtained with the MFA. Darker grey lines correspond to higher correlations. Correlations smaller than .25 are not shown.

System pathway, we analyze studies GSE9891 and GSE20565 separately from the other two using MFA. The AIC chooses a model with $K = 4$ with a total of six latent factors in the former study and seven in the latter one. The results are shown in Figure 4.3.

Figure 4.3 Immune System pathway: Heatmap of the factor loadings obtained with the MFA performed in the two studies, first two rows in Table 2.1.

Again, we performed the GSEA on the estimated factor loadings. The resulting analysis shows that the first common factor is related to the AI system pathway, as was the only

common factor shared between the four studies in the earlier analysis. In fact, the common factor of the four studies analysis is highly correlated with the first common factors of the two-study analysis, $r = 0.60$. The three 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 induced by the specific platform used.

Some checking on the impact of the order-dependence induced by the block lower triangular structure assumed for Ω_s to address the identifiability problem was carried out. In particular, the same analysis was repeated after permuting the genes. Despite some discrepancies, the final conclusion is exactly the same. Namely, one common factor is significantly enriched only with the AI system sub-pathway.

4.2 DNA-repair

We continue to validate our proposed procedures by analyzing the DNA-repair pathway.

As in the previous section the analysis is focused on common genes across studies and included in the sub-pathways "Base Excixion Repair" (BER), "DNA damage" (DD), "Double-Strand Break Repair" (DBR), "Nucleotide Exicixion" (NE), "Fanconi Anemia pathway" (FA) and "Mismatch Repair" (MR) obtained from `reactome.org` belonging to the DNA-repair pathway.

We assess the total latent dimensions through the standard FA techniques, the number of common factors across the studies through the AIC criterion and, consequently, the number of specific factors. The results lead again to one common factor shared across studies. The total latent dimensions and the number of specific factor are showed in Table 4.2.

Table 4.2 Number of total and specific factors for each study in Table 2.1 for the DNA repair pathway.

Study	total latent factors	specific latent factors
GSE9891	6	5
GSE20565	7	6
GSE26712	11	10
TCGA	10	9

Once again, to demonstrate the performance of our method we predict the mean squared error leaving out the 15% of the sample for independent validation. The MSE based on MFA for the four studies is 0.64 whereas the MSE based on FA 0.65. Also in this application, the MFA slightly outperforms the standard FA.

Next, we focus on the analysis of the data presented in Table 2.1 and the estimated factor loadings. The heatmap in Figure 4.4 represents the estimates of the factor loadings, both common (highlighted in the black rectangle) and specific.

To interpret the biological meaning of the common factor, we apply again the GSEA. We considered all the six sub-pathways in the DNA-repair pathway to compare our gene set using the package `Rtopper`. The results lead us to conclude that our gene set is related with the only sub-pathway BER, so that also in this case the biological signal may have been identified.

However, also in the DNA-repair application some specific factor loadings of the GSE9891 study exhibit a similar pattern to that of specific loadings of the GSE20565 study. An example is provided in Figure 4.4 by the second column of the heatmap relative to GSE9891 and the second column of the heatmap relative to GSE20565. They are strongly correlated (> 0.6), while the correlations with other specific factor loadings in GSE26712 and TCGA are less than 0.3. The reason can be related to artifactual source of variations since the studies are measured in the same platform, i.e. Affy U133 Plus2.0. For a better understanding, we analyze the two studies, GSE9891 and GSE20565, separately.

Considering only two studies, as in Figure 4.5, the AIC chooses a model with $K = 5$ with a total of six latent factors in the former study and seven in the latter one. The commonality maybe come from the platform on which gene expressions are measured. In order to further investigate this point, we perform again the GSEA on the estimated factor loadings for the model with $K = 5$.

The resulting analysis shows that the first and the second common factors are not related to the sub-pathways described before: they are not capturing the biological features but, for example, the platform where the genes are measured. The others three common factors are related to these sub-pathways:

- the third common factor is related to DD and MR;
- the fourth common factor is related to MR;
- the fifth common factor is related to DD.

The common factors in the two studies capture the biological signal of two relevant pathways, the DD and MR.

In the DNA-repair pathway, subsets of genes contribute to diverse sub-pathway, in the BER we found large amount (more than the 50%) of genes in common with the DD and MR. Moreover, the common factor of the four studies analysis is highly correlated with the fourth and the fifth common factors of the two-study analysis, $r = 0.52$ and $r = 0.60$.

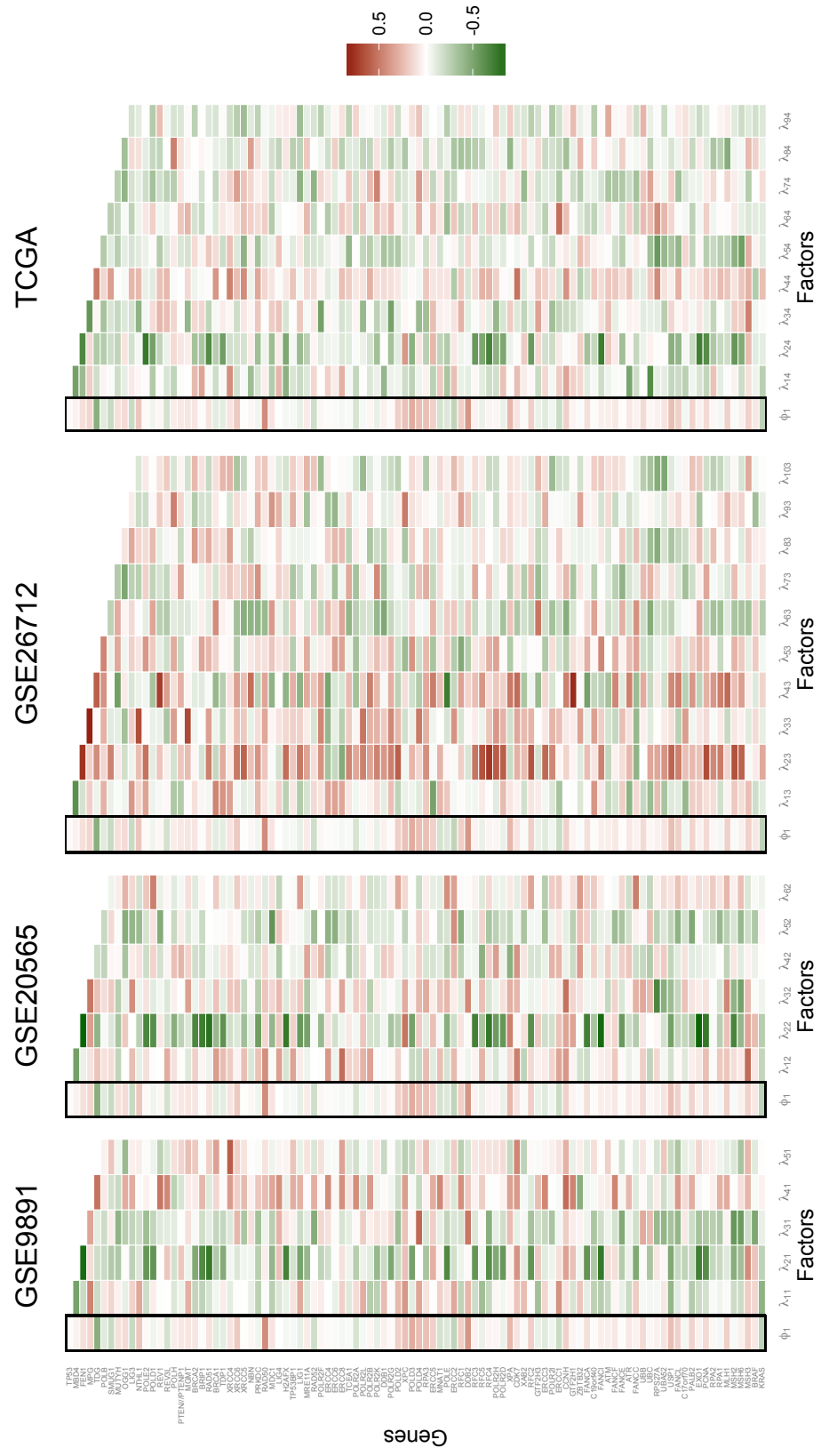

Figure 4.4 DNA Repair pathway: Heatmap of the factor loadings obtained with the MFA in 2.1.

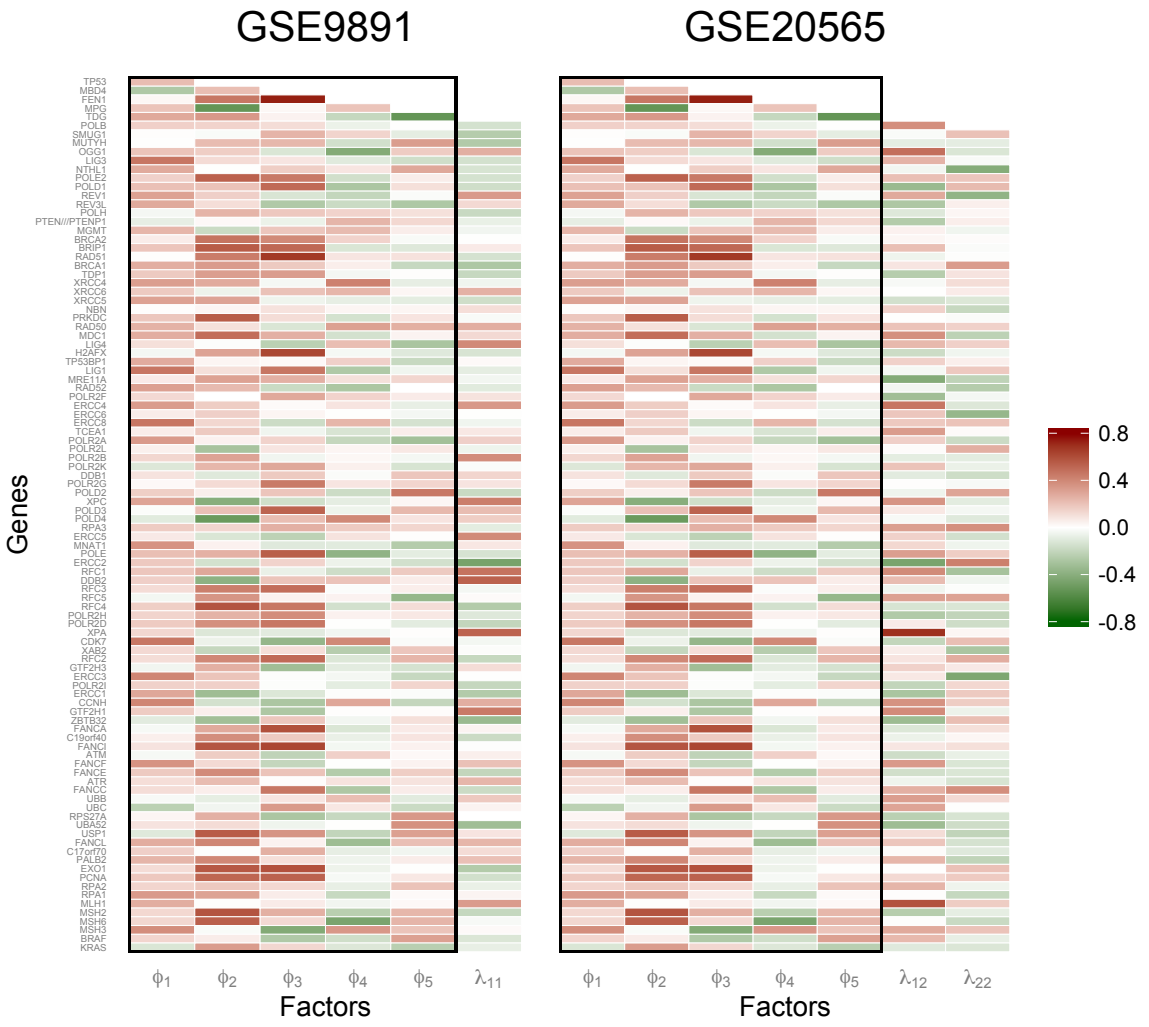

Figure 4.5 DNA Repair pathway: Heatmap of the factor loadings obtained with the MFA performed in the two studies, first two rows in Table 2.1.

4.3 Discussion

The method presented in the last two chapters is able to extract information from multiple studies.

Integration of different studies relies critically on specific methods of analysis. As already mentioned, it is crucial in these kind of analysis to separate the two kind of information, that related to the common part shared among studies and that described by the differences and specificity of each study. Indeed, the most critical step in cross-study analysis is to identify biological factors that are reproducible across studies and to remove idiosyncratic variation that lacks cross-study reproducibility.

The method is simple and it is based on a generalized version of FA able to handle multiple studies simultaneously. Indeed, we develop dimension reduction tools that allow for joint analysis of multiple studies and capture the two types of information.

The simulations of Chapter 3 and real data analysis of this chapter suggest that the model is worthy of serious consideration.

The MFA model can be applied to many settings when the aim is to isolate commonalities and differences across different groups, population or studies. Differently from the focus of the thesis, there might be other applications where the goal is to capture some study-specific features of interest and, instead, remove some common factors shared among studies. Other applications may focus on capturing both common and specific factors, without removing any of them. Indeed, the approach illustrated could be extended in several directions. It will be relevant for a wide variety of genomic platforms (e.g. microarrays, RNA-seq, SNPs, proteomics, metabolomics, epigenomics), as well as datasets in other fields of biomedical research, such as those generated by exposome studies or Electronic Medical Record (EMR).

Overall, this analysis illustrates the strength of this method, namely its ability to capture biological signal and to isolate the source of variation coming, for example, from the different platforms by which gene expressions are measured.

Chapter 5

Sparse Bayesian multi-study factor model

In this chapter we focus on the sparse setting where the number of variables is larger than the number of subjects. In Section 5.1 a general description of the sparsity approach is given, and how the problem of $p > n$ has been tackled by means of Bayesian methods in the literature. In Section 5.2 we consider the sparsity approach in MFA model through the spike and slab prior introduced by George and McCulloch (1993). The details of a Gibbs sampling are provided to sample from the posterior distribution. The analysis of some simulated data are presented in Section 5.3 to study the properties of the proposed approach. In Section 5.4 the methodology is applied to the OC data, while finally Section 5.5 provides a discussion.

5.1 A brief introduction to the sparse setting

In recent years many statistical applications involve high-dimensional data. Here, high dimension refers to settings with $p > n$, with singular covariance matrix for the s^{th} study Σ_s, preventing the application of maximum likelihood estimation. Such settings arise crucially in genomic applications where the number of genes are greater than the number of subjects.

FA models are still used in such settings, but some sort of regularization is required (Carvalho et al., 2008; Engelhardt and Stephens, 2010; Lopes and West, 2004). The relation between the latent factors and the observed variables is described by the coefficient of the factor loadings matrix Ω_s. In the applications where there are more components than subjects it is crucial to include in the formalization some strong regularization on the elements of the factor loadings matrix. In the statistical literature, in order to regularize the factor loadings, priors or penalties are used to induce sparsity, namely selection of some features. Sparsity

removes some entries of the loading matrix, thus assuming that only some of the variables are associated to some of the latent factors. Indeed, in the genomic context, sparsity implies that if we take in consideration a set of genes, only a few are interacting and can reveal relevant biological factors (Tegner et al., 2003). Sparse factor loadings can be used to identify clusters of genes and interpret them as classes of interacting genes (Lucas et al., 2010; Pournara and Wernisch, 2007). This situation is visualized in Figure 5.1.

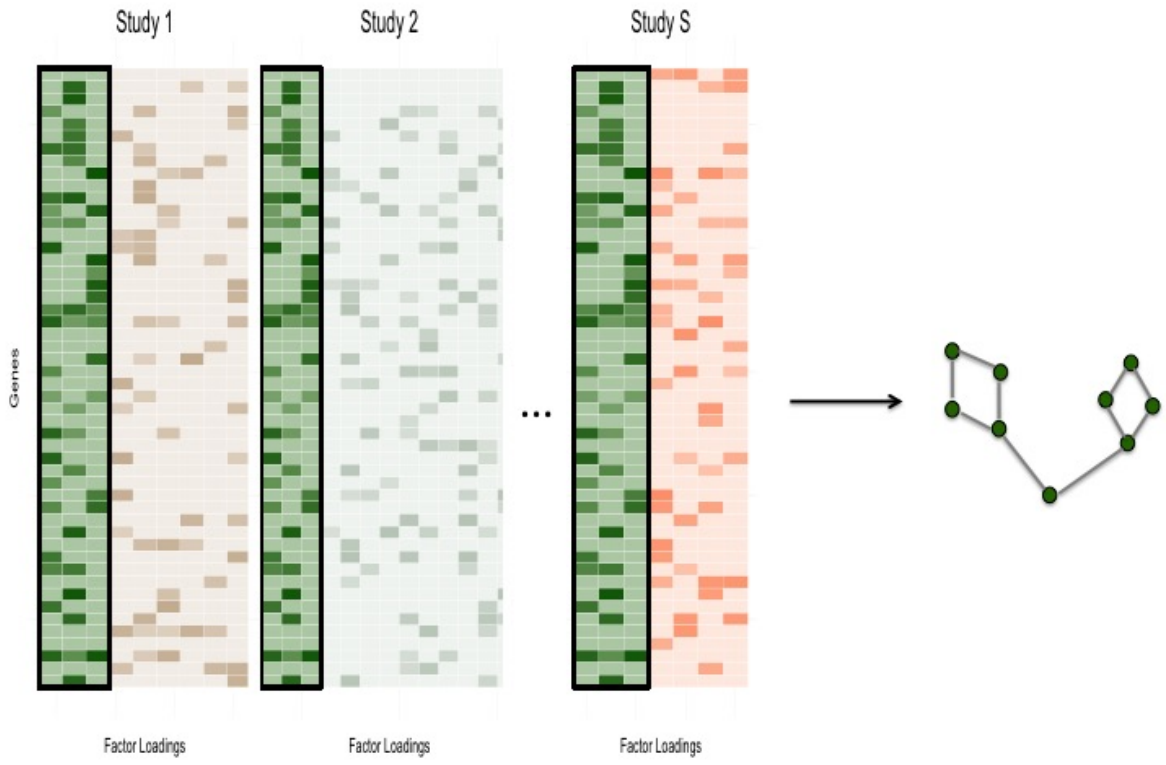

Figure 5.1 Visualization of MFA under the sparsity assumption

As represented in Chapter 3, the MFA has two different matrices of loadings in each study, those related to the common factor and then the study-specific part. In the sparsity context the objective remains the same, i.e. to capture the common biological features and isolating the artifactual and biological variation. Here, an important additional complication is due to the high-dimension involved.

In order to choose the most suitable regularization to induce sparsity in the MFA model, we here summarize the theory and the application of sparse FA developed in the statistical literature.

Sparsity in factor model has been investigated through regularization via L_1-type penalties in frequentist analyses (Witten et al., 2009; Zou and Hastie, 2005). In contrast to the frequentist approach, Bayesian methods model sparsity through the introduction of shrinkage priors.

In the Bayesian approach to sparsity, three main priors has been used and developed. These three approaches have been considered here in turn, in order to choose that most suitable to extension to the MFA setting.

The first one is the Bayesian lasso prior, introduced by Park and Casella (2008) and developed in high-dimensional linear models by Hans (2009).

Based on the Lasso penalty of Tibshirani (1996), the Bayesian lasso prior is a conditional Laplace prior for the loadings

$$\lambda_{pjs}|\psi_{ps} \sim \frac{\tau}{2\sqrt{\psi_{ps}}} e^{-\tau|\lambda_{pjs}|/\sqrt{\psi_{ps}}},$$

where ψ_{ps} is the diagonal element of the covariance error matrix and $\tau > 0$ is the scale hyper parameter. In this modeling setting, the posterior mode of λ_{pjs} is the lasso estimate with the penalty equal to $2\tau\psi_{ps}$, which regulates the amount of shrinkage. Posterior inference is developed via Gibbs sampling.

The major limitation of this approach lies on the lack of unimodality for the posterior distribution of λ_{pjs}. Indeed, the posterior distribution of the factor loadings could present a bimodal trend and this problem leads to point estimates less meaningful (Park and Casella, 2008). This problem can also occur considering the prior of the error variance ψ_{ps} as proper. In some preliminary experiments with the data analyzed in the previous chapter, a clear suggestion of bimodality occurred for the posterior distribution of the elements of Φ, see Figure 5.2.

The second sparse Bayesian approach is taken from the paper of Bhattacharya and Dunson (2011). They develop a multiplicative gamma shrinkage prior on the factor loadings, with increasing shrinkage as the column index increases.

Their model, called the *Sparse Bayesian infinite factor model*, uses the prior

$$\lambda_{pjs}|\zeta_{pjs},\tau_j \sim N(0,\zeta_{pjs}^{-1}\tau_j^{-1}), \quad \zeta_{pjs} \sim Ga(\eta/2,\eta/2),\ \tau_j = \prod_{l=1}^{j}\delta_l,$$

$\delta_1 \sim Ga(a_1,1),\ \delta_l \sim Ga(a_2,1),\ \ l \geq 2$, where $\delta_l (l = 1,\ldots,\infty)$ are independent, τ_j is a global shrinkage parameter for the j^{th} column and ζ_{pjs} are local shrinkage parameters for the elements in the j^{th} column. As Bhattacharya and Dunson (2011) report, the ζ_{pjs} are stochastically increasing only with $a_2 > 1$. So, more shrinkage is obtained as the column

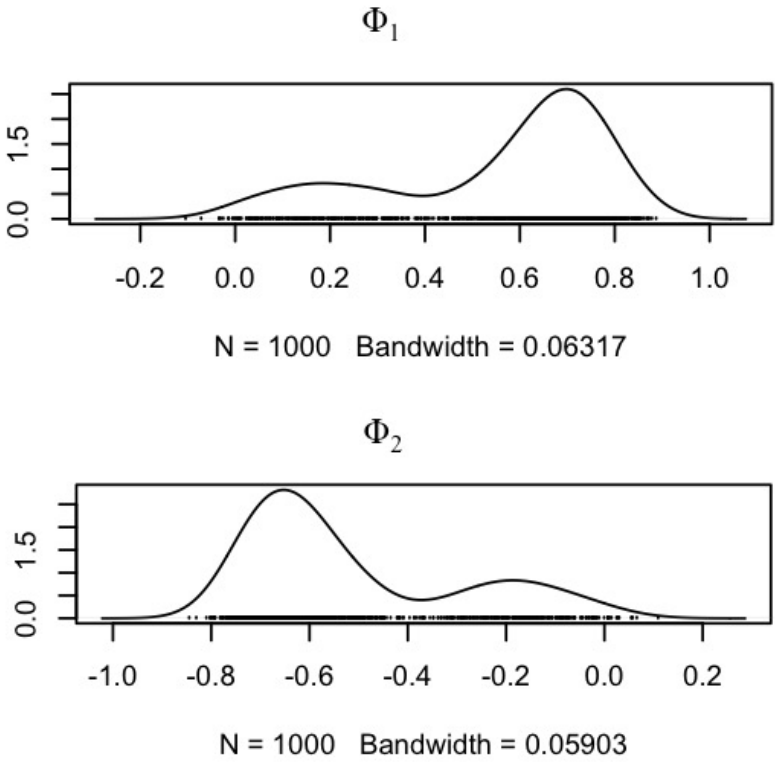

Figure 5.2 Density plot of the full conditional of the first two common factor loadings with the lasso prior

index increases. The idea is rather suitable to high dimensions, where if more factors are added it is crucial to consider an increment of the shrinkage. They develop a Gibbs sampling and study the inferential implications of the prior. In their formalization, the identification problem is not considered since they focus on the estimation of the covariance matrix, and not on the estimation of the factor loadings. An algorithm is provided where the number of factors are chosen adaptively. They adapt with probability

$$p(t) = \exp(\alpha_0 + \alpha_1 t)$$

at t^{ht} iteration, with α_0 and α_1 so that at the beginning adaptation arises every 10 iterations, and then it decreases. They provide a quite useful `Matlab` code, that we ported to `R`.

We use this code to apply the Sparse Bayesian infinite factor model to the ovarian data, on the DNA-repair pathway presented in Section 4.2.

We take in consideration the data showed in Table 2.1 and we fit the Sparse Bayesian infinite factor model in each data set since this method does not consider multiple studies.

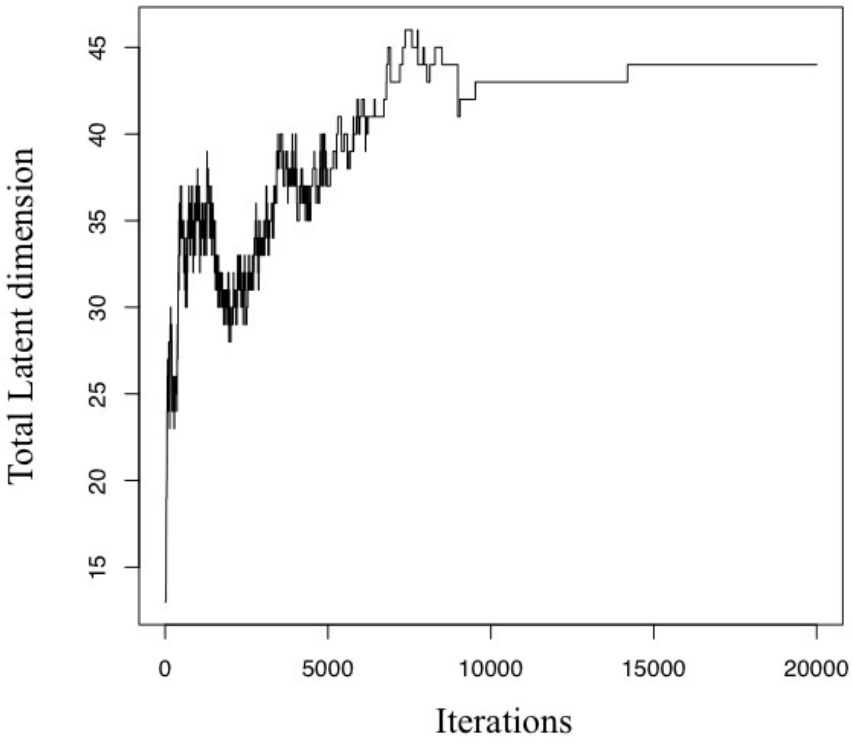

Figure 5.3 Number of selected latent factors in the Bhattacharya and Dunson (2011) approach for the TCGA data on ovarian cancer

In Figure 5.3 the results of the application to the TCGA data set are represented. The figure shows that after 15000 iterations the choice of the total latent dimension T_s is more than 40 latent factors, much higher than what found in Chapter 4. In other words, this method tends to overestimate the number of latent dimension at least in this data set with $p = 100$.

The last approach adopts the spike and slab prior, developed by George and McCulloch (1993, 1997) in the context of variable selection for linear regression.

The prior is a normal mixture defined by the random variable $\delta^s_{pj} = 1$ or 0

$$\lambda_{pjs}|\delta^s_{pj} \sim (1-\delta^s_{pj})N(0,\tau^2_{pjs}) + \delta^s_{pj}N(0,c^2_{\lambda_{pjs}}\tau^2_{pjs}),$$

and $P(\delta^s_{pj} = 1) = p^s_{pj}$.

When $\delta^s_{pj} = 0$, the prior $\lambda_{pjs} \sim N(0,\tau^2_{pjs})$, with τ^2_{pjs} quite small, is nearly a *spike* implying that the related factor loadings are concentrated around zero. Instead, when $\delta^s_{pj} = 1$, $\lambda_{pjs} \sim N(0,c^2_{\lambda_{pjs}}\tau^2_{pjs})$, and for large values of $c^2_{\lambda_{pjs}}$, the prior is a flat distribution, called the *slab*.

George and McCulloch (1993) report some indications for the choice of τ_{pjs}, $c_{\lambda_{pjs}}$ and δ^s_{pj}, and a Gibbs sampling algorithm is provided.

When prior knowledge is absent, as reported by Gelman et al. (2014), it is possible to set $p^s_{pj} = p^s$ and so consider in the formalization an hyperprior $p^s \sim \text{Beta}(\alpha_\text{p}, \beta_\text{p})$. Particular attention should be given to the choice of hyperparameters, especially in the context of $p > n$, where informative choices are mandatory.

This approach has been successfully applied to genetic data by Carvalho et al. (2008); West (2003). This is the approach followed here. In the rest of this chapter, it will be extended to the MFA model.

5.2 Model and prior specification

As showed in Section 3, the MFA assumes that each observed variable in each study is decomposed as follow

$$\mathbf{x}_{is} = \Phi \mathbf{f}_i + \Lambda_s \mathbf{l}_{is} + \mathbf{e}_{is}\,.$$

In order to follow the Bayesian approach, we assume a prior distribution on each element of the factor loading matrices.

A sparsity mixture prior defined by the random variable δ_{pk} is assigned to each element $\phi_{pk}, p = 1,\ldots,P,\ k = 1,\ldots,K$, of the common factor loadings matrix Φ:

$$\phi_{pk} \mid \delta_{pk} \sim (1-\delta_{pk})N(0,\zeta^2_{pk}) + \delta_{pk}N(0, c^2_{\phi_{pk}}\zeta^2_{pk}),$$

and

$$P(\delta_{pk} = 1) = 1 - P(\delta_{pk} = 0) = p_{pk}.$$

If we denote the p^{th} row of Φ by ϕ^t_p, then the ϕ_ps have the prior distributions

$$\phi_p \mid \delta_p \sim N_p(\mathbf{0}, \mathbf{D}_{p\delta}),$$

where $\mathbf{D}_{p\delta} = \text{ diag}(a_{p1}\zeta^2_{p1},\ldots,a_{pK}\zeta^2_{pK})$ and $a_{pk} = 1$ if $\delta_{pk} = 0$ and $a_{pk} = c^2_{\phi_{pk}}$ if $\delta_{pk} = 1$.

A beta prior distribution is used for p_{pk}. When prior information on the expected level of sparsity is scarce, such choice is recommendable.

A mixture prior defined by the variable δ^s_{pj} is assigned to each element $\lambda_{pjs}, p = 1,\ldots,P\ j = 1,\ldots,J_s$ and $s = 1,\ldots,S$ of the specific factor loading matrix Λ_s:

$$\lambda_{pjs} | \delta^s_{pj} \sim (1-\delta^s_{pj})N(0,\tau^2_{pjs}) + \delta^s_{pj}N(0, c^2_{\lambda_{pjs}}\tau^2_{pjs}),$$

and

$$P(\delta_{pj}^s = 1) = 1 - P(\delta_{pj}^s = 0) = p_{pj}^s.$$

As for the δ_{pk}, a beta prior is used for p_{pj}^s.

As before, if we denote the p^{th} row of Λ_s by $\lambda_{ps}^\top$, so that we write the prior distributions

$$\lambda_{ps} \mid \delta_p^s \sim N_p(\mathbf{0}, \mathbf{D}_{p\delta^s}),$$

where $\mathbf{D}_{p\delta^s} = \text{diag}(a_{p1s}\tau_{p1s}^2, \ldots, a_{pJ_ss}\tau_{pJ_ss}^2)$, $a_{pjs} = 1$ if $\delta_{pj}^s = 0$ and $a_{pjs} = c_{\lambda_{pjs}}^2$ if $\delta_{pj}^s = 1$.

The priors for the factor loadings, both common and specific, belong to the class of absolutely continuous spike and slab priors where ζ_{pk}^2 and τ_{pjs}^2 are small constants , thus representing the spike of the common and specific factors, respectively, and so the distributions are concentrated on zero. Instead, $c_{\phi_{pk}}^2$ and $c_{\lambda_{pjs}}^2$ are large constants ($\gg 1$), thus representing the slab part of the mixture of the common and specific factor loadings.

For each of the erratic variance $\psi_{ps},\ p = 1, \ldots, P$ we assume an inverse gamma prior. This choice comes from the standard FA (Bhattacharya and Dunson, 2011; Lopes and West, 2004). In details, the ψ_{ps} are formalized as

$$\psi_{ps} \sim \text{Ga}(\text{a}_\psi, \text{b}_\psi).$$

As we reported in Section 3.2.2, MFA must be further constrained to define a model free from identification problems. As applied in the classical approach of Chapter 3, we use here the block lower triangular matrix constraint and then we adjust the sign of each column of the loading matrices by post processing.

5.2.1 Posterior Computation

We propose a Gibbs sampler for posterior computation, through data augmentation (Gelman et al., 2014; Tanner and Wong, 1987). In the data augmentation, first we proceed conditionally on the latent variables, common and specific factors. The algorithm is given by the following steps, which are performed cyclically

- STEP 1. Sample from the conditional posterior of $l_{is},\ i = 1, \ldots, n_s,\ s = 1, \ldots, S$, the specific factors in each study.
- STEP 2. Sample from the conditional posterior of $f_i,\ i = 1, \ldots n_s$, the common latent factors.
- STEP 3. Sample from the specific factor loadings, $\lambda_{ps},\ p = 1, \ldots, P,\ s = 1, \ldots, S$.

- STEP 4. Sample from the common factor loadings, $\phi_p,\ p = 1,\ldots,P$.
- STEP 5. Update $\delta^s_{pj},\ j = 1,\ldots,J_s$, the Bernoulli random variables for the choice of spike or the slab for the specific factors.
- STEP 6. Update $\delta_{pk},\ k = 1,\ldots,K$, the Bernoulli random variables for the choice of spike or the slab for the common factors.
- STEP 7. Update p^s_{pj} the probability for the spike and slab in the specific factor.
- STEP 8. Update p_{pj} the probability for the spike and slab in the common factors.
- STEP 9. Update Ψ_s, the covariance matrix of the error terms.

For the sampling in each step that we use the complete log-likelihood, as done for the ECM algorithm. Indeed, the data augmentation approach leads to an algorithm similar to the EM algorithm, though here the aim is sampling from the posterior distribution.

The likelihood function has the usual form

$$|\Psi_s|^{-n_s/2}\prod_{s=1}^{S}\prod_{i=1}^{n_s}\exp\left\{-\frac{[\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}]^\top\Psi_s^{-1}[\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}]}{2}\right\}. \tag{5.1}$$

The (5.1) is conditional on the latent variables

$$\mathbf{x}_{is}|\mathbf{f}_i,\mathbf{l}_{is}\sim N(\Phi\mathbf{f}_i+\Lambda_s\mathbf{l}_{is},\Psi_s),$$

and

$$\mathbf{f}_i\sim N(\mathbf{0},\mathbf{I}_K),\qquad \mathbf{l}_{is}\sim N(\mathbf{0},\mathbf{I}_{J_s}).$$

In what follows, we provide the details for the full conditionals employed for the Gibbs sampling.

STEP 1.

The full conditional of $\mathbf{l}_{is}$, with the prior $\mathbf{l}_{is}\sim N(\mathbf{0},\mathbf{I}_{J_s})$ is

$$\pi(\mathbf{l}_{is}|-) \;=\; |\Psi_s|^{-n_s/2}\prod_{s=1}^{S}\prod_{i=1}^{n_s}\exp\left\{-\frac{[\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}]^\top\Psi_s^{-1}[\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}]+\mathbf{l}_{is}^\top\mathbf{l}_{is}}{2}\right\}.$$

Expanding terms in the exponent leads to

$$\begin{aligned}
& -\left[\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}\right]^\top\Psi_s^{-1}\left[\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}\right]+\mathbf{l}_{is}^\top\mathbf{l}_{is}= \\
= \;& \mathbf{x}_{is}^\top\Psi_s^{-1}\mathbf{x}_{is}+\mathbf{f}_i^\top\Phi^\top\Psi_s^{-1}\Phi\mathbf{f}_i+\mathbf{l}_{is}^\top\Lambda_s^\top\Psi_s^{-1}\Lambda_s\mathbf{l}_{is} \\
- \;& 2\mathbf{x}_{is}^\top\Psi_s^{-1}\Phi\mathbf{f}_i-2\mathbf{x}^\top\Psi_s^{-1}\Lambda_s\mathbf{l}_{is}+2\mathbf{f}_i^\top\Phi^\top\Psi_s^{-1}\Lambda_s\mathbf{l}_{is}+\mathbf{l}_{is}^\top\mathbf{l}_{is} \\
= \;& \mathbf{l}_{is}^\top(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi^{-1}\Lambda_s)\mathbf{l}_{is}-2\mathbf{l}_{is}^\top(\Lambda_s^\top\Psi^{-1}\mathbf{x}_{is}-\Lambda_s^\top\Psi^{-1}\Phi\mathbf{f}_i) \\
+ \;& \mathbf{x}_{is}^\top\Psi^{-1}\mathbf{x}_{is}+\mathbf{f}_i^\top\Phi^\top\Psi^{-1}\Phi\mathbf{f}_i-2\mathbf{x}_{is}^\top\Psi^{-1}\Phi\mathbf{f}_i,
\end{aligned}$$

so the mean of $\mathbf{l}_{is}$ is

$$\bar{\mathbf{l}}_s=(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi^{-1}\Lambda_s)^{-1}(\Lambda_s^\top\Psi^{-1}\mathbf{x}_{is}-\Lambda_s^\top\Psi^{-1}\Phi\mathbf{f}_i).$$

We can then write

$$\begin{aligned}
& (\mathbf{l}_{is}-\bar{\mathbf{l}}_s)^\top(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi^{-1}\Lambda_s)^{-1}(\mathbf{l}_{is}-\bar{\mathbf{l}}_s)=\mathbf{l}_{is}^\top(\mathbf{I}_{J_s}+\Lambda_s^t\Psi^{-1}\Lambda_s)^{-1}\mathbf{l}_{is} \\
- \;& 2\mathbf{l}_{is}^\top(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi^{-1}\Lambda_s)^{-1}\bar{\mathbf{l}}_s+\bar{\mathbf{l}}_s^\top(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi^{-1}\Lambda_s)^{-1}\bar{\mathbf{l}}_s,
\end{aligned}$$

adding and subtracting $\bar{\mathbf{l}}_s^\top(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi^{-1}\Lambda_s)^{-1}\bar{\mathbf{l}}_s$ we find

$$\begin{aligned}
& \left[\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}\right]^\top\Psi_s^{-1}\left[\mathbf{x}_{is}-\Phi\mathbf{f}_i-\Lambda_s\mathbf{l}_{is}\right]+\mathbf{l}_{is}^\top\mathbf{l}_{is}=\mathbf{l}_{is}^\top(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi_s^{-1}\Lambda_s)^{-1}\mathbf{l}_{is} \\
- \;& 2\mathbf{l}_{is}^\top\mathbf{l}_{is}^\top(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi_s^{-1}\Lambda_s)^{-1}\bar{\mathbf{l}}_s+\bar{\mathbf{l}}_s^\top(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi_s^{-1}\Lambda_s)^{-1}\bar{\mathbf{l}}_s-\bar{\mathbf{l}}_s^\top(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi_s^{-1}\Lambda_s)^{-1}\bar{\mathbf{l}}_s \\
+ \;& \mathbf{x}_{is}^\top\Psi_s^{-1}\mathbf{x}_{is}+\mathbf{f}_i^\top\Phi^\top\Psi_s^{-1}\Phi\mathbf{f}_i-2\mathbf{x}_{is}^\top\Psi_s^{-1}\Phi\mathbf{f}_i \\
= \;& (\mathbf{l}_{is}-\bar{\mathbf{l}}_s)^\top(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi_s^{-1}\Lambda_s)^{-1}(\mathbf{l}_{is}-\bar{\mathbf{l}}_s)-\bar{\mathbf{l}}_s^\top(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi_s^{-1}\Lambda_s)^{-1}\bar{\mathbf{l}}_s \\
+ \;& \mathbf{x}_{is}^\top\Psi_s^{-1}\mathbf{x}_{is}+\mathbf{f}_i^\top\Phi^\top\Psi_s^{-1}\Phi\mathbf{f}_i-2\mathbf{x}_{is}^\top\Psi_s^{-1}\Phi\mathbf{f}_i.
\end{aligned}$$

Dropping the constant

$$\pi(\mathbf{l}_{is}|-)=\exp\left\{-\frac{1}{2}(\mathbf{l}_{is}-\bar{\mathbf{l}}_s)^\top(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi_s^{-1}\Lambda_s)^{-1}(\mathbf{l}_{is}-\bar{\mathbf{l}}_s)\right\},$$

and finally the full conditional for $\mathbf{l}_s$ is

$$\pi(\mathbf{l}_s|\dots)\sim N\left\{(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi_s^{-1}\Lambda_s)^{-1}(\Lambda_s^\top\Psi_s^{-1}\mathbf{x}_{is}-\Lambda_s^\top\Psi_s^{-1}\Phi\mathbf{f}_i),(\mathbf{I}_{J_s}+\Lambda_s^\top\Psi_s^{-1}\Lambda_s)^{-1}\right\}.$$

STEP 2.

We compute the full conditional for $\mathbf{f}_i$, the common latent factors:

$$\pi(\mathbf{f}_i|-) \quad = \quad |\Psi_s|^{-n_s/2} \prod_{s=1}^{S} \prod_{i=1}^{n_s} \exp\left\{ -\frac{\left[\mathbf{x}_{is} - \Phi\mathbf{f}_i - \Lambda_s\mathbf{l}_{is}\right]^\top \Psi_s^{-1} \left[\mathbf{x}_{is} - \Phi\mathbf{f}_i - \Lambda_s\mathbf{l}_{is}\right] + \mathbf{f}_i^\top \mathbf{f}_i}{2} \right\}.$$

Expanding terms in the exponent leads to

$$\begin{aligned}
& \left[\mathbf{x}_{is} - \Phi\mathbf{f}_i - \Lambda_s\mathbf{l}_{is}\right]^\top \Psi_s^{-1} \left[\mathbf{x}_{is} - \Phi\mathbf{f}_i - \Lambda_s\mathbf{l}_{is}\right] + \mathbf{f}^\top\mathbf{f} = \\
= \quad & \mathbf{x}_{is}^\top\Psi_s^{-1}\mathbf{x}_{is} + \mathbf{f}_i^\top\Phi^\top\Psi_s^{-1}\Phi\mathbf{f}_i + \mathbf{l}_{is}^\top\Lambda_s^\top\Psi_s^{-1}\Lambda_s\mathbf{l}_{is} \\
- \quad & 2\mathbf{x}_{is}^\top\Psi_s^{-1}\Phi\mathbf{f}_i - 2\mathbf{x}^\top\Psi_s^{-1}\Lambda_s\mathbf{l}_{is} + 2\mathbf{f}_i^\top\Phi^t\Psi_s^{-1}\Lambda_s\mathbf{l}_{is} + \mathbf{f}_i^\top\mathbf{f}_i \\
= \quad & \mathbf{f}_i^\top(\mathbf{I}_K + \Phi^\top\Psi_s^{-1}\Phi^\top)\mathbf{f}_i - 2\mathbf{f}_i^\top(\Phi^\top\Psi_s^{-1}\mathbf{x}_{is} - \Phi^\top\Psi_s^{-1}\Lambda_s\mathbf{l}_{is}) \\
+ \quad & \mathbf{x}_{is}^\top\Psi_s^{-1}\mathbf{x}_{is} + \mathbf{l}_{is}^\top\Lambda_s^\top\Psi_s^{-1}\Lambda_s\mathbf{l}_{is} - 2\mathbf{x}_{is}^\top\Psi_s^{-1}\Lambda_s\mathbf{l}_{is}.
\end{aligned}$$

As showed before for the specific latent variables, the full conditional for $\mathbf{f}_i$ is

$$\pi(\mathbf{f}_i|-) \sim N((\mathbf{I}_K + \Phi^\top\Psi_s^{-1}\Phi)^{-1}(\Phi^\top\Psi_s^{-1}\mathbf{x}_{is} - \Phi^\top\Psi_s^{-1}\Lambda_s\mathbf{l}_{is}), (\mathbf{I}_K + \Phi^\top\Psi_s^{-1}\Phi)^{-1}).$$

STEP 3.

The λ_{ps}s have independent conditionally conjugate posteriors,

$$\pi(\lambda_{ps}|-) \sim N\left\{(\mathbf{D}_{p\gamma_s}^{-1} + \psi_{ps}^{-2}\mathbf{l}_s^\top\mathbf{l}_s)^{-1}(\mathbf{l}_s^\top\psi_{ps}^{-2}\mathbf{x}_s^{(p)} - \mathbf{l}_s^\top\psi_{ps}^{-2}\phi_p\mathbf{f}), (\mathbf{D}_{p\gamma_s}^{-1} + \psi_{ps}^{-2}\mathbf{l}_s^\top\mathbf{l}_s)^{-1}\right\}$$

where $\mathbf{f} = (f_1, \ldots, f_{n_s})^\top$, $\mathbf{l}_s = (l_{1s}, \ldots, l_{n_s s})^\top$ and $\mathbf{x}_s^{(p)} = (x_{1ps}, \ldots, x_{n_s ps})$.

With more details, the computation of the posterior for Λ_s is

$$\begin{aligned}
\pi(\lambda_{ps}|-) \quad = \quad & \prod_{s=1}^{S} \prod_{p=1}^{P} \left(\psi_{ps}^{-2}\right)^{n_s/2} \exp\left\{ -\frac{\left[\mathbf{x}_s^{(p)} - \mathbf{f}\phi_p - \mathbf{l}_s\lambda_{ps}\right]^\top \left[\mathbf{x}_s^{(p)} - \mathbf{f}\phi_p - \mathbf{l}_s\lambda_{ps}\right] \psi_{ps}^{-2}}{2} \right\} \\
& \exp\left\{ -\frac{\lambda_{ps}^\top\mathbf{D}_{p\gamma_s}^{-1}\lambda_{ps}}{2} \right\}.
\end{aligned}$$

Expanding terms in the exponent leads to

$$\begin{aligned}
&\left[\mathbf{x}_s^{(p)} - \mathbf{f}\phi_p - \mathbf{l}_s\lambda_{ps}\right]^\top \left[\mathbf{x}_s^{(p)} - \mathbf{f}\phi_p - \mathbf{l}_s\lambda_{ps}\right] \psi_{ps}^{-2} + \lambda_{ps}^\top \mathbf{D}_{p\gamma_s}^{-1}\lambda_{ps} = \\
= \;& \psi_{ps}^{-2}\mathbf{x}_s^{(p)^\top}\mathbf{x}_s^{(p)} + \psi_{ps}^{-2}\mathbf{f}^\top\phi_p^\top\phi_p\mathbf{f} + \psi_{ps}^{-2}\mathbf{l}_s^\top\lambda_{ps}^\top\lambda_{ps}\mathbf{l}_s \\
- \;& 2\mathbf{x}_s^{(p)^\top}\psi_{ps}^{-2}\phi_p\mathbf{f} - 2\mathbf{x}_s^{(p)^\top}\psi_{ps}^{-2}\lambda_{ps}\mathbf{l}_s + 2\mathbf{f}^\top\phi_p^\top\psi_{ps}^{-2}\lambda_{ps}\mathbf{l}_s + \lambda_{ps}^\top\mathbf{D}_{p\gamma_s}^{-1}\lambda_{ps} \\
= \;& \lambda_{ps}^\top(\mathbf{D}_{p\gamma_s}^{-1} + \psi_{ps}^{-2}\mathbf{l}_s^\top\mathbf{l}_s)\lambda_{ps} - 2\lambda_{ps}^\top(\mathbf{l}_s^\top\psi_{ps}^{-2}\mathbf{x}_s^{(p)} - \mathbf{l}_s^\top\psi_{ps}^{-2}\phi_p\mathbf{f}) \\
+ \;& \psi_{ps}^{-2}\mathbf{x}_s^{(p)^\top}\mathbf{x}_s^{(p)} + \psi_{ps}^{-2}\mathbf{f}^\top\phi_p^\top\phi_p\mathbf{f} - 2\mathbf{x}_s^{(p)^\top}\psi_{ps}^{-2}\phi_p\mathbf{f}.
\end{aligned}$$

STEP 4.

ϕ_ps have independent conditionally conjugate posteriors:

$$\pi(\phi_p|-) \sim N\left\{\sum_{s=1}^{S}(\mathbf{D}_{p\delta}^{-1} + \psi_{ps}^{-2}\mathbf{f}^\top\mathbf{f})^{-1}(\mathbf{f}^\top\psi_{ps}^{-2}\mathbf{x}_s^{(p)} - \mathbf{f}^\top\psi_{ps}^{-2}bsl_{ps}\mathbf{l}_s), \sum_{s=1}^{S}(\mathbf{D}_{p\delta}^{-1} + \psi_{ps}^{-2}\mathbf{f}^\top\mathbf{f})^{-1}\right\},$$

where $\mathbf{f} = (f_1, \ldots, f_p)^\top$, $\mathbf{l}_s = (l_{1s}, \ldots, l_{ps})^\top$ and $\mathbf{x}_s^{(p)} = (x_{1ps}, \ldots, x_{n_s ps})$.

The calculation of the full conditional for Φ is

$$\begin{aligned}
\pi(\phi_p|-) \;=\; & \prod_{s=1}^{S}\prod_{p=1}^{P}\left(\psi_{ps}^{-2}\right)^{n_s/2}\exp\left\{-\frac{\left[\mathbf{x}_s^{(p)} - \mathbf{f}\phi_p - \mathbf{l}_s\lambda_{ps}\right]^\top\left[\mathbf{x}_s^{(p)} - \mathbf{f}\phi_p - \mathbf{l}_s\lambda_{ps}\right]\psi_{ps}^{-2}}{2}\right\} \\
& \exp\left\{\frac{\phi_p^\top\mathbf{D}_{p\delta}^{-1}\phi_p}{2}\right\}.
\end{aligned}$$

Expanding terms in the exponent leads to

$$\begin{aligned}
&\sum_{s=1}^{S}\left[\mathbf{x}_s^{(p)}-\mathbf{f}\phi_p-\mathbf{l}_s\lambda_{ps}\right]^\top\left[\mathbf{x}_s^{(p)}-\mathbf{f}\phi_p-\mathbf{l}_s\lambda_{ps}\right]\psi_{ps}^{-2}+\phi_p^\top\mathbf{D}_{p\delta}^{-1}\phi_p = \\
= &\sum_{s=1}^{S}\psi_{ps}^{-2}\mathbf{x}_s^{(p)\top}\mathbf{x}_s^{(p)}+\psi_{ps}^{-2}\mathbf{f}^\top\phi_p^\top\phi_p\mathbf{f}+\psi_{ps}^{-2}\mathbf{l}_s^\top\lambda_{ps}^\top\lambda_{ps}\mathbf{l}_s \\
- &\ 2\mathbf{x}_s^{(p)\top}\psi_{ps}^{-2}\phi_p\mathbf{f}-2\mathbf{x}_s^{(p)\top}\psi_{ps}^{-2}\lambda_{ps}\mathbf{l}_s+2\mathbf{f}^\top\phi_p^\top\psi_{ps}^{-2}\lambda_{ps}\mathbf{l}_s+\phi_p^\top\mathbf{D}_{p\delta}^{-1}\phi_p \\
= &\sum_{s=1}^{S}\phi_p^\top(\mathbf{D}_{p\delta}^{-1}+\psi_{ps}^{-2}\mathbf{f}^\top\mathbf{f})\phi_p-2\phi_p^\top(\mathbf{f}^\top\psi_{ps}^{-2}\mathbf{x}_s^{(p)}-\mathbf{f}^\top\psi_{ps}^{-2}\lambda_{ps}\mathbf{l}_s) \\
+ &\ \psi_{ps}^{-2}\mathbf{x}_s^{(p)\top}\mathbf{x}_s^{(p)}+\psi_{ps}^{-2}\mathbf{l}_s^\top\lambda_{ps}^\top\lambda_{ps}\mathbf{l}_s-2\mathbf{x}_s^{(p)\top}\psi_{ps}^{-2}\lambda_{ps}\mathbf{l}_s.
\end{aligned}$$

STEP 5.

We compute the full conditional of δ_{pj}^s for the specific factor, noticing that its distribution does not depend on $\mathbf{x}_s$ (George and McCulloch, 1993). Each distribution is Bernoulli with probability

$$\pi(\delta_{pj}^s=1|-)=\frac{a^s}{a^s+b^s},$$

where

$$\begin{aligned}
a_s &= \pi(\lambda_{pjs}|\ldots,\delta_{pj}^s=1)\,p_{pj}^s \\
b_s &= \pi(\lambda_{pjs}|\ldots,\delta_{pj}^s=0)\,(1-p_{pj}^s).
\end{aligned}$$

STEP 6.

We compute the full conditional of δ_{pk} for the common factors. The result is similar to that of Step 5, and we obtain the Bernoulli distribution with probability

$$\pi(\delta_{pk}=1|-)=\frac{a}{a+b},$$

where

$$\begin{aligned}
a &= \pi(\phi_p|-,\delta_{pk}=1)\,p_{pk} \\
b &= \pi(\phi_p|-,\delta_{pk}=0)\,(1-p_{pk}).
\end{aligned}$$

STEP 7.

The full conditional posterior distribution for p^s_{pj} is (Gelman et al., 2014, Section 20.2)

$$\pi(p^s_{pj}|-) \sim \text{Beta}\left(\alpha^{\text{s}} + \sum_{\text{p}=1}^{\text{p}} \delta^{\text{s}}_{\text{pj}}, \beta^{\text{s}} + \sum_{\text{p}=1}^{\text{p}} (1 - \delta^{\text{s}}_{\text{pj}})\right).$$

STEP 8.

We apply the same approach to the probability p_{pk} of the common factor loadings with the prior $p_{pk} \sim \text{Beta}(\alpha_{\text{p}}, \beta_{\text{p}})$. Like in the above step, we obtain

$$\pi(p_{pk}|-) \sim \text{Beta}\left(\alpha + \sum_{\text{p}=1}^{\text{p}} \delta_{\text{pk}}, \beta + \sum_{\text{p}=1}^{\text{p}} (1 - \delta_{\text{pk}})\right).$$

STEP 9.

Finally we compute the posterior distribution for the covariance matrix of error term Ψ_s. Taking in consideration each element of Ψ_s, $\psi_{ps},\ p = 1, \ldots, P$

$$\begin{aligned}
\pi(\Psi_s|-) &= \prod_{s=1}^{S}\prod_{p=1}^{P} \left(\psi_{ps}^{-2}\right)^{n_s/2} \exp\left(\frac{\mathbf{C}_{x_s x_s}}{2}\psi_{ps}^{-2}\right) \left(\psi_{ps}^{-2}\right)^{\frac{\alpha_\psi}{2}-1} \exp\left(-\frac{\beta_\psi}{2}\psi_{ps}^{-2}\right) \\
&= \prod_{s=1}^{S}\prod_{p=1}^{P} \left(\psi_{ps}^{-2}\right)^{\frac{n_s+\alpha_\psi}{2}-1} \exp\left(-\frac{\mathbf{C}_{x_s x_s} + \beta_\psi}{2}\psi_{ps}^{-2}\right).
\end{aligned}$$

Thus, as in FA the full conditional for each ψ^{-2}_{ps} in the study s is

$$\pi(\psi^{-2}_{ps}|-) \sim \text{Ga}\left(\frac{\text{n}_\text{s} + \alpha_\psi}{2}, \frac{\mathbf{C}_{\text{x}_\text{s}\text{x}_\text{s}} + \beta_\psi}{2}\right).$$

5.3 Analysis of simulated data

In order to validate our procedure, we perform a small simulation experiment. Here, we present two different studies with $p = 60$. The number of samples is set to $n_s = \{25, 30\}$ to better understand the behavior of the algorithm proposed in the $p > n$ settings. We consider the first study having a total latent dimension of 4 and the second one a total latent dimension

of 5. Table 5.2 provides the details. For simulated data we have the advantage of knowing exactly the sparsity pattern. Here we set around 70% of the loadings to zero.

Table 5.1 Settings for the simulation study.

S	n_s	$K+J_s$
1	25	4
2	30	5

The data come from a MFA model with $K = 1$.

As denoted by George and McCulloch (1993), the choice of the parameters of the spike and slab prior is really crucial. Moreover, the choice is made more complex by the presence of the common and specific factors. Adopting an informative approach, which is essential in sparse settings, the beta hyper prior on the probability if belonging to the non-sparse set of Bernoulli simplifies in part this problem. For the specific factor loadings we set $\tau^2_{pjs} = 10^{-4}$ and $c^2_{\lambda_{pjs}} = 1000$, and we choose the same value for the common factor loadings, $\zeta^2_{pk} = 10^{-4}$ and $c^2_{\phi_{pk}} = 1000$. Note that setting $\tau_{pjs} = 0.01$ implies that when $\delta^s_{pj} = 0$, then λ_{pjs} will (nearly) assume values in $(-0.03, 0.03)$, a range of values compatible with loadings that can be taken as nil. We place a Beta$(25, 200)$ on p^s_{pj} and on p_{pk}. As showed in Figure 5.4, the beta prior leads to the p^s_{pj} and p_{pk} concentrated around 0.1. This amounts to higher sparsity level than that used to generate the data, but we checked that other choices, such as a prior concentrated around 0.2, lead to similar results.

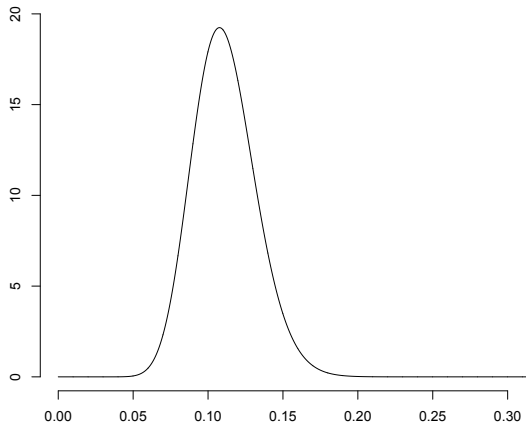

Figure 5.4 Density function obtained by a Beta(25,200)

We choose the same beta priors for the common and specific factor loadings since the sparsity is similar in the common factors and in the specific ones. In other applications where there is the possibility to have different sparsity level, different beta priors could be chosen.

Finally we choose a Ga(1,0.3) for a_ψ and b_ψ, as customary in FA (Bhattacharya and Dunson, 2011; Lopes and West, 2004; West, 2003) .

We run the Gibbs sampler described before for 10000 iterations with a burn-in of 1000.

First, we check the convergence properties of the algorithm. The convergence is suggested by some indexes and tests performed, such as Gelman and Rubin Multiple Sequence Diagnostic (Brooks and Gelman, 1998), and Geweke Diagnostic (Cowles and Carlin, 1996). The chains hint to stationarity and we proceed with the analysis. In order to validate our procedure we check whether the posterior results reproduce the true sparsity pattern.

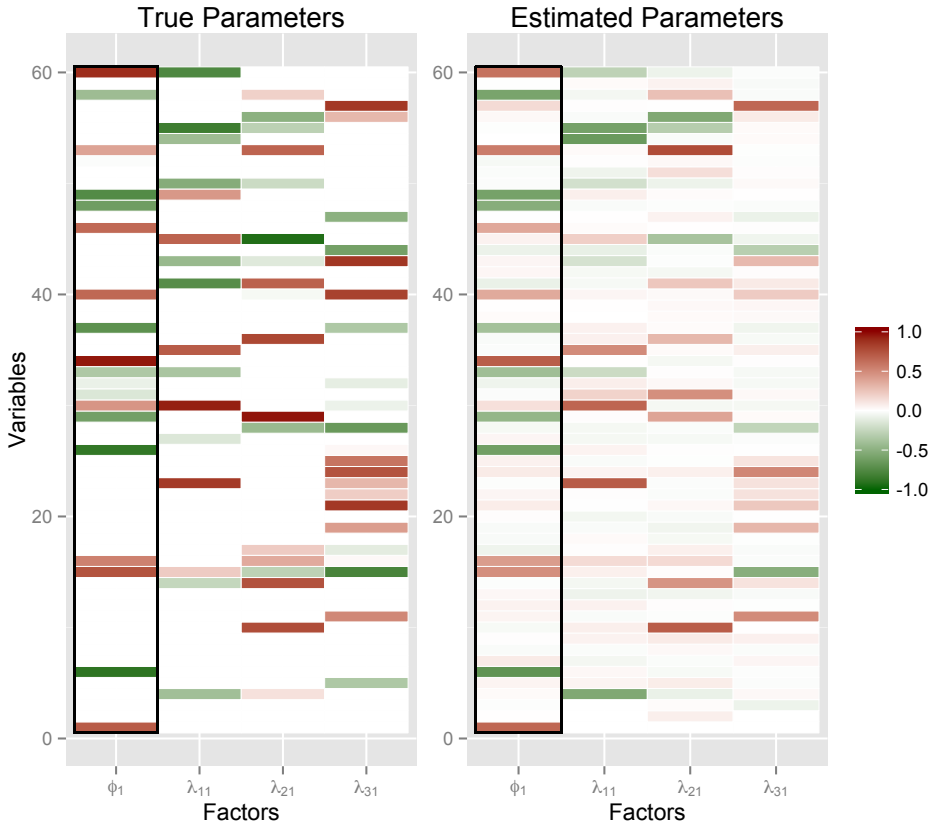

Figure 5.5 Comparison of the two factor loading matrices in the first study, one obtained with the true parameters (on the left) and the other one obtained by the mean of the posterior sampling (on the right).

This is depicted in Figure 5.5 in which, after adjusting the sign of a column, we perform a comparison of the posterior medians with the true parameters, reporting only the first study

for brevity. As reported by Figure 5.5, the proposed method is able to capture the sparsity components of the Φ and Λ_1, and also to broadly identify the elements of the loadings different from zero.

We turn then to the problem of selecting the latent dimensions of common and specific factors. What we expect, especially with large dimensions, is to have some loadings with no contribution to the relation between the observed and the latent components. So when the number of the fitted latent dimension is too large, we expect to find some columns in the factor loadings matrix with all the elements close to zero.

We note in passing that some experiments with the usage of the Deviance Information Criteria (DIC) (Spiegelhalter et al., 2002) for model selection were performed. Not surprisingly for sparsity settings, the results were not encouraging, and this route abandoned.

We perform the Bayesian MFA in the settings define in Table 5.2 with seven common factors, six specific factors for the first study and eight for the second one.

In order to assess the total latent dimension, we pay attention to the following two steps.

- Monitor the columns of the loadings. In this process, we use the posterior median to estimate the loadings, since when the dimension of the loading matrix increases the estimation becomes more demanding. Indeed, when the dimension of the latent factors increases, the trace plots of MCMC iterations for some columns of the loading matrices may exhibit some unsatisfactory mixing, with high uncertainty about the sparsity pattern. The posterior median is therefore a more cautious estimate.

- Take as a indication of non-sparsity for a column j of Φ (and similarly for the other loading matrices) if

$$\frac{1}{p^*}\sum_{i=1}^{p^*} \mid \hat{\Phi}_{pk} \mid \geq \varepsilon, \quad \varepsilon = 0.03,$$

  where p^* denotes the number of loadings of the p^{th} column. Here we exclude the elements set to zero due to identify the model, so that $p^* < p$.

By doing this, we obtained a suggestion of $k = 1$, $j_1 = 3$ and $j_2 = 4$, so that the fitted latent dimension are actually the same as the true one.

Some further considerations are in order. The threshold ε is chosen equal to 0.03 accordingly to the prior chosen for the loadings, but a sensitive analysis is warranted considering also other values. Moreover, for higher-dimensional settings other choices might be sensible. Furthermore, a further outcome of the proposed approach is that Bernoulli random variables δ_{pk} and δ^s_{pj} provide some useful information (see also Gelman et al., 2014, p.491). Indeed, Carvalho et al. (2008) make an intense use of the posterior probabilities $P(\delta_{pk} = 1|\text{data})$

in order to define the total latent dimensions and also to check about the inclusion of further variables in the model. In particular, their approach identifies a significant gene-factor interaction when such posterior probability is higher than a given threshold.

5.4 Application in a $p > n$ context

In this section some analyses to the ovarian cancer are performed to validate the procedure described before. We consider the first two studies presented in Table 2.1, GSE9891 and GSE20565, performing the analysis with the two pathways described in Chapter 4. So we have $n_s = 285, 140$ and $p = 163$, thus resulting in a $p > n$ setting.

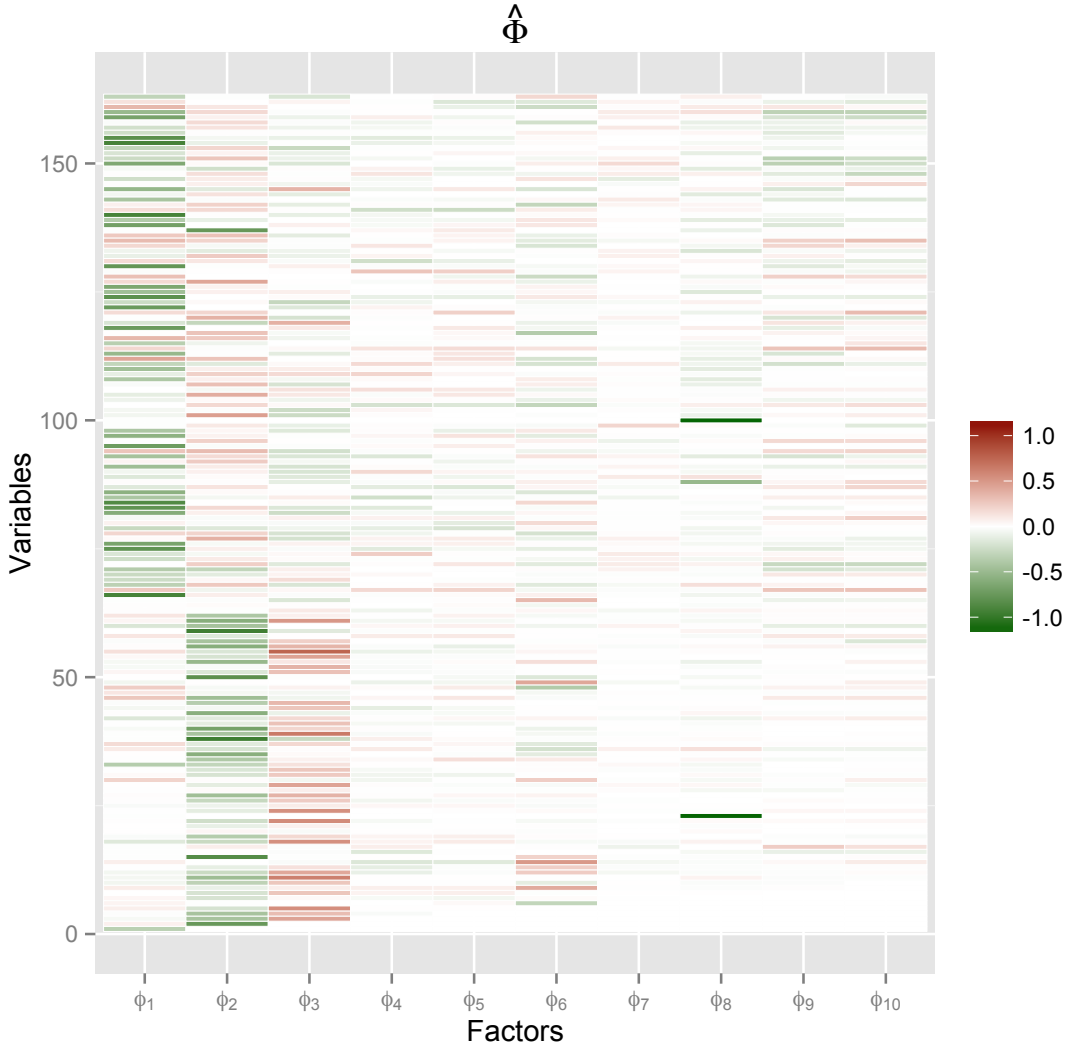

Figure 5.6 Heatmap of the posterior median for Φ with k=10.

For the common factor loadings, we set $\zeta_{pk}^2 = 10^{-4}$ and $c_{\phi_{pk}}^2 = 5000$, and we choose the same value for the specific factor loadings, $\tau_{pjs}^2 = 10^{-4}$ and $c_{\lambda_{pjs}}^2 = 5000$. We place a Beta(25,200) on p_{pj}^s and on p_{pk} in order to have a prior concentrated on 0.1. This choice corresponds to higher sparsity level than that obtained by the frequentist analysis of Chapter 4 as the latter is based on a non-singular covariance matrix. Furthermore, the Beta(25,200) gave satisfactory results in recovering the sparsity pattern of the simulated data of the previous section.

Some preliminary analyses with large latent dimensions, such as $k = 15$, $j_1 = 10$ and $j_2 = 10$, readily suggest that there are several sparse specific factor loadings, not significantly contributing to the relation between the observed and latent components. We then decided to switch to a model with $k = 10$, $j_1 = 5$, $j_2 = 3$. Such values are similar to those obtained in the frequentist analysis of Chapter 4. More precisely, in the frequentist analysis with $p = 63$ we obtained $k = 4$, $j_1 = 2$ and $j_2 = 3$, and with $p = 100$ we obtained $k = 5$, $j_1 = 1$ and $j_2 = 2$.

We run a Gibbs sampling with these latent dimensions for 50000 iterations with a burn-in of 10000. The results are represented in Figure 5.6, restricted to the common factor loadings Φ.

Here, an adjustment to the sign of columns 7 and 10 has been performed. Table 5.2

Table 5.2 Non-sparsity index values for Φ with k=10.

K	Non-sparsity index
1	0.209
2	0.189
3	0.137
4	0.050
5	0.050
6	0.080
7	0.025
8	0.057
9	0.062
10	0.062

summarizes the posterior non-sparsity index $\frac{1}{p^*}\sum_{i=1}^{p^*} \mid \hat{\Phi}_{pk} \mid$. These results hint to a model with just nine common latent factors, five specific factors for the first study and three for the second study. For the common factor we get the same dimension of the frequentist analysis.

Figure 5.7 shows the elements of Φ for which $P(\delta_{pk} = 1|\text{data}) \geq 0.95$ on the right and for $P(\delta_{pk} = 1|\text{data}) \geq 0.99$ on the left. Following Carvalho et al. (2008), they can be taken as expressing significant gene-factor associations.

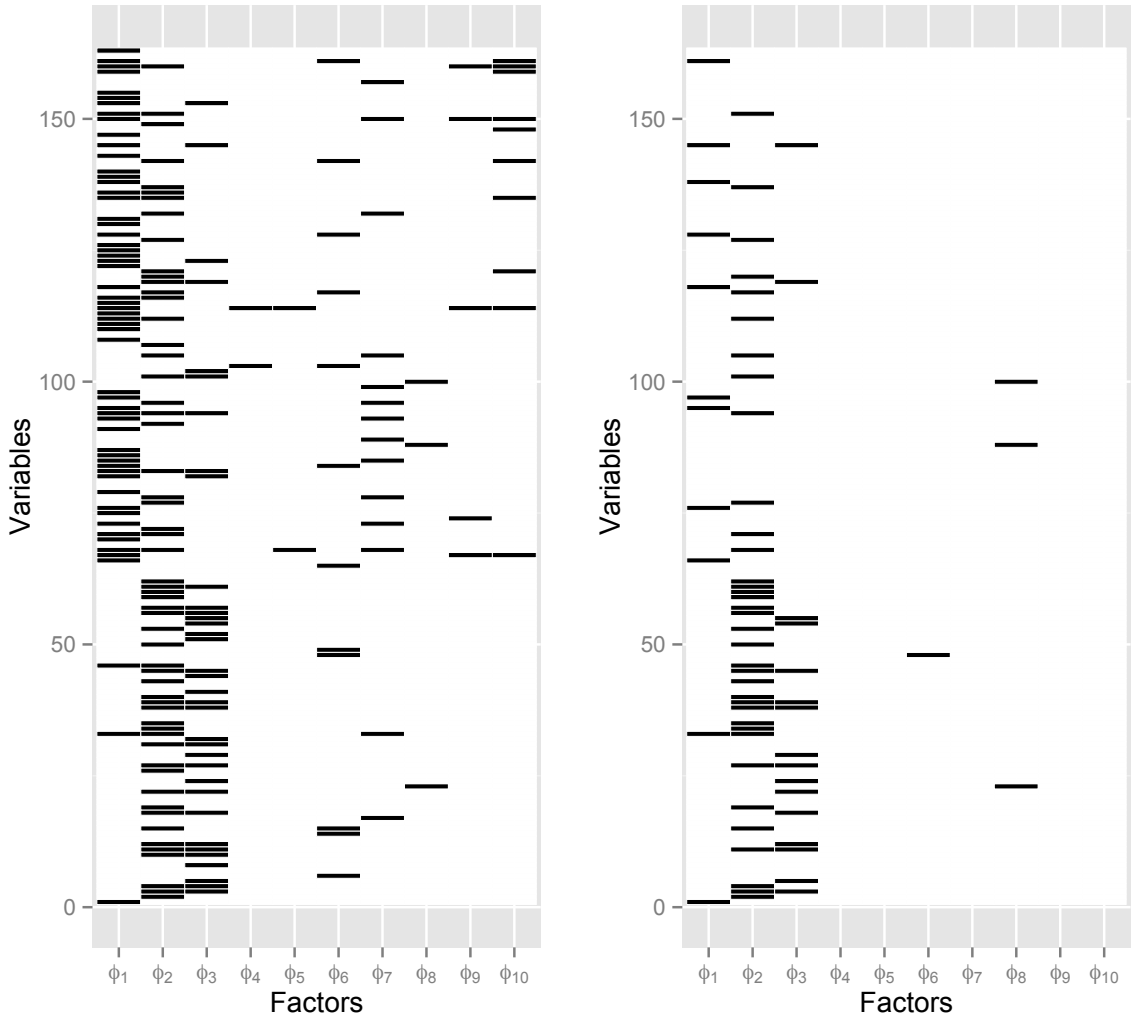

Figure 5.7 Heatmap of $P(\delta_{pk} = 1|\text{data}) \geq 0.95$ (left) and $P(\delta_{pk} = 1|\text{data}) \geq 0.99$ (right) with $k = 10$.

5.5 Discussion

Modeling sparsity settings becomes crucial in contexts with more variables than subjects, which are common with microarray gene expression data. The proposal of this chapter allows to fit the MFA model also to this kind of settings, and the results obtained seem encouraging, but there are some points worth noting.

Indeed, the settings considered here involve more variables than subjects, yet the two dimensions are not too dissimilar. In a suitable sense, this is demonstrated by the fact that for the the analysis of the real data of Section 5.4 we could compare the results obtained employing the Bayesian sparse prior with those obtained with $p = 100$ (or $p = 63$) variables

by means of maximum likelihood estimation. Extensions to settings with $p >> n$, for which frequentist analyses of a very small portion of the available variables would be much less meaningful, would require more care.

Fitting sparse MFA adopting a spike-and-slab prior requires quite informative prior distributions, and very high ratios p/n would require to adopt a much stronger prior level of sparsity than that chosen here. Furthermore, the uncertainty about the nullity of some loadings alluded to in Section 5.3, leading to unsatisfactory mixing in the MCMC results, would be even greater in such settings. To this end, the proposal of Carvalho et al. (2008), who adopted a further spike-and-slab structure for the beta prior on the success probability for the δs, seems worth investigating for possible extension to the MFA model. This is actually a point deserving further study, but whose implementation appears within reach. We end by noting that, on the computational side, the `Julia` programming language used for the analyses of this chapter would be a rather good choice also for this kind of extension.

Appendix A

Computational tools

The thesis deals with models for high-dimensional data, involving a large number of parameters. The `R` statistical software, used for most of the data-cleaning and graphical analyses, turned out to be inadequate for the simulation studies of Chapter 3 or the Bayesian analyses on the sparse settings of Chapter 5. The high-performance `Julia` programming language was then employed for such tasks. More details on this powerful programming tool are available at `http://julialang.org`. For the Bayesian analysis of Chapter 5, the public `Matlab` code associated to the paper by Bhattacharya and Dunson (2011) turned out to be useful, and we thanks these Authors to make it available. Finally, several models were also implemented in `JAGS` (`http://mcmc-jags.sourceforge.net`), which was rather valuable at the developing stage, for cross-checking the results obtained both in `R` and `Julia`.

Bibliography

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

Kohei Adachi. Factor analysis with EM algorithm never gives improper solutions when sample covariance and initial parameter matrices are proper. *Psychometrika*, 78(2):380–394, nov 2012. doi: 10.1007/s11336-012-9299-8.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

B. Alberts. *Molecular Biology of the Cell: Reference edition*. Number v. 1 in Molecular Biology of the Cell: Reference Edition. Garland Science, 2008. ISBN 9780815341116.

Anirban Bhattacharya and David B Dunson. Sparse bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011.

Yuna Blum, Guillaume Le Mignon, Sandrine Lagarrigue, and David Causeur. A factor model to analyze heterogeneity in gene expression. *BMC Bioinformatics*, 11(1):368, 2010.

Dankmar Böhning, Ekkehart Dietz, Rainer Schaub, Peter Schlattmann, and Bruce G Lindsay. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46(2):373–388, 1994.

Tomas Bonome, Douglas A Levine, Joanna Shih, Mike Randonovich, Cindy A Pise-Masison, Faina Bogomolniy, Laurent Ozbun, John Brady, J Carl Barrett, Jeff Boyd, et al. A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer Research*, 68(13):5478–5486, 2008.

Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

Kenneth P Burnham and David R Anderson. *Model Selection and Multimodel Inference: a Practical Information-theoretic Approach*. Springer Science & Business Media, New York, second edition, 2002.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Carlos M Carvalho, Jeffrey Chang, Joseph E Lucas, Joseph R Nevins, Quanli Wang, and Mike West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

Raymond B Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45(10):1127–1133, October 2013.

Erin M Conlon, Joon J Song, and Anna Liu. Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics*, 8(1):80, 2007.

Leslie Cope, Daniel Q Naiman, and Giovanni Parmigiani. Integrative correlation: Properties and relation to canonical correlations. *Journal of Multivariate Analysis*, 123:270–280, January 2014.

Mary Kathryn Cowles and Bradley P Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91 (434):883–904, 1996.

Aedín C Culhane, Guy Perrière, and Desmond G Higgins. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, 4(1): 59, 2003.

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

Amit G Deshwar and Quaid Morris. Plida: cross-platform gene expression normalization using perturbed topic models. *BMC Bioinformatics*, 30(7):956–961, 2014.

Kim-Anh Do, Peter Müller, and Marina Vannucci. *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, New York, 2006.

F Dominici, G Parmigiani, K H Reckhow, and R L Wolpert. Combining Information from Related Regressions. *Journal of Agricultural, Biological, and Environmental Statistics*, 2: 313–332, 1997.

F Dominici, G Parmigiani, R L Wolpert, and Vic Hasselblad. Meta-analysis of Migraine Headache Treatments: Combining information from heterogeneous designs. *Journal of the American Statistical Association*, 94:16–28, 1999.

Barbara E Engelhardt and Matthew Stephens. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genetics*, 6(9), 2010.

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

Sylvia Frühwirth-Schnatter and Hedibert Freitas Lopes. Parsimonious bayesian factor analysis when the number of factors is unknown. *Unpublished Working Paper, Booth Business*, 2010.

Benjamin Frederick Ganzfried, Markus Riester, Benjamin Haibe-Kains, Thomas Risch, Svitlana Tyekucheva, Ina Jazic, Xin Victoria Wang, Mahnaz Ahmadifar, Michael J Birrer, Giovanni Parmigiani, et al. curatedovariandata: clinically annotated data for the ovarian cancer transcriptome. *Database*, 2013, 2013.

Jianjiong Gao, Giovanni Ciriello, Chris Sander, and Nikolaus Schultz. Collection, integration and analysis of cancer genomic profiles: from data to insight. *Current Opinion in Genetics & Development*, 24:92–98, February 2014.

E Garrett-Mayer, G Parmigiani, X Zhong, L Cope, and E Gabrielson. Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics*, 9(2):333–354, July 2007.

Elizabeth Garrett-Mayer, Giovanni Parmigiani, Xiaogang Zhong, Leslie Cope, and Edward Gabrielson. Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics*, 9(2):333–354, 2008.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, and Donald B Rubin. *Bayesian Data Analysis*. Taylor & Francis, Florida, third edition, 2014.

Robert Gentleman, Vincent Carey, Wolfgang Huber, Rafael Irizarry, and Sandrine Dudoit. *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer Science & Business Media, 2006.

Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Edward I George and Robert E McCulloch. Approaches for bayesian variable selection. *Statistica Sinica*, 7(2):339–373, 1997.

J. Geweke and G. Zhou. Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies*, 9(2):557–587, apr 1996. doi: 10.1093/rfs/9.2.557.

Louis Guttman. Some necessary conditions for common-factor analysis. *Psychometrika*, 19 (2):149–161, 1954.

Chris Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.

Wolfgang Härdle and Léopold Simar. *Applied Multivariate Statistical Analysis*, volume 2. Springer, 2003.

D Neil Hayes, Stefano Monti, Giovanni Parmigiani, C Blake Gilks, Katsuhiko Naoki, Arindam Bhattacharjee, Mark A Socinski, Charles Perou, and Matthew Meyerson. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *Journal of Clinical Oncology*, 24(31):5079–5090, November 2006.

Kei Hirose and Michio Yamamoto. Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics & Data Analysis*, 79:120–132, 2014.

Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

John L Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

Lloyd G Humphreys and Richard G Montanelli. An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, 10(2):193–205, 1975.

Curtis Huttenhower, Matt Hibbs, Chad Myers, and Olga G Troyanskaya. A scalable method for integration and functional analysis of multiple microarray datasets. *BMC Bioinformatics*, 22(23):2890–2897, December 2006.

Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, Terence P Speed, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

Karl G Jöreskog. Simultaneous factor analysis in several populations. *Psychometrika*, 36(4): 409–426, 1971.

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

Yihan Li and Debashis Ghosh. Assumption weighting for incorporating heterogeneity into meta-analysis of genomic data. *BMC Bioinformatics*, 28(6):807–814, 2012.

Hedibert Freitas Lopes and Mike West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–68, 2004.

Joseph E Lucas, Hsiu-Ni Kung, and Jen-Tsan A Chi. Latent factor analysis to discover pathway-associated putative segmental aneuploidies in human cancers. *PLoS Computational Biology*, 6(9), 2010.

Geoffrey McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons, New Jersey, second edition, 2007.

Gàbor Méhes, Andrea Luegmayr, Claudia M Hattinger, Thomas Lörch, Inge M Ambros, Helmut Gadner, and Peter F Ambros. Automatic detection and genetic profiling of disseminated neuroblastoma cells. *Medical and Pediatric Oncology*, 36(1):205–209, 2001.

Chen Meng, Bernhard Kuster, Aedín C Culhane, and Amin M Gholami. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, 15(1):162, 2014.

William Meredith. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543, 1993.

Jean-Philippe Meyniel, Paul H Cottu, Charles Decraene, Marc-Henri Stern, Jérôme Couturier, Ingrid Lebigot, André Nicolas, Nina Weber, Virginie Fourchotte, Séverine Alran, et al. A genomic and transcriptomic approach for a differential diagnosis between primary and secondary ovarian carcinomas in patients with a previous history of breast cancer. *BMC Cancer*, 10(1):222, 2010.

Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, et al. Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, 2003.

Stanley A Mulaik. *Foundations of Factor Analysis*. CRC press, Florida, second edition, 2009.

P Muller, G Parmigiani, J Schildkraut, and L Tardella. A Bayesian hierarchical approach for combining case-control and prospective studies. *Biometrics*, 55(3):858–866, September 1999.

National Academy of Sciences. Statistical challenges in assessing and fostering the reproducibility of scientific results: A workshop. "http://sites.nationalacademies.org/DEPS/BMSA/DEPS_153236", 2015.

Cancer Genome Atlas Research Network et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 2011.

Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Giovanni Parmigiani, Elizabeth S Garrett-Mayer, Ramaswamy Anbazhagan, and Edward Gabrielson. A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 10(9):2922–2927, May 2004.

Paul D P Pharoah, Ya-Yu Tsai, Susan J Ramus, Catherine M Phelan, Ellen L Goode, Kate Lawrenson, Melissa Buckley, Brooke L Fridley, Jonathan P Tyrer, Howard Shen, et al. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nature Genetics*, 45(4):362–370, April 2013.

Iosifina Pournara and Lorenz Wernisch. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, 8(1):61, 2007.

Kristopher J Preacher and Edgar C Merkle. The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*, 17(1):1, 2012.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

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

Markus Riester, Wei Wei, Levi Waldron, Aedin C Culhane, Lorenzo Trippa, Esther Oliva, Sung-Hoon Kim, Franziska Michor, Curtis Huttenhower, Giovanni Parmigiani, and Michael J Birrer. Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *Journal of the National Cancer Institute*, 106(5), May 2014.

Donald B Rubin and Dorothy T Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, 1982.

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

RB Scharpf, H Tjelmeland, G Parmigiani, and AB Nobel. A Bayesian model for cross-study differential gene expression. *Journal of the American Statistical Association*, 104(488): 1295–1310, 2009a.

Robert B Scharpf, Håkon Tjelmeland, Giovanni Parmigiani, and Andrew B Nobel. A Bayesian model for cross-study differential gene expression. *Journal of the American Statistical Association*, 104(488), 2009b.

Almut Schulze and Julian Downward. Navigating gene expression using microarrays—a technology review. *Nature Cell Biology*, 3(8):E190–E195, 2001.

Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

Andrey A Shabalin, Håkon Tjelmeland, Cheng Fan, Charles M Perou, and Andrew B Nobel. Merging two gene-expression studies via cross-platform normalization. *BMC Bioinformatics*, 24(9):1154–1160, 2008.

Ronglai Shen, Debashis Ghosh, and Arul M Chinnaiyan. Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics*, 5 (1):94, 2004.

Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, Shawn C Baker, Patrick J Collins, Francoise De Longueville, Ernest S Kawasaki, Kathleen Y Lee, et al. The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, 2006.

Rebecca Siegel, Deepa Naishadham, and Ahmedin Jemal. Cancer statistics, 2012. *A Cancer Journal for Clinicians*, 62(1):10–29, 2012.

Terry Speed. *Statistical Analysis of Gene Expression Microarray Data*. CRC Press, Florida, 2003.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

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

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. 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(43):15545–15550, 2005.

Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

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

Lewis L Thurstone. Multiple factor analysis. *Psychological Review*, 38(5):406, 1931.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Richard W Tothill, Anna V Tinker, Joshy George, Robert Brown, Stephen B Fox, Stephen Lade, Daryl S Johnson, Melanie K Trivett, Dariush Etemadmoghadam, Bianca Locandro, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*, 14(16):5198–5208, 2008.

Svitlana Tyekucheva, Luigi Marchionni, Rachel Karchin, and Giovanni Parmigiani. Integrating diverse genomic data using gene sets. *Genome Biology*, 12(10):R105, 2011.

Charles F Van Loan. The ubiquitous kronecker product. *Journal of Computational and Applied Mathematics*, 123(1):85–100, 2000.

Levi Waldron, Benjamin Haibe-Kains, Aedin C Culhane, Markus Riester, Jie Ding, Xin Victoria Wang, Mahnaz Ahmadifar, Svitlana Tyekucheva, Christoph Bernau, Thomas Risch, Benjamin Frederick Ganzfried, Curtis Huttenhower, Michael Birrer, and Giovanni Parmigiani. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *Journal of the National Cancer Institute*, 106(5), May 2014a.

Levi Waldron, Benjamin Haibe-Kains, Aedín C Culhane, Markus Riester, et al. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *Journal of the National Cancer Institute*, 106(5):dju049, 2014b.

Jing Wang, Kevin R Coombes, W Edward Highsmith, MJ Keating, and Lynne V Abruzzo. Differences in gene expression between b-cell chronic lymphocytic leukemia and normal b cells: a meta-analysis of three microarray studies. *BMC Bioinformatics*, 20(17):3166–3178, 2004.

Xin Victoria Wang, RG Verhaak, Elizabeth Purdom, Paul T Spellman, and Terence P Speed. Unifying gene expression measures from multiple platforms using factor analysis. *PloS One*, 6(3), 2011.

Mike West. Bayesian factor regression models in the "large p small n" paradigm. *Bayesian Statistics*, 7:1–11, 2003.

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

Yang Xie, Wei Pan, Kyeong S Jeong, Guanghua Xiao, and Arkady B Khodursky. A bayesian approach to joint modeling of protein-dna binding, gene expression and sequence data. *Statistics in Medicine*, 29(4):489, 2010.

J-H Zhao, LH Philip, and Qibao Jiang. ML estimation for factor analysis: EM or non-EM? *Statistics and Computing*, 18(2):109–123, 2008.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

William R Zwick and Wayne F Velicer. Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99(3):432, 1986.

Roberta de Vito

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistical Sciences
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4174
e-mail: devito@stat.unipd.it

Current Position

Since January 2013; (expected completion: March 2016)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Multi-study factor models for high-dimensional biological data
Supervisor: Prof. Bellio
Co-supervisor: Prof. Parmigiani.

Research interests

- Biostatistics
- High-dimensional data
- Latent variable models

Education

2009 – 2012
M.S. in Statistical Science.
University La Sapienza, Rome
Title of dissertation: "Joint alalysis of efficacy and toxicity in clinical trials"
Supervisor: Prof. Ludovico Piccinato
Final mark: 110/110 cum laude

2006 – 2009
B.S. in Statistics, Population and Society.
University La Sapienza, Rome
Title of dissertation: "Determinant Factors in the Anemia in Malawi "
Supervisor: Prof. Marco P. Pacifico
Final mark: 110/110

Visiting periods

January 2014 – November 2015
Department of Biostatistics and Computational Biology,
Harvard T.H.Chan School of Public Health, Boston, USA.
Supervisor: Giovanni Parmigiani

September – December 2011
Department of Biostatistics and Computational Biology,
Harvard T.H.Chan School of Public Health, Boston, USA.
Supervisor: Giovanni Parmigiani

Further education

December 2014
Modeling Ordinal Categorical Data (*A. Agresti*)
Harvard University

March 2014
Frequentist Accuracy of Bayesian Estimates (*B. Efron*)
Harvard University

February 2014
Decision Curve analysis (*G. Parmigiani*)
Harvard University

Awards and Scholarship

2011
MSc Thesis Award to conduct research at Harvard T.H. Chan School of Pub- lic Health, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute.
2011 and 2010
Study Fellowship, University La Sapienza, Rome , Italy.
2008
Erasmus Fellowship, Southampton University, UK, focus on Population and En- vironment, Multivariate Data Analysis and Demography.

Computer skills

- R, Julia, Jags, SAS, SPSS, Matlab, Mathematica.

Language skills

Italian native; English fluent.

Academic Papers

de Vito R., Trippa L., Bellio R., Parmigiani G. Multi-study Factor Analysis for High Dimensional Biological Data, in preparation.
de Vito R., Trippa L., Bellio R., Parmigiani G. Sparse Bayesian Multi-study Factor Models for

High-Dimensional Data, in preparation

Oral Presentations

Factor analysis in high-dimensional data, Computer Science Department, Center for Statistics and Machine Learning, Princeton University, November 2015.

Multi-study Factor Analysis for Biological Data, speaker at the Genomic Meeting, Department of Biostatistic and Computational Biology Department, Dana Farber Cancer Institute, October 2015.

Multi-study Factor Analysis in High Dimensional Biological Data, Biostatistic and Computational Biology Department, Dana Farber Cancer Institute, March 2015.

Advanced Multi-study Techniques in High Dimensional Data, Seminar at the Department of Statistics, University of Padua, February 2015 .

References

Prof. Ruggero Bellio
Department of Economics and Statistics
University of Udine
Via Tomadini, 30/a
33100 Udine - Italy.
e-mail: ruggero.bellio@uniud.it

Prof. Giovanni Parmigiani
Biostatistics and Computational Biology
Harvard Uiverisity
44 Binney Street
02115 Boston-USA
e-mail: gp@jimmy.harvard.edu