Learning from high-dimensional multivariate signals

by

Arnau Tibau-Puig

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2012

Doctoral Committee:

Professor Alfred O. Hero III, Chair
Professor Anna Catherine Gilbert
Assistant Professor Rajesh Rao Nadakuditi
Assistant Professor Clayton D. Scott

"Als meus pares, per haver-me donat el tresor de la vida i haver-me ensenyat a gaudir-ne."

To my parents, for giving me the treasure of life and teaching me how to enjoy it.

ACKNOWLEDGEMENTS

This work would not have been possible without the advice, guidance and patience of several individuals. Chief among them is Professor Hero, who has guided me throughout the last four years with dedication, and has taught me that, in research, there is always an aspect worth exploring in more detail, a paragraph deserving additional editing, or a proof waiting to be simplified.

At the beginning of this journey, I must confess, I was quite lost: geographically (after moving to a new country), climatically (Michigan's January is quite different from Spain's), and also intellectually (what to do with four years of research ahead!). Several people helped me relocate my spirit and my body, and contributed to the success of this project. By order of appearance, Ami Wiesel taught me how to start and, more importantly, how to finish my first conference paper, and also half of what I have learnt in convex optimization. My numerous other colleagues in the Hero Research Group (Mark Kliger, Patrick Harrington, Yongsheng Juang, Kevin Xu, Greg Newstadt, Yilun Chen, Kumar Shridharan, Sung Jin Huang, Fra Bassi, Denis Wei, Mark Shiao, Zaoshi Meng) made my arrival and my life at the University of Michigan much easier, both professionally and personally. On the other side of the Atlantic Ocean, Gilles Fleury, Laurent Le Brusquet and Arthur Tenenhaus made me feel at home during my summer stays at Supelec, France, and gave me valuable insight and comments during each of our frequent videoconferences.

At the University of Michigan, I have had the unvaluable opportunity to learn about Sparse Approximmation/Compressive Sensing and Random Matrix Theory from two of the leading experts in the field, Anna Gilbert and Raj Rao Nadakuditi, who gracefully accepted to sit in my dissertation committee. Together with Clayton Scott, they gave me insightful comments and advice towards the transformation of my dissertation proposal into the present document.

I would also like to express my gratitude to all the good friends I have made during the last years. I came to the US socially empty-handed, and I will probably leave with many life-long friends from whom I learnt and enjoyed a great deal. To those who gave me support during each of my winter crises: Cristina Reina and Ricardo Ferrera,

Nicolas Lamorte, Remy Elbez, Emilie Bourdonnay, Aghapi Mordovanakis, Myriam Affeiche, Paul Samaha, Hannah Darnton, Emmanuel Bertrand, Laith Alattar, Adrien Quesnel, Wylde Norellus, Sofiya and Jamie Dahman, Le Nguyen, German Martinez and Maria Vega, Rahul Ahlawat, Matt Reyes and Awlok Josan: Thanks to you all!

I would finally like to thank the many people from the Electrical Engineering and Computer Science department that are always willing to assist graduate students in need of help: Becky Turanski, Ann Pace, Michel Feldman, and the DCO team, whose members always managed to reply to my frequent demands within hours.

I can't finish this section without dedicating a special paragraph to Maria Luz Alvarellos, who gently put up with me during the hard-working months that preceded my oral defense, always ready to give a word (or a kiss) of encouragement. Gracias, Maria Luz!

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

NOTATION

The following notation convention is adopted throughout this dissertation. Boldface upper case letters denote matrices, boldface lower case letters denote column vectors, standard lower case letters denote scalars and standard upper case letters denote random variables. In addition, we define:

Sets and generalized inequalities:

$\mathbb{R}$: the field of real numbers.
$\mathbb{C}$: the field of complex numbers.
$\mathbb{R}^p$: the set of p-dimensional real-valued vectors.
$\mathbb{R}^p_+$: the non-negative orthant, i.e., the set $\{\boldsymbol{x} \in \mathbb{R} : x_i \geq 0\}$.
$\mathbb{R}^p_{++}$: the positive orthant, i.e., the set $\{\boldsymbol{x} \in \mathbb{R} : x_i > 0\}$.
$\mathbb{R}^{p\times n}$: the set of $p \times n$ real-valued matrices.
$\mathbb{S}^p$: the set of $p \times p$ symmetric matrices.
$\mathbb{S}^p_+$: the subset of semi-positive definite matrices, i.e.

$$\mathbb{S}^p_+ = \left\{ \boldsymbol{X} \in \mathbb{S}^p : \boldsymbol{x}^T \boldsymbol{X} \boldsymbol{x} \geq 0, \forall \boldsymbol{x} \in \mathbb{R}^p \right\}.$$

$\mathbb{S}^p_{++}$: the subset of positive definite matrices, i.e.

$$\mathbb{S}^p_{++} = \left\{ \boldsymbol{X} \in \mathbb{S}^p : \boldsymbol{x}^T \boldsymbol{X} \boldsymbol{x} > 0, \forall \boldsymbol{x} \in \mathbb{R}^p \right\}.$$

$\boldsymbol{x} \succeq \boldsymbol{0}$ ($\boldsymbol{x} \succ \boldsymbol{0}$) means that $\boldsymbol{x} \in \mathbb{R}^p_+$ ($\boldsymbol{x} \in \mathbb{R}^p_{++}$).
$\boldsymbol{X} \succeq \boldsymbol{0}$ ($\boldsymbol{X} \succ \boldsymbol{0}$) means that $\boldsymbol{X} \in \mathbb{S}^p_+$ ($\boldsymbol{X} \in \mathbb{S}^p_{++}$).
$||\cdot||_2$: the ℓ_2 norm of a vector, $||\boldsymbol{x}||_2 = \sqrt{\sum_{i=1}^p x_i^2}$.
$||\cdot||_F$: the Frobenius norm of a matrix, $||\boldsymbol{X}||_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^n x_{i,j}^2}$.
Given a proper cone $C \subset \mathbb{R}^p$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, $\boldsymbol{x} \succeq_C \boldsymbol{y}$ means that $\boldsymbol{x} - \boldsymbol{y} \in C$.
K: the second order (Lorentz) cone,

$$K = \left\{ \left[\begin{array}{c} \boldsymbol{x} \\ t \end{array} \right] \in \mathbb{R}^{N+1} : ||\boldsymbol{x}||_2 \leq t \right\}.$$

Matrix notation and operators:

^{T}: matrix transpose operator.
†: matrix pseudoinverse operator.
$\otimes$: Kronecker product.

$[\cdot]_{S,T}$: submatrix operator, returns the matrix constructed from the rows indexed by S and the columns indexed by T. We will drop the brackets whenever this is possible. Similarly, $[\cdot]_{S,*}$ denotes the submatrix obtained from the row indices in S and all its columns.
$\operatorname{diag}(\boldsymbol{x})$ returns a diagonal matrix with the elements of $\boldsymbol{x}$ in its diagonal.
$\operatorname{tr}(\cdot)$: the trace operator.
$\det(\cdot)$: the determinant.
$\operatorname{supp}(\boldsymbol{X})$ returns the subset of column indices corresponding to columns with at least one non-zero element.
$\langle \cdot, \cdot \rangle$ denotes the euclidean inner product between two matrices or vectors, defined as $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \operatorname{tr}\left(\boldsymbol{X}^T \boldsymbol{Y}\right)$.
$(\cdot)_+$ denotes the projection onto the non-negative orthant.
$\mathcal{R}(\boldsymbol{X})$ denotes the range of a matrix $\boldsymbol{X}$.
$\boldsymbol{I}_n$ is the $n \times n$ identity matrix.
$\mathbf{0}_{n \times p}$ and $\mathbf{1}_{n \times p}$ denote the $n \times p$ matrices of all-zeroes and all-ones respectively. We will omit the dimensions whenever they are clear from the context.

Spectrum of symmetric matrices

$\lambda_i(\boldsymbol{X})$ and $\boldsymbol{v}_i(\boldsymbol{X})$: the i-th eigenvalue and eigenvector of $\boldsymbol{X} \in \mathbb{S}^p$.
$\lambda_{\max}(\boldsymbol{X})$ and $\lambda_{\min}(\boldsymbol{X})$ are defined as $\max_i \lambda_i(\boldsymbol{X})$ and $\min_i \lambda_i(\boldsymbol{X})$, respectively.
$\aleph(\boldsymbol{X}) := \frac{\lambda_{\min}(\boldsymbol{X})}{\lambda_{\max}(\boldsymbol{X})}$ is the condition number of $\boldsymbol{X} \in \mathbb{S}^p$.
$\mathcal{V}_F(\boldsymbol{X})$ and $\Lambda_F(\boldsymbol{X})$ denote the $p \times F$ matrix and $F \times F$ diagonal matrix obtained from the first F eigenvectors and the first F eigenvalues of $\boldsymbol{X} \in \mathbb{S}^p$, considering the eigenvalues in decreasing order and $F \leq p$.

Functions:

$f(\boldsymbol{x})$: a generic real-valued function, $f : \mathbb{R}^p \to \mathbb{R}$.
$\boldsymbol{f}(\boldsymbol{x})$: a generic multi-dimensional real function, $\boldsymbol{f} : \mathbb{R}^p \to \mathbb{R}^n$.
∇f: the gradient of a differentiable real-valued function,

$$\nabla f = \left[\frac{df(\boldsymbol{x})}{dx_1}, \cdots, \frac{df(\boldsymbol{x})}{dx_p}\right]^T.$$

$\nabla^2 f$: the Hessian of a twice-differentiable real-valued function,

$$\left[\nabla^2 f\right]_{i,j} = \frac{d^2 f(\boldsymbol{x})}{dx_i dx_j}.$$

For any two real-valued functions $f(x)$, $g(x)$, we write

$$f(x) = \mathcal{O}(g(x)) \text{ if } \limsup_{x \to \infty} \left|\frac{f(x)}{g(x)}\right| < \infty$$
$$f(x) = o(g(x)) \text{ if } \lim_{x \to \infty} \frac{f(x)}{g(x)} = 0.$$

ABSTRACT

Learning from high-dimensional multivariate signals
by
Arnau Tibau Puig

Chair: Alfred O. Hero III

Modern measurement systems monitor a growing number of variables at low cost. In the problem of statistically characterizing the observed measurements, budget limitations usually constrain the number n of samples that one can acquire, leading to situations where the number p of variables is much larger than n. In this situation, classical statistical methods, founded on the assumption that n is large and p is fixed, fail both in theory and in practice. A successful approach to overcome this problem is to assume a parsimonious generative model characterized by a number k of free parameters, where k is much smaller than p.

In this dissertation we develop algorithms to fit low-dimensional generative models and extract relevant information from high-dimensional, multivariate signals. First, we define extensions of the well-known Scalar Shrinkage-Thresholding Operator, that we name Multidimensional and Generalized Shrinkage-Thresholding Operators, and show that these extensions arise in numerous algorithms for structured-sparse linear and non-linear regression. Using convex optimization techniques, we show that these operators, defined as the solutions to a class of convex, non-differentiable, optimization problems have an equivalent convex, low-dimensional reformulation. Our equivalence results shed light on the behavior of a general class of penalties that includes classical sparsity-inducing penalties such as the LASSO and the Group LASSO. In addition, our reformulation leads in some cases to new efficient algorithms for a variety of high-dimensional penalized estimation problems.

Second, we introduce two new classes of low-dimensional factor models that account for temporal shifts commonly occurring in multivariate signals. Our first con-

tribution, called Order Preserving Factor Analysis, can be seen as an extension of the non-negative, sparse matrix factorization model to allow for order-preserving temporal translations in the data. We develop an efficient descent algorithm to fit this model using techniques from convex and non-convex optimization. Our second contribution extends Principal Component Analysis to the analysis of observations suffering from arbitrary circular shifts, and we call it Misaligned Principal Component Analysis. We quantify the effect of the misalignments in the spectrum of the sample covariance matrix in the high-dimensional regime and develop simple algorithms to jointly estimate the principal components and the misalignment parameters.

All our algorithms are validated with both synthetic and real data. The real data is a high-dimensional longitudinal gene expression dataset obtained from blood samples of individuals inoculated by different types of viruses. Our results demonstrate the benefit of applying tailored, low-dimensional models to learn from high-dimensional multivariate temporal signals.

CHAPTER I

Introduction

1.1 From petal lengths to mRNA abundance

The text-book famous "Iris flower dataset" (Fis36) was collected by E. Anderson and was popularized by R.A. Fisher in 1936 to illustrate his method of Linear Discriminant Analysis. This dataset consists of 4 variables (Sepal length and width, Petal length and width) measured over 50 plants of three different species. To the 21st century statistics practitioner, a natural question is the following: if E. Anderson's intention was to characterize the morphological features of different species of Iris in a specific geographical area, why did he limit the number of variables to only four?

A plausible explanation is that E. Anderson was trying to strike a balance between the minimum number of features that could discriminate between species, and the number of different replicates he needed in order to obtain a representative sample. Since Anderson's time budget for data collection was probably limited, measuring an additional feature would have likely implied a reduction in the total number of replicates for each feature.

Fisher and Anderson's was an era where data collection was a manual or semi-automatized process, and the cost of measuring an additional variable was the same as the cost incurred in monitoring each of the previous ones. This was the data collection paradigm until the end of the 20th century: From medical research to communications systems, the number of measured variables was limited by the fact that the cost of the measurement system grew strongly with the number of observables. As a consequence, one had to limit the number of variables in order to allocate enough budget to the collection of replicates.

The advent of modern manufacturing techniques brought this limitation to an end. In essence, new technologies have allowed measurement and computing systems to do

economies of scale in the number of sensing devices. The cost of adding an additional measurement sensor decreases with the number of sensors already integrated, giving rise to measurement devices monitoring many orders of magnitude more variables than in the past[1].

This shift has profoundly changed the process of data collection and analysis: now it is not necessary anymore for the biologist, the astronomist, the marketing specialist or the antenna in the receptor of a communication system to know in advance what the relevant variables are in order to statistically characterize a physical process. Instead, one obtains measurements from a large pool of candidate features, and then relies on computational power to process the data and select the variables relevant to the study.

For example, in the data analysis problem that motivates this work, we are interested in extracting gene expression temporal patterns that drive the immune system response of a cohort to upper respiratory tract viral infections. Unfortunately, the specific genes that are involved in this process are unknown. During most of the past century, we would have had to use medical and biological a priori knowledge to determine a pool of candidate genes related to immune response and then perform costly and sensitive gene expression assays for each of the tissue samples. In contrast, modern Affymetrix mRNA microarray technology allows us to monitor tens of thousands of genes at low cost, and extract relevant information exclusively from the data using modern statistical techniques.

Unfortunately, the increase in the number of variables has not been accompanied by a proportional increase in the number of replicates or samples that one is able to record. As an example, very few mRNA microarray-based gene expression studies collect more than tens of replicates, usually due to budget constraints. This phenomenon is not limited to situations where the ratio between the available budget and the cost of each sample caps the number of available replicates. For instance, in wireless communications or in internet traffic data analysis (LBC$^+$06), the process under measurement is time-varying and hence one can only take few snapshots before violating the usual stationarity assumption. In conclusion, modern data sets are usually characterized by having a much larger number of variables (denoted by p)

[1]Quoting the great statistician Jerome H. Friedman (Fri01): "Twenty years ago most data was still collected manually. The cost of collecting it was proportional to the amount collected. [...] Now much (if not most) data is automatically recorded with computers. There is a very high initial cost [...] that is incurred before any data at all is taken. After the system is set up and working, the incremental expense of taking the data is only proportional to the cost of the magnetic medium on which it is recorded. This cost has been exponentially decreasing with time."

than replicates (denoted by n). This has a number of statistical and computational consequences, which we briefly explore in the following section.

1.2 Finding needles in a haystack

As the great statistician and applied mathematician David Donoho wrote (Don00),

"[...] we are in the era of massive automatic data collection, systematically obtaining many measurements, not knowing which ones will be relevant to the phenomenon of interest. Our task is to find a needle in a haystack, teasing the relevant information out of a vast pile of glut."

In other words, the naive intuition according to which "the more data, the better" seems to fail. For a fixed number of samples n, increasing the dimension of the observables, p, (by, say, adding more sensors to our measurement system) effectively increases the amount of data: $p \times n$. However, following Donoho's metaphor, an increase in p is equivalent to an increase in the size of the haystack which is not necessarily followed by an increase in the amount of needles. Indeed, there is practical and theoretical evidence (JL08; FFL08) showing that an increase in the ratio $\frac{p}{n}$ sometimes blurs or even completely suppresses the informative part of a noisy signal. This paradox is usually known as the "curse of dimensionality". It is also worth mentioning that there are other consequences of the $p \gg n$ regime which have been dubbed the "blessing" (as opposed to the "curse") of dimensionality. Indeed, recent results in probability and statistics show that there is much more structure in high dimensional random data than one would expect. Moreover, this structure only manifests itself in the high-dimensional setting, hence one can only *take advantage of it* in such regime. Examples of this phenomena are the concentration of measure (Mas07) of (well-behaved) functions of high-dimensional random variables around its mean, the convergence of the distribution of eigenvalues of large random matrices to a simple asymptotic distribution (Wig55), or the asymptotic uncorrelation phenomenon, by which large sequences of random variables behave as if their terms were uncorrelated, as the number of terms increases.

In order to overcome the curse of dimensionality, one popular approach is to assume that most physical phenomena can be characterized by only a few variables. For instance, in the Iris dataset example, E. Anderson *knew*, probably thanks to his training and experience, that petal and sepal dimensions are good discriminatory variables for the Iris subspecies. Among infinitely many other morphological features, E. Anderson chose those four in order to perform his study. In contrast, in modern

datasets one does not know a priori what the relevant features are, but it is often still reasonable to assume that only a handful of the variables, or low-dimensional linear combinations of them, are relevant. For instance, in a seminal paper (GST+99), Golub et al. showed that only 50 genes out of 6817 screened genes sufficed to build a simple classifier that discriminated between acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). This practical assumption is sometimes philosophically justified by the principle of parsimony, which is often (and perhaps wrongly) identified[2] with Occam's razor: "entia non sunt multiplicanda praeter necessitatem" (entities must not be multiplied beyond necessity.) Whether the parsimony assumption is accurate or useful to describe reality is an important epistemological question. In this dissertation we embrace K. Popper's view (Pop02), who argues that simpler models are preferable because they are better testable and less likely to be falsifiable[3].

1.3 Parsimonious statistical models

Mathematically, the notion of parsimony is formalized by assuming a low-dimensional generative model. In statistical learning, this is equivalent to restricting the class of probability distributions that model the observations. In this dissertation we adopt a parametric approach, which implies that the distribution classes we consider are completely characterized by a finite-dimensional set of parameters. Denoting by $p_{\boldsymbol{y}}(\boldsymbol{x};\boldsymbol{\theta})$ the joint distribution of a multivariate observation $\boldsymbol{y} \in \mathbb{R}^p$, we will assume that:

$$p_{\boldsymbol{y}}(\boldsymbol{x};\boldsymbol{\theta}) \in \left\{ f(\boldsymbol{x},\boldsymbol{\theta}) : \mathbb{R}^{2p} \to \mathbb{R}_+, \int f(\boldsymbol{x},\boldsymbol{\theta})\, d\boldsymbol{x} = 1, \boldsymbol{\theta} \in \mathcal{M} \right\}, \tag{1.1}$$

where $\mathcal{M} \subset \mathbb{R}^p$ is the parameter space, which we assume to be a low-dimensional manifold of $\mathbb{R}^p$. This characterization of the observations is theoretically attractive because for many families of $\mathcal{M}$ and $f(\boldsymbol{x},\boldsymbol{\theta})$, the problem of estimating $\boldsymbol{\theta}$ is computationally tractable and amenable to mathematical analysis. We proceed to illustrate a few instances of this general model that will appear throughout this dissertation.

- *Sparse Generalized Linear models:* Generalized Linear Models (GLM) are supervised learning models that characterize the distribution of a label or response

[2] In fact the history of what is often called "Occam's razor" might have a funny twist. W.H. Thorburn thoroughly argues in (Tho18) that "Occam's Razor is a modern myth. There is nothing medieval in it [...]". According to (Tho18), the often-quoted latin statement was never written by Occam and was first utilized much later, in 1639, by John Ponce of Cork.

[3] "Simple statements [...] are to be prized more highly than less simple ones because they tell us more; because their empirical content is greater; and because they are better testable." Chapter 7, Section 43 in (Pop02)

variable y as a function of a set of covariates, denoted by $\boldsymbol{x} \in \mathbf{R}^p$. In a Generalized Linear Model, $p_{y,\boldsymbol{x}}(y, \boldsymbol{x})$ is assumed to belong to the exponential family, and the mean of the response variable y is modeled as:

$$E(y|\boldsymbol{x}) = g^{-1}\left(\boldsymbol{x}^T\boldsymbol{\theta}\right)$$

where $g(x)$ is called the link function and $\boldsymbol{\theta} \in \mathcal{M}$ are the model parameters. The estimation of $\boldsymbol{\theta}$ is usually done by maximizing the likelihood of y given $\boldsymbol{x}$. In our context, we are interested in sparse GLMs, which means that $\mathcal{M}$ specializes to:

$$\mathcal{M}_{k-\text{sparse}} = \{\boldsymbol{\theta} \in \mathbf{R}^p : ||\boldsymbol{\theta}||_0 \leq k\} \tag{1.2}$$

where we define:

$$||\boldsymbol{\theta}||_0 := |\text{supp}(\boldsymbol{\theta})|,$$

and hence k is an upper bound on the number of non-zero elements in the parameter vector. The estimate of $\boldsymbol{\theta}$ in Sparse GLMs has to verify the model constraints, which leads to a constrained Maximum Likelihood (ML) problem. Given a collection of n independent observations $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$, the constrained ML estimator of $\boldsymbol{\theta}$ is defined as:

$$\begin{aligned} \hat{\boldsymbol{\theta}} := \quad & \arg \quad \max \sum_{i=1}^{n} p_{y|\boldsymbol{x}}(y_i, \boldsymbol{x}_i). \\ & \text{s.t.} \quad ||\boldsymbol{\theta}||_0 \leq k \end{aligned} \tag{1.3}$$

In this work we consider two types of generalized linear models, the linear regression and the logistic regression model. In the linear regression model, $g(x)$ is taken to be the identity and $p_{y|\boldsymbol{x}}(y, \boldsymbol{x})$ is the Gaussian distribution with unit variance. In such case, the ML estimate of θ is given by:

$$\begin{aligned} \hat{\boldsymbol{\theta}} = \quad & \arg \quad \min \sum_{i=1}^{n} \left|y_i - \boldsymbol{x}_i^T\boldsymbol{\theta}\right|^2 \\ & \text{s.t.} \quad ||\boldsymbol{\theta}||_0 \leq k, \end{aligned} \tag{1.4}$$

which is readily identified as the usual k-sparse signal recovery problem appear-

ing in inverse problems or compressive sensing (Tro06):

$$\begin{aligned} \hat{\boldsymbol{\theta}} = & \quad \arg \quad \min ||\boldsymbol{y} - \boldsymbol{D}\boldsymbol{\theta}||_2^2 \\ & \text{s.t.} \quad ||\boldsymbol{\theta}||_0 \leq k. \end{aligned} \tag{1.5}$$

Here $\boldsymbol{D}$ is a matrix of dictionary elements (a basis or an overcomplete dictionary) and $\boldsymbol{\theta}$ is the representation of the signal over this dictionary.

The second class of GLMs we will consider is the logistic regression model, where $p_{y|\boldsymbol{x}}(y, \boldsymbol{x})$ is the binomial distribution and $g(x)$ is the logit function (HTF05). In this case, the constrained MLE takes the form:

$$\begin{aligned} \hat{\boldsymbol{\theta}} = & \quad \arg \quad \min \sum_{i=1}^{n} y_i \boldsymbol{x}_i^T \boldsymbol{\theta} - \log\left(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\theta}}\right) \\ & \text{s.t.} \quad ||\boldsymbol{\theta}||_0 \leq k, \end{aligned} \tag{1.6}$$

We will consider variants of these Sparse GLM's in Chapter 2 and 3, in the application of finding groups of genes that discriminate the temporal responses over different populations.

- *Low-rank factor models*: The models considered in the last section characterized a response variable $\boldsymbol{y}$ as the output of a sparse linear model:

$$\boldsymbol{y} \approx \boldsymbol{D}\boldsymbol{\theta} \quad , ||\boldsymbol{\theta}||_0 \leq k,$$

where $\boldsymbol{\theta} \in \mathbf{R}^p$ is sparse but $\boldsymbol{D} \in \mathbf{R}^{n \times p}$ is usually full column rank and/or overcomplete, with p much larger than n. In contrast, in this section we consider low rank linear models of the type:

$$\boldsymbol{y} = \boldsymbol{F}\boldsymbol{a}^T + \boldsymbol{n}, \quad \text{rank}\,(\boldsymbol{F}) = f \ll p,$$

where p is the dimension of $\boldsymbol{y}$, $\boldsymbol{F} \in \mathbf{R}^{p \times f}$ is a low-rank matrix of factors with rank much smaller than the dimension of the ambient space, $\boldsymbol{a} \in \mathbf{R}^f$ is a vector of coefficients and $\boldsymbol{n} \in \mathbf{R}^p$ is a small, random residual error. Depending on the nature of the coefficient vector $\boldsymbol{a}$ and the residual $\boldsymbol{n}$, this model specializes to different paradigms. For instance, if $\boldsymbol{a}$ and $\boldsymbol{n}$ are assumed to be zero-mean Gaussian random vectors, with isotropic covariance and unit variance, we have

that:

$$p_{\boldsymbol{y}}\left(\boldsymbol{x};\boldsymbol{\Sigma}\right)=\frac{1}{\sqrt{2\pi}\left|\boldsymbol{\Sigma}\right|^{1/2}}\exp\left(-\frac{1}{2}\boldsymbol{x}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right),\quad\boldsymbol{\Sigma}\in\mathcal{M}_{f-\text{rank}},$$

where

$$\mathcal{M}_{f-\text{rank}}=\left\{\boldsymbol{\Sigma}\in\mathbf{R}^{p\times p}:\boldsymbol{\Sigma}=\boldsymbol{F}\boldsymbol{F}^{T}+\boldsymbol{I}\succ 0,\text{rank}\left(\boldsymbol{F}\right)=f\right\},\tag{1.7}$$

which is a low-dimensional subset of the cone of positive definite matrices. This model is also known as the Probabilistic PCA model (TB99). We will adapt a similar model to the problem of estimating temporal patterns from misaligned signals in Chapter 5.

Another class of low rank models stems from the assumption that $\boldsymbol{a}$ is non-random parameter of interest. In this case, assuming the residual $\boldsymbol{n}$ is centered and normal with isotropic covariance, we have

$$\begin{aligned}p_{\boldsymbol{y}}\left(\boldsymbol{x};\boldsymbol{H},\boldsymbol{a}\right) &= \frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left(-\frac{1}{2\sigma_n^2}\left(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{a}^{T}\right)^{T}\left(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{a}^{T}\right)\right),\\ &\quad\begin{cases}\boldsymbol{H}\in\mathcal{M}_H\subset\mathbf{R}^{p\times f}\\ \boldsymbol{a}\in\mathcal{M}_a\end{cases}.\end{aligned}$$

Under this model, a common estimate of $\boldsymbol{H}$ and $\boldsymbol{a}$ from observations $\{\boldsymbol{y}_i\}_{i=1}^n$ is the constrained Maximum Likelihood Estimator, defined as:

$$\begin{aligned}\hat{\boldsymbol{\theta}}= &\quad\arg\quad\min\sum_{i=1}^{n}\left|\left|\boldsymbol{y}_i-\boldsymbol{H}\boldsymbol{a}_i^{T}\right|\right|^2\\ &\text{s.t.}\quad\begin{cases}\boldsymbol{H}\in\mathcal{M}_H\\ \boldsymbol{a}_i\in\mathcal{M}_a,\quad i=1,\cdots,n.\end{cases}\end{aligned}\tag{1.8}$$

Depending on the specific choice of $\mathcal{M}_H$ and $\mathcal{M}_a$ this problem relates to k-means clustering (HTF05), Sparse Coding/Dictionary learning (OF97; KDMR+03), Non-Negative Matrix Factorization (LS99a) or Tensor Decomposition (KB09), to name a few. We develop a special instance of this class of models in Chapter 4, in the problem of estimating order-preserving temporal factors from longitudinal gene expression data.

In this dissertation we propose three different low-dimensional models for the analysis of high-dimensional, multivariate signals which are extensions or combina-

tions of the models listed above. Our models build on the special characteristics of high-dimensional multivariate signals, which we proceed to describe in the following section.

1.4 The special flavor of (high-dimensional) multivariate signals

We define a multivariate signal as a finite collection of multivariate random variables indexed by a set of increasing real numbers:

$$\left\{ \left[Y_1^t, Y_2^t, \cdots, Y_G^t \right] \in \mathbb{R}^G, t \in \left\{ \{t_1, \cdots, t_T\}, t_i \leq t_{i-1}, 1 \leq i \leq T \right\} \right\} \tag{1.9}$$

We will generally denote the realizations of the random variables (1.9) at a specific time point t by a row vector $\boldsymbol{y}_t^T$. A realization of the multivariate signal will be identified with a $T \times G$ matrix:

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \boldsymbol{y}_2^T \\ \cdots \\ \boldsymbol{y}_T^T \end{bmatrix}. \tag{1.10}$$

In general, we will work with a collection of S realizations of (1.9), which we identify with the vertical slices of a data cube, denoted by $\{\boldsymbol{Y}_s\}_{s=1}^S$ and shown in Figure 1.1.

The main difference between a multivariate signal and an ordinary collection of multivariate random vectors is the existence of an ordered structure, $t_i \leq t_{i-1}, 1 \leq i \leq T$[4]. The order assumption is important because it is closely related to the correlation structure of Y_i^t. Indeed, many time-varying physical processes exhibit some kind of regularity over time: continuity, smoothness or other. For example, in video data, one expects a certain degree of continuity between the images of adjacent frames. Another example is gene expression longitudinal data, e.g., the one we describe in the next section, where the expression values of a given gene are not expected to change abruptly across adjacent time points. Statistically, this regularity manifests itself in the form of a temporal correlation between Y_i^t and its temporal neighbors Y_i^{t-1}, Y_i^{t+1}, $\cdots$,Y_i^{t+f}. If we are to learn a statistical model from realizations $\{\boldsymbol{Y}_s\}_{s=1}^S$, it seems reasonable to enforce these properties in our estimators.

[4]This is more generally the definition of *longitudinal data*, which includes time series, spatial and others types of data.

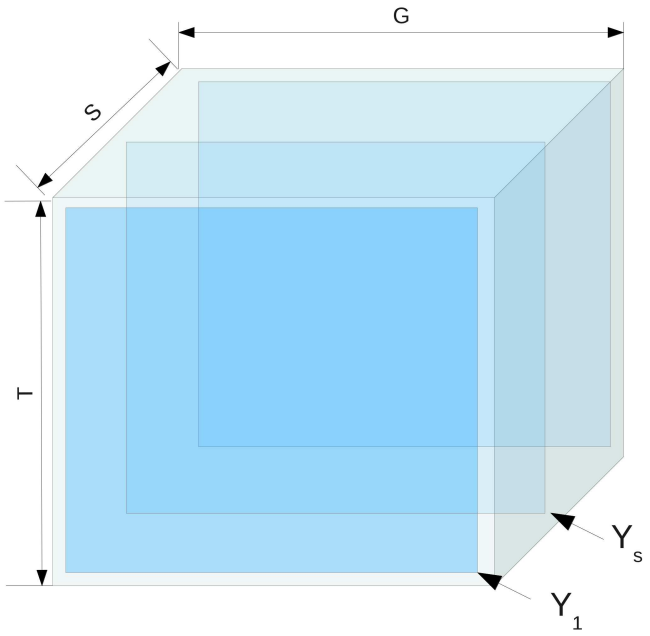


Figure 1.1: Multivariate signal data cube.

From the discussion above, it follows that we can interpret multivariate signals as a class of multivariate random variables for which we have a regularity prior along (at least) one of the dimensions. Unfortunately, this prior comes at the price of a type of sensitivity that is seldom taken into account. In many situations, one can not obtain replicates $\{\boldsymbol{Y}_s\}_{i=1}^n$ from (1.9), but rather a transformed version of them,

$$\boldsymbol{X}_s = \mathcal{T}_s\left(\boldsymbol{Y}_s\right). \tag{1.11}$$

It is obvious that except for the case where $\mathcal{T}_s\left(\cdot\right)$ is the identity operator, the statistical properties of $\boldsymbol{X}_s$ will be different from those of[5] $\boldsymbol{Y}_s$. Consequently, any estimator building on the temporal regularity of the underlying signal can be severely affected by operators that modify its temporal correlation structure, such as permutations or simple cyclic translations. More generally, the statistical properties of the signal along the time axis are not invariant to transformations that alter its ordered structure. We propose in this thesis two models that seek to compensate for the effects of two classes of transformations $\mathcal{T}_s\left(\cdot\right)$ that commonly occur in practice. In Chapter 4, we consider a transformation model that applies order-preserving circular shifts to the

[5]This is in general the case whenever the statistical properties of Y_s^t happen to be invariant with respect to the operator $\mathcal{T}_s\left(\cdot\right)$.

basis elements of a generative linear model. In Chapter 5, we consider the simpler class of cyclic translation operators, in which case $\mathcal{T}_s(\cdot)$ is a permutation matrix. This is a special case of the order-preserving model of Chapter 4, in which the shifts on each basis element are assumed to be the same.

There is another aspect of multivariate signals that we have not yet explored, and that has to do precisely with the "multivariate" part of the nomenclature. In many situations, the random variables Y_i^t represent particular features of a large complex system and hence correlation will exist not only along neighboring time points, but also across its multivariate structure (without restriction to neighboring indices in this case). For example, in the context of gene expression data analysis, it is well known that groups of genes belonging to the same signaling pathway exhibit similar expression patterns. Another example arises in the context of MIMO communication systems, which use correlation across signals received in different antennas at the receiver to take advantage of spatial diversity.

If one has such prior knowledge, it would be foolish not to use it to our advantage. In Chapter 2 and 3, we develop efficient optimization schemes to fit classes of Sparse GLMs that enforce a low-dimensional model constructed from a-priori knowledge of the underlying correlation structure.

We turn now to the description of the dataset which motivates most of the developments in this work.

1.5 Predictive Health and Disease (PHD)

Despite its relatively low mortality rate, Acute Respiratory Infections such as Rhinovirus (HRV), influenza (H3N2 and H1N1A), and respiratory syncytial virus (RSV) have an important societal and economical impact (ZCV$^+$09). Today's detection and classification of this family of diseases is largely based on physicians' diagnostic expertise, which relies on the assessment of the symptoms displayed by infected individuals. The DARPA-funded Predictive Health and Disease project[6] aims at developing novel detection and classification schemes for such pathologies *before* the symptoms appear, through the measurement of a pool of mRNA abundances in peripheral blood. The underlying assumption is that peripheral blood is a good proxy for the immune system response of an individual to a viral entity, which starts hours before symptoms appear.

[6] `http://www.darpa.mil/Our_Work/DSO/Programs/Predicting_Health_and_Disease_(PHD).aspx`

In order to study the feasibility of the project's goal, the PHD gene expression data set was collected as follows. After receiving appropriate Institutional Review Board approval, the authors in (ZCV$^+$09) performed separate challenge studies with two strains of influenza (H3N2 and H1N1), human rhino virus (HRV) and respiratory syncytial virus (RSV). For each such challenge study, roughly 20 healthy individuals were inoculated with one of the above viruses, and blood samples were collected at regular time intervals until the individuals were discharged. The blood obtained at each of these time points was assayed with Affymetrix Genechip mRNA microarray technology, yielding a matrix of gene expression values for each subject, such as the one depicted as a heatmap in Figure 1.3. Stacking each subject's gene expression matrix yields the data cube in Figure 1.2.

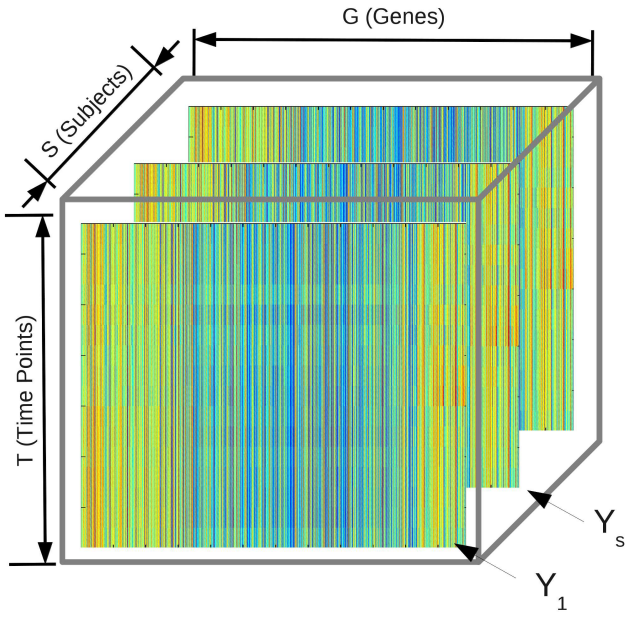


Figure 1.2: Predictive Health and Disease data cube.

The raw Genechip array data was pre-processed using robust multi-array analysis (IHC$^+$03) with quantile normalization (BIAS03). For each of the individuals and each time point, experienced physicians determined whether symptoms existed and correspondingly assigned a set of labels. These labels constitute the ground truth for our supervised learning tasks.

In this dissertation we address two major statistical challenges arising from the PHD project goals. The first one, which we address in Chapters 2 and 3, is to determine which genes are discriminatory of the different states of disease progression,

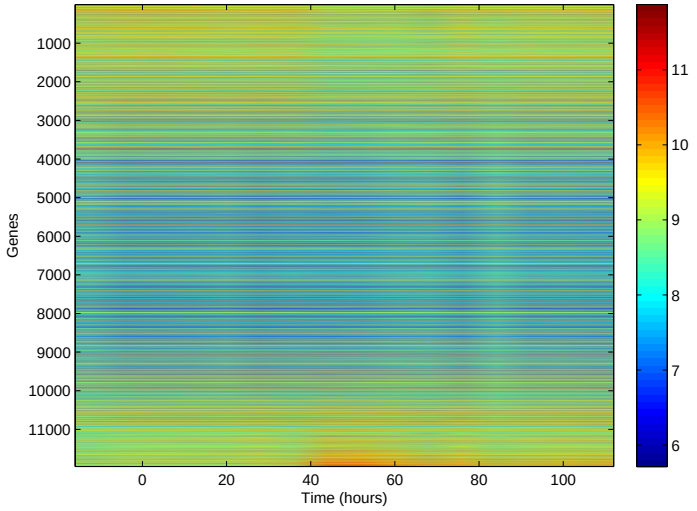


Figure 1.3: Heatmap of the logarithm of the 11961×16 normalized temporal gene expression matrix for a subject inoculated by H3N2 (influenza).

by using only early time samples with respect to the onset time, which is the time when the first symptoms are recorded. The number of samples available for this purpose is equal to the number of symptomatic subjects $S = 9$ times the number of time points used to train our predictors, which is smaller or equal than $T = 16$. On the other hand, the number of genes under our normalization scheme is equal to $G = 11961$, meaning that in this tasks we are in a high dimensional regime, where $\frac{\#\text{ variables}}{\#\text{ samples}} = \frac{G}{S \times T} \geq 83$.

Our second challenge, which we undertake in Chapters 4 and 5, is to discover temporal gene expression patterns that characterize the immune system response to viral infection. A quick glimpse at Figure 1.4, which plots the normalized expression values of 5000 genes with high temporal variability for a subject inoculated with H3N2, demonstrates that this is not an easy feat. In addition, as we will explore in Chapter 4, there is evidence that the immune system responses of each individual have different latencies, and hence the matrices $\boldsymbol{X}_s$ corresponding to each subject can not be taken as realizations from the same multivariate distribution. Instead, it will be convenient to model the observations $\boldsymbol{X}_s$ as in (1.11), where they are characterized as the result of applying a certain transformation to the common, underlying immune system expression response denoted by $\boldsymbol{Y}_s$.

We outline next the contributions of this dissertation towards achieving the aforementioned goals.

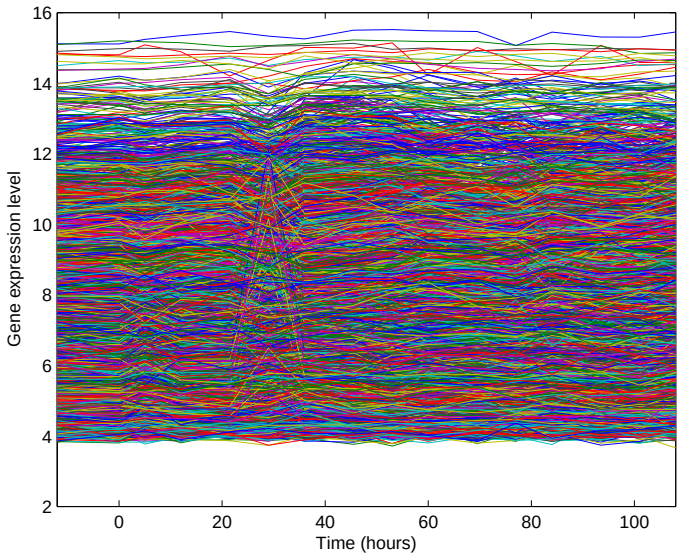


Figure 1.4: Expression values for 5000 genes with high relative temporal variability over 16 time points for a subject inoculated by H3N2. Underlying temporal patterns are clearly not visible due to noise and high-dimensionality.

1.6 Structure of this dissertation

In this section we relate the different chapters of this dissertation to the statistical learning problems associated to the Predictive Health and Diagnose project and dataset described above.

1.6.1 Penalized estimation and shrinkage-thresholding operators

We have seen in Section 1.2 that a number of estimation and approximation problems can be posed as a constrained optimization problem of the form:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\min \quad \mathcal{L}\left(\left\{\boldsymbol{x}^i\right\}_{i=1}^n, \boldsymbol{\theta}\right), \\ &\text{s.t.} \qquad \boldsymbol{\theta} \in \mathcal{M} \subset \mathbb{R}^p \end{aligned}$$

where $\mathcal{L}(\cdot,\cdot)$ is usually a *smooth and convex* loss function measuring the fit of the data to the model parameterized by $\boldsymbol{\theta}$, $\{\boldsymbol{x}^i\}_{i=1}^n$ are independent, identically distributed samples, and $\mathcal{M}$ denotes the subset of $\mathbb{R}^p$ characterizing the model constraints. In

particular, for the important class of k-sparse models, we had:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\min \quad \mathcal{L}\left(\left\{\boldsymbol{x}^i\right\}_{i=1}^n, \boldsymbol{\theta}\right) \\ &\text{s.t.} \qquad ||\boldsymbol{\theta}||_0 \leq k. \end{aligned} \tag{1.12}$$

For instance, when $\mathcal{L}$ is the squared ℓ_2 loss,

$$\mathcal{L}\left(\{\boldsymbol{x}_i\}_{i=1}^n, \boldsymbol{\theta}\right) = \sum_{i=1}^n \left|\left| x_1^i - \left[\begin{array}{cccc} x_2^i & x_3^i & \cdots & x_p^i \end{array}\right] \boldsymbol{\theta} \right|\right|_2^2$$

the solution to (1.12) yields the best k-sparse least squares approximation of x_1 from $x_2, x_3, \cdots, x_p$. It is reasonable to think that such an estimator would have better statistical properties than the unconstrained least squares (LS) estimator, specially in the high dimensional setting where $p > n$ and the least squares estimator is known to be inconsistent. (Note for instance that we can add any vector from the null-space of the design matrix to the LS solution without modifying the LS objective value).

Unfortunately, (1.12) is a combinatorial problem even when $\mathcal{L}(\cdot, \cdot)$ is the squared ℓ_2 loss, and no efficient computational method is known to solve it. In fact, it is known that the related problem of finding the smallest sparse approximation of a linear system is NP-Hard (Nat95).

One approach to circumvent the computational burden of solving (1.12) is to relax the combinatorial constraint $||\boldsymbol{\theta}||_0 \leq k$ to a *convex* constraint of the form $||\boldsymbol{\theta}||_1 \leq \tau$ (TBM79; Tib96; Tro06). Then one usually considers the penalized version of the constrained problem[7]:

$$\hat{\boldsymbol{\theta}} = \arg\min \quad \mathcal{L}\left(\left\{\boldsymbol{x}^i\right\}_{i=1}^n, \boldsymbol{\theta}\right) + \lambda ||\boldsymbol{\theta}||_1 \tag{1.13}$$

where λ is a parameter in one-to-one relationship with the sparsity parameter τ. Since the loss function $\mathcal{L}$ is usually convex (as in the GLM examples of Section 1.2), then (1.13) is a convex optimization problem. This seemingly simple relaxation has two surprising properties: (i) it transforms a combinatorial problem to a polynomially solvable one (for instance, by interior point methods which are known to have polynomial time complexity for self-concordant objectives (Wri)), and (ii) it has been shown that under relatively mild assumptions the solution to (1.12) and the solution to (1.13) are very close (Tro06).

[7] One can formalize the equivalence between the constrained and the penalized problems through Lagrange duality theory (BV).

The worst-case polynomial complexity of (1.13) represents a huge advantage with respect to the combinatorial nature of (1.12). Notwithstanding, in most modern statistical problems the dimension of the optimization domain, denoted by p, is often in the order of the tens or hundreds of thousands of variables. This precludes the usage of interior point (IP) methods which require the storage of $p \times p$ matrices and the solution of p-dimensional systems of equations. Numerous efforts have been devoted to developing algorithms that are better adapted to this large-dimensional regime. Roughly speaking, these algorithms sacrifice the fast convergence properties of IP methods in exchange of a very low per-iteration cost and sometimes weaker convergence guarantees[8] (DDDM04a; WNF09; CW06; BT09; BBC09).

In addition, the appropriate value of τ (or λ) is usually not known in advance. A common approach is to estimate this tuning parameter via cross-validation (HTF05), which requires the computation of the entire *solution path*, that is, the solution to (1.13) as a function of λ. For the least squares loss, homotopy algorithms (OPT00; EHJT04) take advantage of the piece-wise linearity of the solutions to (1.13) with respect to λ to efficiently compute the solution path at little cost.

The LASSO problem, which is the specialization of (1.13) to the least squares loss, has been proved effective for a variety of applications (MB06; THNC02; Can06). However, it is well known that consistency of the LASSO estimator is only possible under low correlation conditions which are not necessarily reasonable in practice (Bac08; NRWY10). To overcome this limitation, there has been a push to incorporate penalties that enforce structures other than simple sparsity while maintaining the convexity of the optimization problem. One way to incorporate these structures is to consider a generalization of the ℓ_1 penalized estimation problem which takes the following form:

$$\hat{\boldsymbol{\theta}} \quad = \arg\min \quad \mathcal{L}\left(\left\{\boldsymbol{x}^i\right\}_{i=1}^n, \boldsymbol{\theta}\right) + \lambda\Omega\left(\boldsymbol{\theta}\right), \tag{1.14}$$

where $\Omega\left(\boldsymbol{\theta}\right)$ is a sparsity-inducing penalty defined as:

$$\Omega\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{m} \sqrt{c_i}\left|\left|\boldsymbol{\theta}_{G_i}\right|\right|_2, \tag{1.15}$$

where $c_i > 0$, and G_i are subsets of indices such that $G_i \subseteq \{1, \cdots, p\}$ and $\cup_{i=1}^m G_i = \{1, \cdots, p\}$. This class of penalties enforce more fine-grained sparse models, in that

[8]By weak convergence we refer here to convergence of the sequence of objective values but not necessarily of the optimization parameters.

they require the complement of the support set of the solution $\hat{\boldsymbol{\theta}}$ to be the union of the groups of active subsets:

$$\operatorname{supp}\left(\hat{\boldsymbol{\theta}}\right) = \{1, \cdots, p\} \setminus \cup_{i \in \mathcal{A}} G_i, \tag{1.16}$$

for some $\mathcal{A} \subseteq \{1, \cdots, m\}$. When the groups are formed by individual indices, problem (1.14) specializes to the LASSO estimator, otherwise it is generally known as the Group LASSO, with (ZRY09; SRSE10b; JMOB10) or without overlap (YL06a). There is practical and theoretical evidence that such penalties are statistically superior to the ℓ_1 approach when the correlation structure of the data and the structure of the sparsity enforced by the penalty agree (OWJ08; SRSE10a; JOB10). It is worth mentioning, however, that this class of penalties is by no means exhaustive and that other schemes based on different convex functionals have been devised (Bac10; JOV09; CT07; NW08).

Unfortunately, the increase in modeling possibilities brought about by the penalties in (1.15) is accompanied by an increase in the computational burden for solving the associated penalized learning problem (1.14). On one hand, the proximal operator associated to the penalty (1.15), defined as:

$$\mathcal{P}_{\tau,\Omega}\left(\boldsymbol{x}\right) = \min_{\boldsymbol{\theta}} \frac{1}{2\tau} \left|\left|\boldsymbol{\theta} - \boldsymbol{x}\right|\right|_2^2 + \Omega\left(\boldsymbol{\theta}\right), \tag{1.17}$$

which constitutes the main building block for fast first order algorithms such as those in (ABDF11; WNF09; CW06; BT09; BBC09), is only easy to evaluate in the *separable* case, where the supports of the groups G_i do not overlap, or in the *hierarchical* case, where the supports only overlap in a hierarchical manner. On the other hand, in general, the solutions to (1.14) are no longer linear with respect to λ, rendering the computation of the entire regularisation path more computationally demanding than in the ℓ_1 case.

In the second and third chapter of this dissertation we address the aforementioned problems. First, we consider a generalization of (1.15) that incorporates the possibility of enforcing sparcity in a different coordinate system than the one where the regression fit is performed. This class of penalties is of the form:

$$\Omega\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{m} \sqrt{c_i} \left|\left|\boldsymbol{A}_{G_i,\cdot}\theta\right|\right|_2, \tag{1.18}$$

and, associated to them, we define the Generalized Shrinkage Thresholding Operator

(GSTO):

$$\mathcal{T}_{\lambda,\boldsymbol{H}}(\boldsymbol{g}) = \arg\min \quad \tfrac{1}{2}\boldsymbol{\theta}^T\boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{g}^T\boldsymbol{\theta} + \lambda\sum_{i=1}^m c_i \left|\left|\boldsymbol{A}_{G_i,\cdot}\theta\right|\right|_2. \tag{1.19}$$

In Chapter 2 and 3 we show that the GSTO arises naturally in a number of algorithms for group-sparse penalized linear and non-linear regression. For example, if $\boldsymbol{A}_{G_i,\cdot}$ are indicator matrices such that $[\boldsymbol{A}_{G_i,\cdot}]_{i,i} = 1$, $[\boldsymbol{A}_{G_i,\cdot}]_{k,l} = 0$ for $k, l \neq i$, and we let $\boldsymbol{H} = \boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{g} = -\boldsymbol{X}^T\boldsymbol{y}$, this problem specializes to the ℓ_1 penalized linear regression problem (1.13). Other choices of $\boldsymbol{A}_{G_i,\cdot}$ lead to group sparse solutions as in (1.16).

More generally, in analogy to the Scalar Shrinkage Thresholding Operator (SSTO), the GSTO shrinks or thresholds the input vector $\boldsymbol{g}$ and returns a thresholded vector in the following sense:

$$\mathcal{T}_{\lambda,\boldsymbol{H}}(\boldsymbol{g}) \in \mathrm{Ker}\,(\boldsymbol{A}_{D,\cdot}) \tag{1.20}$$

where $D = \cup_{i\in\mathcal{A}} G_i$ for some active set $\mathcal{A} \subseteq \{1, \cdots, m\}$. In addition, we show that the convex, non-differentiable problem (1.19) is approximately equivalent to a smooth (differentiable) optimization problem over a much smaller domain of dimension m (as opposed to p). This reformulation leads to efficient evaluations of Overlapping Group LASSO problem and the proximal operator (1.17), which in some cases outperform state-of-the-art first order methods. Finally, the smoothness of our reformulation over the active set of variables allows us derive first-order updates with respect to perturbations in λ, which enhance the computation of the regularization path over a grid of penalty parameters.

1.6.2 Order-Preserving Factor analysis

We have seen in Section 1.2 that one possible way to avoid overfitting in the high-dimensional regime is to enforce a low dimensional model. In unsupervised learning tasks, such as the one of extracting immune system-related temporal responses from gene expression data, a popular approach is to find a simultaneous factorization of the temporal slices of the data cube in Figure 1.2:

$$\boldsymbol{Y}_s \approx \boldsymbol{F}\boldsymbol{A}_s, \quad s = 1, \cdots, S. \tag{1.21}$$

Here $\boldsymbol{F}$ is a $T \times F$ factor (or loading) matrix common to all observations and $\boldsymbol{A}_s \in \mathbb{R}^{F\times G}$ are the coordinates or scores of the s-th observation on the factor matrix. In low-rank factor models, we have $F << G$ and choose F by Cross-Validation

over a test set of entries that were treated as missing during the training phase. In the application of this model to the gene expression example, $\boldsymbol{F}$ contains the temporal patterns that explain away the correlations between different genes, and each $\boldsymbol{A}_s$ describes the association of each subject's genes with the temporal patterns in $\boldsymbol{F}$. Such a simultaneous matrix decomposition achieves several goals: it enhances interpretability, it reduces the variance due to noise and it can be useful in imputating potentially missing entries. The underlying assumption here is that there is strong correlation among the columns of $\boldsymbol{Y}_s$, and that this correlation is persistent for all $s = 1, \cdots, S$, hence a few prototype patterns constructed from the columns of $\boldsymbol{F}$ are enough to approximate well the data. This is a reasonable assumption for multivariate signals, as we explained in Section 1.4.

Despite its relatively small number of degrees of freedom, factor models such as (1.21) often suffer from identifiability issues. Notice for instance that there is a scale ambiguity within $\boldsymbol{F}$ and $\boldsymbol{A}_s$. In addition, the null space of $\boldsymbol{F}$ is relatively high-dimensional and hence there exist multiple ways to represent $\boldsymbol{Y}_s$ on $\boldsymbol{F}$. To escape from these issues while maintaining the advantages of the low-rank structure, several authors have proposed to add additional constraints to $\boldsymbol{F}$ and $\{\boldsymbol{A}_s\}_{s=1}^{S}$. These include, non-negativity (LS99a), sparsity (JOB10; WTH09) or both. These approaches work well in practice, and they have also been theoretically justified under a rather stringent setting (DS04).

Unhappily, in many cases, models of the form (1.21) fall short and fail to accommodate meaningful intersubject variability within the data matrices. In the context of multivariate signals, a major problem is that outlined in Section 1.4, where we only have access to a transformed version of the data we are interested in. For example, in the analysis of immune system-related temporal patterns from gene expression data, we need to account for the natural temporal variability across different subjects. As we show in Chapter 4, this variability manifests itself as a difference in the temporal latencies each individual shows after viral infection. Indeed, the sequence of gene expression responses of each individual is similar, but the moment at which the responses occur vary by up to 24 hours. This motivates the following order-preserving factor model,

$$\boldsymbol{Y}_s \approx \boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right) \boldsymbol{A}_s, \quad s = 1, \cdots, S. \tag{1.22}$$

where $\boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}\right)$ is a matrix valued function that applies a circular shift to each column of $\boldsymbol{F}$ according to the vector of shift parameters $\boldsymbol{d}$, and $\boldsymbol{d}$ is a set of order-preserving

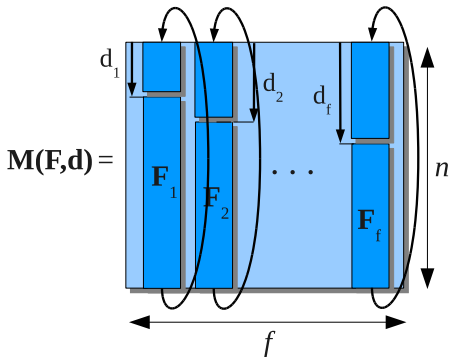


Figure 1.5: Construction of the factor matrix $\boldsymbol{M}(\boldsymbol{F}, \boldsymbol{d})$ by applying a circular shift to a common set of factors $\boldsymbol{F}$ parameterized by a vector $\boldsymbol{d}$.

latencies (see Figure 1.5 for an explanatory diagram). This is an extension of the usual factor model, to which it specializes by constraining $\boldsymbol{d} = \boldsymbol{0}$.

The complexity of this model increases with respect to the simpler model in (1.21), and so does the computational effort required to fit it. In Chapter 4, we propose a block coordinate descent algorithm to fit the order preserving model (1.22) according to a least squares criteria. Special care is devoted to the step concerning the estimation of the order-preserving $\boldsymbol{d}$, which is a non-convex global optimization problem. We take advantage of the structure of the least squares objective to design a fast branch-and-bound procedure to solve this problem that avoids a potentially costly exhaustive search.

Our algorithm, combining convex and non-convex optimization techniques leads to fitting a sparse, non-negative order-preserving factor model in only a matter of minutes, Cross-Validation of all tuning parameters included. Our OPFA decompositions are shown to outperform simpler sparse factor analysis models when order-preserving misalignments are present. We also show how the OPFA decomposition is a valuable tool for extracting order-preserving patterns that are related to the immune system response to viral infection in symptomatic individuals inoculated by influenza.

1.6.3 Misaligned Principal Component Analysis

One of the multiple advantages of the digital age is that rare and old scientific documents are only a few clicks away from the comfort of our working space. Perhaps surprisingly, this now easily accessible bibliographical evidence demonstrates that many of the challenges that occupy today's statisticians and engineers are anything but new. One such example is found in the work of Karl Pearson, a British mathematician credited with the invention of a number of classical statistical tools.

In his 1901 paper "On lines and planes of closest fit to systems of points in the space" (Pea01), K. Pearson considered the problem of representing a systems of points in a high dimensional vector space by the means of a low-dimensional subspace. His solution to this problem, preceding Hotelling's work (Hot33) on the subject by a few decades, was based on computing the leading eigenvector of the correlation matrix, and is currently known as Principal Component Analysis (PCA).

PCA is nowadays routinely used as a dimensionality reduction method, for interpretation, compression or representation purposes, and its multiple variants are still the subject of current research. For instance, significant efforts have been devoted to incorporating prior information to the PCA estimates, in the form of smoothness for functional PCA (Ram97), or in the form of sparsity in the eigenvector estimates in order to enhance interpretability (JL08; dEGJL07; WTH09).

In the problem that concerns us, the system of points is in fact a collection of realizations of a (possibly multivariate) signal. Estimating a low-dimensional subspace amounts then to finding latent temporal patterns that characterize the observations. As in the previous section, this is equivalent to the problem of finding an approximate decomposition:

$$\boldsymbol{Y}_s \approx \boldsymbol{F}\boldsymbol{A}_s, \quad s = 1, \cdots, S, \tag{1.23}$$

with the exception that in the PCA framework, one assumes that $\boldsymbol{A}_s$ is a random matrix with i.i.d. Gaussian elements. This assumption enables the estimation of $\boldsymbol{F}$ through the covariance of $\boldsymbol{Y}_s$, decreasing the computational complexity with respect to the Order Preserving Factor Analysis problem of Chapter 4 and rendering the problem amenable to mathematical analysis. Analogously to the previous section, when $\boldsymbol{Y}_s$ contains correlated gene expression temporal responses, PCA yields a basis for the temporal patterns that are common across genes.

It is well known that PCA can be interpreted as the Maximum Likelihood estimate of the covariance matrix under a gaussian, low-rank factor model assumption (TB99) such as the one in (1.7). Given a large number of independent identically distributed observations and fixing the dimension, the MLE is known to be asymptotically consistent, and hence so is the PCA estimate. For multivariate signals, as we suggested in Section 1.4, independent identically distributed (i.i.d.) replicates of a signal are not always easy to obtain. Specifically, it is common to have misalignments between batches of observations, due to sampling or physical limitations. In such case, we find ourselves in the situation modeled by equation (1.11), where the

observations are not i.i.d. and hence consistency can no longer be expected to hold. A straightforward way to overcome this limitation, at least in the asymptotic regime, is to compute the PCA estimate separately for each batch of observations having the same degree of misalignment and then align the PCA estimates together to construct a global estimate.

In the high dimensional regime, there is a more subtle and negative effect of misalignments that sometimes will hinder this straightforward approach. In order to understand it, we will consider the case where there is no misaligment, but the number of variables p and the number of samples n are of the same order of magnitude. In this regime, as illustrated by Figure 1.6 in a 2-dimensional example, the PCA estimate can be severely off unless the Signal-to-Noise Ratio (SNR) is high enough. Intuitively, when one has only a few noisy observations, the number of subspaces which span the directions of higher variance is large unless the observations are very well aligned, which only happens when the SNR is high enough. This phenomenon is known as a phase transition effect, meaning that the estimation goes from being impossible to practically perfect by increasing the SNR by only a few dBs (Pau07; BBAP05). As a consequence, computing a PCA estimate for each batch of misaligned data is not a good approach when the SNR or the number of replicates is small.

In Chapter 5, we will use recent developments in the characterization of the eigenvectors of random Wishart matrices to show that misalignments increase the phase transition SNR from which estimation is possible. We also asymptotically quantify this degradation as a function of a few parameters related to the underlying signal spectrum. Our results highlight the advantage of considering all the observations together, *despite* the misalignments, whenever the SNR is close to the phase transition point.

These results will also motivate us to consider the Misaligned PCA (MisPCA) problem of simultaneously estimating the principal components and the misalignment parameters. Unfortunately, this problem is combinatorial in nature, and the search space grows exponentially fast with the number of misaligned observations. We will propose instead two simple algorithms that approximate the MisPCA solution, while offering substantial advantage with respect to the traditional PCA estimator.

1.7 Publications

The work presented in this dissertation has led to the following publications.

Journals:

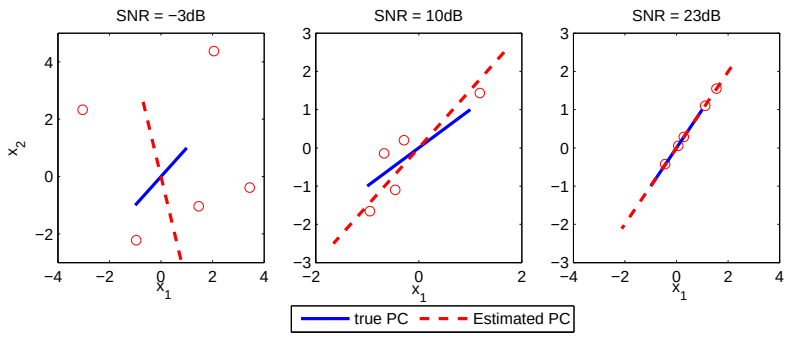


Figure 1.6: Estimated and true principal components, for 2 variables x_1 and x_2 and 5 samples, and increasing SNR. It is clear that the PCA estimate is pretty accurate at 10dBs, while it is almost orthogonal to the true one at -3dBs.

1. L. Carin, A.O. Hero, J. Lucas, D. Dunson, M. Chen, R. Henao, A. Tibau-Puig, A. Zaas, C.W. Woods, and G.S. Ginsburg, "High Dimensional Longitudinal Genomic Data: An analysis used for monitoring viral infections," IEEE Signal Processing Magazine, vol. 29, no. 1, pp. 108-123, Jan. 2012.

2. A. Tibau-Puig, A. Wiesel, A. Zaas, C. W. Woods, G. S. Ginsburg, G. Fleury and A. O. Hero, "Order-preserving factor analysis - application to longitudinal gene expression", IEEE Transactions on Signal Processing, vol. 59, no. 9, pp. 4447-4458, Sept. 2011.

3. A. Tibau-Puig, A. Wiesel, G. Fleury and A. O. Hero, "Multidimensional shrinkage-thresholding operator and Group LASSO penalties," IEEE Signal Processing Letters, vol. 18, no. 6, pp. 363 - 366, Jun. 2011.

Manuscripts in preparation:

4 A. Tibau-Puig and A. O. Hero, "Misaligned Principal Component Analysis", December 2011.

5 A. Tibau-Puig and A. O. Hero, "Generalized Shrinkage-Thresholding Operators", December 2011.

Conference proceedings:

6 A. Tibau-Puig, A. Wiesel, R. R. Nadakuditi and A. O. Hero, "Misaligned Principal Component Analysis (Mis-PCA) for Gene Expression Time Series Analysis", Asilomar Conference on Signals, Systems, and Computers, Pacific Groove, CA, Nov. 2011.

7 A. Tibau-Puig, A. Wiesel, A. Zaas, C. W. Woods, G. S. Ginsburg, G. Fleury and A. O. Hero, "Order-preserving factor discovery from misaligned data," Sensor Array and Multichannel Signal Processing Workshop (SAM), 2010 IEEE, pp.209-212, 4-7 Oct. 2010.

8 A. Tibau-Puig, A. Wiesel, and A. O. Hero, "A multidimensional shrinkage-thresholding operator," Statistical Signal Processing, 2009. SSP '09. IEEE/SP 15th Workshop on, pp.113-116, Aug. 31 2009-Sept.3 2009.

CHAPTER II

A multidimensional shrinkage-thresholding operator

2.1 Introduction

Variable selection is a crucial step in modern statistical signal processing, where oftentimes the number of variables largely exceeds the number of available samples. In genomic signal processing, for instance, RNA microarray data consists of gene expression levels for tens of thousands of genes, while the number of available samples rarely exceeds the hundreds. In this sample-starved situation, it is common practice to perform a preprocessing step to select the variables that are most relevant with respect to the biological process under study (THNC03). The scalar shrinkage-threshold operator is central to variable selection algorithms such as Iterative Thresholding (DDDM04a) for image deblurring (BT09), wavelet-based deconvolution (NF01) or predictive analysis of gene expression microarrays (THNC03).

In this chapter, we first introduce a multidimensional generalization of the scalar shrinkage thresholding operator. We define this operator as the minimization of a convex quadratic form plus a (non-squared) Euclidean (ℓ_2) norm penalty. We analyze this non-differentiable optimization problem and discuss its properties. In particular, in analogy to the scalar shrinkage operator, we show that this generalization yields a Multidimensional Shrinkage Thresholding Operator (MSTO) which takes a vector as an input and shrinks it or thresholds it depending on its Euclidean norm. Our results relies on a reformulation of the problem as a constrained quadratic problem with a conic constraint. Using conic duality theory, we transform this multidimensional optimization problem into a simple line search which can be efficiently implemented. We propose a simple algorithm to evaluate the MSTO and show by simulations that it outperforms other state-of-the-art algorithms.

In the second part of this chapter we discuss applications of the MSTO to several estimation problems. First, we consider the Euclidean-norm penalized least squares and discuss its relation to ridge regression (TAJ77) and robust regression (EGL97). Using the MSTO formulation, we show that this problem leads to a solution which is either the zero vector or the ridge-penalized least squares solution where the optimal shrinkage is chosen through a line search.

The second application we consider is the problem of estimating the mean of a Gaussian distribution under a block diagonal covariance and block sparse structure. This is a variant of the Regularized Linear Discriminant Analysis (RLDA) problem which seeks to construct a linear classifier using as little variables as possible (TP07), (HTF05). We give an exact solution to this estimation problem in terms of the MSTO and give an application of Block Sparse RLDA to the problem of selecting genes that classify two different populations across different time points. Our implementation using the MSTO allows to jointly process several RNA microarrays in a matter of minutes.

Finally, we consider two applications in the context of group-sparsity penalized regression, with disjoint and non-disjoint, hierarchical groups. This class of problems appears in many signal processing applications where the problem suggests enforcing a structured-sparse estimate rather than a simple sparse estimate. Examples of this situation occur in spectrum cartography for cognitive radio (BMG10), jointly-sparse signal recovery (WDS$^+$05), regression with grouped variables (YL06a), source localization (MCW05) or model-based compressive sensing (SRSE10b). We give a block-wise descent algorithm for group-sparse linear regression and show that the MSTO arises naturally in fast proximal algorithms for large-scale non-differentiable convex optimization.

This chapter is organized as follows. In Section 2.2, we first define the MSTO and introduce our first theoretical result. Second, we discuss how to efficiently evaluate the MSTO. We illustrate applications of the MSTO in statistical signal processing problems in Section 2.3. In Section 2.4 we present numerical experiments and an application of the MSTO to the problem of finding genes whose time course discriminate two population of individuals. We finally conclude the chapter in Section 2.5.

2.2 Multidimensional Shrinkage-Thresholding Operator

The scalar shrinkage-thresholding operator is usually defined as:

$$\begin{aligned} \mathcal{T}_{\lambda,h}(g) &:= \arg\min_x \frac{1}{2}hx^2 + gx + \lambda|x| \\ &= \begin{cases} -\frac{|g|-\lambda}{h}\text{sign}(g) & \text{if } |g| > \lambda \\ 0 & \text{otherwise.} \end{cases}, \end{aligned} \tag{2.1}$$

where $h, \lambda > 0$ and $g \in \mathbb{R}$. This operator takes a scalar g as an input and thresholds or shrinks its magnitude. A natural generalization is the following *Multidimensional Shrinkage Thresholding Operator* (MSTO):

$$\mathcal{T}_{\lambda,\boldsymbol{H}}(\boldsymbol{g}) := \arg\min_{\boldsymbol{x}} \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} + \lambda\|\boldsymbol{x}\|_2, \tag{2.2}$$

where $\boldsymbol{H} \in \mathbb{S}_+^N$, $\lambda > 0$ and $\boldsymbol{g} \in \mathbb{R}^N$. This is a convex optimization problem and can be cast as a standard Second Order Cone Program (SOCP) (LVBL98):

$$\begin{aligned} \min \quad & \tfrac{1}{2}t_1 + \lambda t_2 + \boldsymbol{g}^T\boldsymbol{x} \\ \text{s.t.} \quad & \begin{bmatrix} \boldsymbol{V}^T\boldsymbol{x} \\ \frac{t_1-1}{2} \\ \frac{t_1+1}{2} \end{bmatrix} \succeq_K 0, \quad \begin{bmatrix} \boldsymbol{x} \\ t_2 \end{bmatrix} \succeq_K 0, \end{aligned} \tag{2.3}$$

where $\boldsymbol{V}$ is such that $\boldsymbol{H} = \boldsymbol{V}\boldsymbol{V}^T$. SOCPs can be solved efficiently using interior point methods (LVBL98). The next theorem shows that, as in the scalar case, the MSTO shrinks or thresholds the norm of the input vector $\boldsymbol{g}$ and that the corresponding SOCP (2.3) can be solved using a simple line search.

Theorem II.1. *Let $N < \infty$, $\boldsymbol{H} \in \mathbb{S}_+^N$, $\boldsymbol{W} \in \mathbb{S}_{++}^N$, $\boldsymbol{g} \in \mathcal{R}(\boldsymbol{H})$ and $\lambda > 0$. The optimal value of the N-dimensional, non-differentiable problem:*

$$\min_{\boldsymbol{x}} \frac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} + \lambda\|\boldsymbol{W}\boldsymbol{x}\|_2 \tag{2.4}$$

is equal to the optimal value of the convex one-dimensional problem:

$$\min_{\eta \geq 0} \quad \left(1 - \tfrac{1}{2}\boldsymbol{g}^T\boldsymbol{B}^{-1}(\eta)\,\boldsymbol{g}\right)\eta, \tag{2.5}$$

where

$$\boldsymbol{B}(\eta) := \eta\boldsymbol{H} + \frac{\lambda^2}{2}\boldsymbol{W}^2 \succ 0, \tag{2.6}$$

and the solutions of (2.4) and (2.5) are related by:

$$\mathcal{T}_{\lambda,\boldsymbol{H}}(\boldsymbol{g}) = \begin{cases} -\eta \boldsymbol{B}^{-1}(\eta^*)\,\boldsymbol{g} & \text{if } \|\boldsymbol{W}^{-1}\boldsymbol{g}\|_2 > \lambda \\ \mathbf{0} & \text{otherwise.} \end{cases}, \tag{2.7}$$

where η^* is the solution to (2.5). Furthermore, if $\lambda_{\min}(\boldsymbol{H}) > 0$, the solution to (2.5) satisfies:

$$\eta \in \frac{\lambda}{2}\left(\|\boldsymbol{W}^{-1}\boldsymbol{g}\|_2 - \lambda\right)\left[\frac{1}{\lambda_{\max}(\boldsymbol{W}^{-1}\boldsymbol{H}\boldsymbol{W}^{-1})}, \frac{1}{\lambda_{\min}(\boldsymbol{W}^{-1}\boldsymbol{H}\boldsymbol{W}^{-1})}\right] \tag{2.8}$$

Proof. Let us assume momentarily that $\boldsymbol{W} = \boldsymbol{I}$. Since $\boldsymbol{H} \succeq 0$ and $\|\cdot\|_2$ is a norm, it follows that $\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x}$ and $\|\boldsymbol{x}\|_2$ are convex functions of $\boldsymbol{x}$. Also, (2.4) is equivalent to the following quadratic program with a second order conic constraint:

$$\begin{aligned} \min_{x,t} \quad & \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} + t \\ \text{s.t.} \quad & \begin{bmatrix} -\lambda\boldsymbol{x} \\ -t \end{bmatrix} \preceq_K 0. \end{aligned} \tag{2.9}$$

Slater's condition for generalized inequalities is verified and strong duality holds. Since K is self-dual, the conic dual can be written as ((BV), Section 5.9.1):

$$\max q(\boldsymbol{u}, \mu) \quad \text{s.t.} \quad \begin{bmatrix} \boldsymbol{u} \\ \mu \end{bmatrix} \succeq_K 0, \tag{2.10}$$

where the dual function is defined as

$$q(\boldsymbol{u}, \mu) = \min_{x,t} \frac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} + t - \boldsymbol{u}^T(\lambda\boldsymbol{x}) - \mu t. \tag{2.11}$$

This inner minimization is unbounded in t unless $\mu = 1$ and in $\boldsymbol{x}$ unless $\boldsymbol{u} \in \mathcal{R}(\boldsymbol{H})$. Otherwise, its optimum satisfies:

$$\boldsymbol{x} = -\boldsymbol{H}^{\dagger}(\boldsymbol{g} - \lambda\boldsymbol{u}). \tag{2.12}$$

Plugging (2.12) in (2.10), and using the fact that a non differentiable dual conic constraint $\begin{bmatrix} \boldsymbol{u}^T, & 1 \end{bmatrix}^T \succeq_K 0$ is equivalent to a standard quadratic constraint $\|\boldsymbol{u}\|_2^2 \leq$

1, we obtain the following dual concave maximization:

$$\max_{\|\boldsymbol{u}\|_2^2 \leq 1, \boldsymbol{u} \in \mathcal{R}(\boldsymbol{H})} -\frac{1}{2} (\boldsymbol{g} - \lambda \boldsymbol{u})^T \boldsymbol{H}^\dagger (\boldsymbol{g} - \lambda \boldsymbol{u}). \tag{2.13}$$

The standard Lagrange dual of this problem is:

$$\min_{\eta \geq 0} \max_{\boldsymbol{u} \in \mathcal{R}(\boldsymbol{H})} -\frac{1}{2} (\boldsymbol{g} - \lambda \boldsymbol{u})^T \boldsymbol{H}^\dagger (\boldsymbol{g} - \lambda \boldsymbol{u}) - \eta \left(\boldsymbol{u}^T \boldsymbol{u} - 1\right). \tag{2.14}$$

Since $\boldsymbol{H} \succeq 0$ and $\boldsymbol{H}^\dagger \boldsymbol{g} \in \mathcal{R}\left(\boldsymbol{H}^\dagger\right)$, the inner maximization is a simple quadratic problem in $\boldsymbol{u}$ with solution:

$$\boldsymbol{u} = \frac{\lambda}{2} \boldsymbol{B}^{-1}(\eta) \boldsymbol{g}, \tag{2.15}$$

where $\boldsymbol{B}(\eta)$ is defined in (2.6). This leads to the following line search over the Lagrange multiplier η:

$$\min_{\eta \geq 0} \left(1 - \frac{1}{2} \boldsymbol{g}^T \boldsymbol{B}^{-1}(\eta) \boldsymbol{g}\right) \eta, \tag{2.16}$$

which proves the equivalence between (2.4) and (2.5) and is convex by Lagrange's duality properties.

The eigenvalues of $\boldsymbol{B}^{-1}(\eta)$ are real and can be characterized as:

$$\lambda_i \left(\boldsymbol{B}^{-1}(\eta)\right) = \frac{1}{\eta \lambda_i(\boldsymbol{H}) + \frac{\lambda^2}{2}}. \tag{2.17}$$

Since $\eta \geq 0$, $\lambda_i(\boldsymbol{H}) \geq 0$ and $\lambda > 0$, it holds that $0 < \lambda_i\left(\boldsymbol{B}^{-1}(\eta)\right) \leq \frac{2}{\lambda^2}$. Therefore, if $\|\boldsymbol{g}\|_2 \leq \lambda$ then $\frac{1}{2} \boldsymbol{g}^T \boldsymbol{B}^{-1}(\eta) \boldsymbol{g} \leq 1$ and

$$\eta \left(1 - \frac{1}{2} \boldsymbol{g}^T \boldsymbol{B}^{-1}(\eta) \boldsymbol{g}\right) \geq 0. \tag{2.18}$$

This implies that if $\|\boldsymbol{g}\|_2 \leq \lambda$ the minimum in (2.16) is attained by choosing $\eta = 0$. Plugging (2.15) into (2.12) yields (2.7). Using this and plugging (2.15) in (2.12) yields (2.7). To obtain the bounds on the solution to (2.16), we let $\boldsymbol{u} = \frac{\lambda}{2} \boldsymbol{B}^{-1}(\eta) \boldsymbol{g}$ and use the following inequalities:

$$\frac{\lambda}{2} \frac{\|\boldsymbol{g}\|_2}{\eta \lambda_{\min}(\boldsymbol{H}) + \frac{\lambda^2}{2}} \geq ||\boldsymbol{u}||_2 \geq \frac{\lambda}{2} \frac{\|\boldsymbol{g}\|_2}{\eta \lambda_{\max}(\boldsymbol{H}) + \frac{\lambda^2}{2}}. \tag{2.19}$$

Since we have assumed $\|\boldsymbol{g}\|_2 > \lambda$, $\boldsymbol{u}$ solving (2.13) has to verify the complementary slackness condition, namely $\|\boldsymbol{u}\|_2^2 = 1$. Setting $\|\boldsymbol{u}\|_2 = 1$ in the inequalities above yields the following bounds in η:

$$\eta \in \frac{\lambda}{2}\left(\|\boldsymbol{g}\|_2 - \lambda\right)\left[\frac{1}{\lambda_{\max}\left(\boldsymbol{H}\right)}, \frac{1}{\lambda_{\min}\left(\boldsymbol{H}\right)}\right] \tag{2.20}$$

where we define $\frac{1}{0} = \infty$ if $\lambda_{\min}\left(\boldsymbol{H}\right) = 0$. This concludes the proof when $\boldsymbol{W} = \boldsymbol{I}$. To extend the results to a general $\boldsymbol{W} \succ 0$, we just need to use the bijective change of variables $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ and solve (2.4) with respect to $\boldsymbol{y}$. Applying the results above and undoing the change of variable finalizes the proof. □

2.2.1 Evaluating the MSTO

According to Theorem II.1, evaluating the MSTO reduces to solving (2.5) when $\|\boldsymbol{g}\|_2 > \lambda$. In the special case where $\boldsymbol{H} = k\boldsymbol{I}$ for some $k > 0$, the optimality condition for (2.5) leads to a simple solution for its positive root:

$$\eta^* = \frac{\lambda}{2k}\left(\|\boldsymbol{g}\|_2 - \lambda\right), \tag{2.21}$$

which yields the following closed form expression for the MSTO:

$$\mathcal{T}_{\lambda,k\boldsymbol{I}}\left(\boldsymbol{g}\right) = -\frac{1}{k}\left(\|\boldsymbol{g}\|_2 - \lambda\right)_+ \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}. \tag{2.22}$$

where $\left(x\right)_+ = \max\left(x, 0\right)$. This is equivalent to (2.1) if we define the multidimensional sign function as $\operatorname{sign}\left(\boldsymbol{x}\right) = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}$ and coincides with the vectorial soft-threshold in (WNF09). If $\boldsymbol{H} \neq k\boldsymbol{I}$ and $\|\boldsymbol{g}\|_2 > \lambda$, evaluating the MSTO is non trivial and requires the numerical solution of the line-search in (2.5). In particular, we propose to use a Projected Newton approach with Goldstein step-length rule (Dun80) which incorporates the advantages of second order methods while respecting the constraint $\eta \geq 0$ in (2.5). Let

$$w\left(\eta\right) \;\; := \left(1 - \tfrac{1}{2}\boldsymbol{g}^T\boldsymbol{B}^{-1}\left(\eta\right)\boldsymbol{g}\right)\eta,$$

where $\boldsymbol{B}(\eta)$ is defined in (2.6). At iteration t, the Goldstein Projected Newton iteration for problem (2.5) is given by (Dun80):

$$\begin{aligned} \hat{\eta}^t &= \left(\eta^t - \frac{w'(\eta^t)}{w''(\eta^t)}\right)_+, \\ \eta^{t+1} &= \eta^t + \omega_n\left(\hat{\eta}^t - \eta^t\right), \end{aligned} \tag{2.23}$$

where $w'(\eta)$, $w''(\eta)$ are the first and second derivatives of $w(\eta)$ respectively. Letting $\delta \in (0, .5)$, the step length $\omega_n \in [0,1]$ is determined according to the Goldstein scheme (Dun80):

$$\omega_n \in \begin{cases} \{0\} & \text{if } w'(\eta^t)(\eta^t - \hat{\eta}^t) = 0 \\ \{1\} & \text{if } w'(\eta^t)(\eta^t - \hat{\eta}^t) > 0, h(\eta^t, \hat{\eta}^t, 1) \geq \delta \\ \Omega_\delta(\eta^t, \hat{\eta}^t) & \text{if } w'(\eta^t)(\eta^t - \hat{\eta}^t) > 0, h(\eta^t, \hat{\eta}^t, 1) < \delta, \end{cases}$$

where $h(\eta, \hat{\eta}, \omega) = \frac{w(\eta) - w(\eta + \omega(\hat{\eta} - \eta))}{\omega w'(\eta)(\eta - \hat{\eta})}$ and $\Omega_\delta(\eta, \hat{\eta}) = \{\omega \in [0,1], \delta \leq h(\eta, \hat{\eta}, \omega) \leq 1 - \delta\}$. Notice that for η^t close enough to the optimum, $\omega_n = 1$, which corresponds to the regular Newton regime. Here, $w'(\eta)$ and $w''(\eta)$ are given by the following formulae (see Appendix to Chapter 1 for the derivation):

$$\begin{aligned} w'(\eta) &:= 1 - \tfrac{\lambda^2}{4}\boldsymbol{g}^T\boldsymbol{B}^{-2}(\eta)\,\boldsymbol{g}, \\ w''(\eta) &:= \tfrac{\lambda^2}{2}\boldsymbol{g}^T\boldsymbol{C}(\eta)\,\boldsymbol{g}, \end{aligned} \tag{2.24}$$

where $\boldsymbol{C}(\eta) := \boldsymbol{B}^{-3}(\eta)\,\boldsymbol{H}$. Convergence analysis for this line-search technique is available in (Dun80).

Remark II.2 (Numerical implementation)**.** *To avoid inverting large matrices, in our implementation we compute the update (2.23) as follows. First we compute the Cholesky factorization of the positive definite matrix $\boldsymbol{B}(\eta^{t-1})$, which we denote by $\boldsymbol{R}$. Then, we solve six triangular systems of equations of the type $\boldsymbol{Q}\boldsymbol{z}_k = \boldsymbol{z}_{k-1}$, where $\boldsymbol{Q}$ is $\boldsymbol{R}$ if k is odd and $\boldsymbol{R}^T$ if k is even, and $z_0 = \boldsymbol{g}$. Finally we compute (2.23) as:*

$$\eta^t \leftarrow \left(\eta^{t-1} - \frac{1 - \frac{\lambda^2}{4}\boldsymbol{z}_1^T\boldsymbol{z}_1}{\frac{\lambda^2}{2}\boldsymbol{z}_6\boldsymbol{H}\boldsymbol{g}}\right)_+ \tag{2.25}$$

2.3 Applications

Here we illustrate the MSTO by considering a few applications in statistical signal processing.

2.3.1 Linear regression with ℓ_2 norm penalty

Given a vector of n observations $\boldsymbol{y}$ and an $n \times p$ design matrix $\boldsymbol{X}$, we consider the following class of problems:

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^r + \lambda \|\boldsymbol{\theta}\|_2^q. \tag{2.26}$$

Depending on r and q, this problem specializes to ridge regression ($r = 2$, $q = 2$), robust least-squares ($r = 1$, $q = 1$) [Theorem 3.2, (EGL97)] or ℓ_2-penalized least squares ($r = 2$, $q = 1$). The following corollary of Theorem II.1 characterizes the solution of the latter.

Corollary II.3. *The solution to the ℓ_2-penalized least squares*

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2, \tag{2.27}$$

is:

$$\hat{\boldsymbol{\theta}} = \begin{cases} \left(\boldsymbol{X}^T\boldsymbol{X} + \epsilon \boldsymbol{I}\right)^{-1} \boldsymbol{X}^T\boldsymbol{y} & \text{if } \|\boldsymbol{X}^T\boldsymbol{y}\|_2 > \frac{\lambda}{2} \\ \boldsymbol{0} & \text{otherwise,} \end{cases} \tag{2.28}$$

where the shrinkage parameter $\epsilon = \frac{\lambda^2}{4\eta^*}$ is such that $\eta^* > 0$ solves:

$$\min_{\eta > 0} \left(1 - \boldsymbol{y}^T\boldsymbol{X}\left(\eta \boldsymbol{X}^T\boldsymbol{X} + \frac{\lambda^2}{4}\right)^{-1} \boldsymbol{X}^T\boldsymbol{y}\right)\eta. \tag{2.29}$$

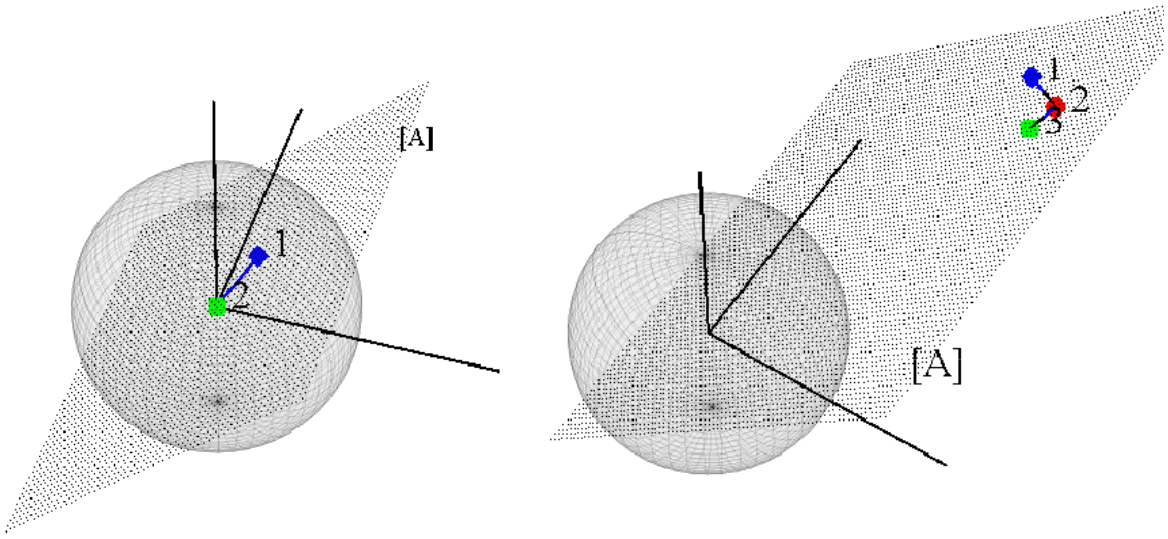
In the special case where $\boldsymbol{X}$ is orthogonal ($2\boldsymbol{X}^T\boldsymbol{X} = k\boldsymbol{I}$) then (2.26) has the closed form solution (2.28) with $\epsilon = \frac{\lambda k}{2(k\|\boldsymbol{y}\|_2 - \lambda)}$.

The proof of this Corollary follows immediately from Theorem II.1 by observing that $\hat{\boldsymbol{\theta}} = \mathcal{T}_{\lambda, 2\boldsymbol{X}^T\boldsymbol{X}}\left(-2\boldsymbol{X}^T\boldsymbol{y}\right)$.

Figure 2.1 depicts the geometrical interpretation of the result in Corollary II.3 in a 3-dimensional space.

2.3.2 Group Regularized Linear Discriminant Analysis

We consider here the problem of estimating the mean of a Gaussian distribution under the assumption of block-diagonal covariance and a block-sparse structure. Since block-sparsity constraints are non-convex, a usual approach is to relax the constraint to a convex Group LASSO (YL06a) penalty added to the maximum likelihood objec-



(a) $\mathcal{T}_{\lambda,2\boldsymbol{A}^T\boldsymbol{A}}\left(-2\boldsymbol{A}^T\boldsymbol{y}\right)$, $\|2\boldsymbol{A}^T\boldsymbol{y}\|_2 \leq \lambda$ (b) $\mathcal{T}_{\lambda,2\boldsymbol{A}^T\boldsymbol{A}}\left(-2\boldsymbol{A}^T\boldsymbol{y}\right)$, $\|2\boldsymbol{A}^T\boldsymbol{y}\|_2 > \lambda$

Figure 2.1: Three-dimensional example of the result of applying the MSTO to a vector $\boldsymbol{y}$ (denoted by (1) in the figure). The sphere (of radius λ) represents the boundary of the region in which $-2\boldsymbol{A}^T\boldsymbol{y}$ gets thresholded to 0, the plane represents the subspace $[\boldsymbol{A}]$. Point (2) on the right plot is the projection of $\boldsymbol{y}$ onto $[\boldsymbol{A}]$ and point (3) is the projected point after the shrinkage. Notice that as predicted by Theorem II.1, the amount of shrinkage is small compared to the norm of $\mathcal{T}_{\lambda,2\boldsymbol{A}^T\boldsymbol{A}}\left(-2\boldsymbol{A}^T\boldsymbol{y}\right)$, since the point $-2\boldsymbol{A}^T\boldsymbol{y}$ is far from the threshold boundary λ.

tive. Thus, given n independent realizations and m non-overlapping sets of indices G_i, we seek to solve:

$$\min_{\boldsymbol{\mu}\in R^N} \sum_{i=1}^{m} \frac{1}{2}\boldsymbol{\mu}_{G_i}^T \boldsymbol{\Sigma}_{G_i,G_i}^{-1}\boldsymbol{\mu}_{G_i} - \bar{\boldsymbol{x}}_{G_i}^T \boldsymbol{\Sigma}_{G_i,G_i}^{-1}\boldsymbol{\mu}_{G_i} + \lambda\|\boldsymbol{W}_i\boldsymbol{\mu}_{G_i}\|_2,$$

where $\boldsymbol{\Sigma}_{G_i,G_i}$ is the covariance of the elements in G_i and $\bar{\boldsymbol{x}}$ is the empirical mean of the observations. Here, the $\boldsymbol{W}_i$ are weighting matrices of the adequate size that allow us to determine what features of the data are to be more heavily penalized, as we will exemplify in Section 2.4.2. Applying (2.2), we can give the solution to this problem in terms of the MSTO applied to each block of variables:

$$\hat{\boldsymbol{\mu}}_{G_i} = \mathcal{T}_{\lambda,\boldsymbol{\Sigma}_{G_i,G_i}^{-1}}\left(\boldsymbol{\Sigma}_{G_i,G_i}^{-1}\bar{\boldsymbol{x}}_{G_i}\right), \quad i = 1,\cdots,m. \tag{2.30}$$

This regularized mean estimation problem arises in the context of Regularized Linear Discriminant Analysis (HTF05), (TP07), where we seek to build the linear discriminant function[1]:

$$\delta_{k,l}\left(\tilde{\boldsymbol{x}}\right) = \sum_{i=1}^{g} \tilde{\boldsymbol{x}}_{G_i}^T \boldsymbol{\Sigma}_{G_i,G_i}^{-1} \left(\boldsymbol{\mu}_{G_i}^k - \boldsymbol{\mu}_{G_i}^l\right)$$

which predicts whether $\tilde{\boldsymbol{x}}$ belongs to class k ($\delta_{k,l}\left(\tilde{\boldsymbol{x}}\right) > 0$) or class l ($\delta_{k,l}\left(\tilde{\boldsymbol{x}}\right) < 0$). Here $\boldsymbol{\mu}^k$ is an estimate of the mean for the k-th class, obtained through equation (2.30), and $\boldsymbol{\Sigma}_{G_i,G_i}$ is an estimate of the covariance of the elements in G_i, which is assumed the same across different classes. In (TP07), the authors set $\boldsymbol{W} = \boldsymbol{I}$ and propose to approximate (2.30) by:

$$\hat{\boldsymbol{\mu}}_{G_i} = \begin{cases} \left(1 - \frac{\lambda}{\left|\left|\boldsymbol{\Sigma}_{G_i,G_i}^{-1}\bar{\boldsymbol{x}}_{G_i}\right|\right|_2}\right)\bar{\boldsymbol{x}}_{G_i} & \text{if } \left|\left|\boldsymbol{\Sigma}_{G_i,G_i}^{-1}\bar{\boldsymbol{x}}_{G_i}\right|\right|_2 > \lambda \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

This approximation only coincides with the optimal solution of (2.30) when $\boldsymbol{x}_{G_i}$ is an eigenvector of $\boldsymbol{\Sigma}_{G_i,G_i}^{-1}$, a situation which is not likely to occur in practice.

2.3.3 Block-wise optimization for Group LASSO Linear Regression

In this section we consider the problem of solving the Group LASSO penalized Linear Regression problem. Given a vector of n observations $\boldsymbol{y}$ and an $n \times p$ design matrix $\boldsymbol{X}$ and q *disjoint* groups of indices $G_i \subseteq \{1, ..., N\}$ satisfying $\cup_{i=1}^{q} G_i = \{1, ..., N\}$, the Group LASSO linear regression problem (YL06a) is defined as:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in R^N} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \sum_{i=1}^{q} \lambda_i \|\boldsymbol{\theta}_{G_i}\|_2, \tag{2.31}$$

where λ_i are fixed penalty parameters which we assume known. For an arbitrary design matrix $\boldsymbol{X}$, problem (2.31) can be solved using a Block Coordinate Descent (BCD) algorithm. The main idea of the BCD method is to iteratively solve (2.31) for each block G_i, letting the parameters corresponding to the other blocks remain fixed. Defining $\boldsymbol{H} = 2\boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{g} = -2\boldsymbol{X}^T\boldsymbol{y}$ and using the MSTO (2.2) we can obtain the following update rule for each block G_i at iteration t:

$$\boldsymbol{\theta}_{G_i}^t \leftarrow \mathcal{T}_{\lambda_i, \boldsymbol{H}_{G_i,G_i}} \left(\boldsymbol{\theta}_{\bar{G}_i}^{t-1} \boldsymbol{H}_{\bar{G}_i,G_i} + \boldsymbol{g}_{G_i}\right), \tag{2.32}$$

[1]Here we assume that the prior probabilities for each class are equal.

where $\bar{G}_i$ is the complementary set of indices with respect to G_i. Convergence of this algorithm is guaranteed for this cost function (Tse01).

2.3.4 MSTO in proximity operators

The proximity operator of a (possibly non-differentiable) convex function $\Omega(\boldsymbol{x})$ is defined as (Mor65), (CW06):

$$\mathcal{P}_{\tau,\Omega}(\boldsymbol{g}) := \arg\min_{\boldsymbol{x}} \frac{1}{2\tau} ||\boldsymbol{x}-\boldsymbol{g}||_2^2 + \Omega(\boldsymbol{x}).$$

Proximity operators are the main ingredient of proximal algorithms (CW06), (BT09), which arise in LASSO and Group LASSO penalized linear regression (DDDM04a), (BT09), (ZRY09), collaborative sparse modeling (SRSE10b) and hierarchical dictionary learning (JMOB10). In these applications, proximal algorithms can be understood as a generalization of quasi-Newton methods to non-differentiable convex problems. An important example is the Iterative Thresholding procedure (DDDM04a), (BT09) which solves problems of the form:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) + \Omega(\boldsymbol{x}),$$

with $f(\boldsymbol{x})$ differentiable and with Lipschitz gradient, by generating the sequence

$$\boldsymbol{x}^{t+1} \leftarrow \mathcal{P}_{k,\Omega}\left(\boldsymbol{x}^t - k\nabla f\left(\boldsymbol{y}^t\right)\right),$$

for an appropriate $k > 0$ and a carefully chosen $\boldsymbol{y}^t$. These algorithms are suitable for applications where the evaluation of $\mathcal{P}_{k,\Omega}$ can be done at low cost.

In some cases, the proximity operator $\mathcal{P}_{k,\Omega}$ can be evaluated in closed form. This is the case for instance when $\Omega(\boldsymbol{x}) = ||\boldsymbol{x}||_2$, where it is given by the orthogonal MSTO (2.22), or in general when $\Omega(\boldsymbol{x}) = \sum_i ||\boldsymbol{A}_i\boldsymbol{x}||_2$ and the supports of $\boldsymbol{A}_i$ are disjoint. Another interesting example is the case of Group LASSO penalties with overlapping hierarchical groups. Given $\lambda > 0$, q groups of indices $G_i \subseteq \{1, ..., N\}$ and a partial order $\mathcal{O} = (o_1, ..., o_q)$ of the groups such that $G_{o_{i+1}} \cap G_{o_i} \neq \emptyset$ only if $G_{o_i} \subseteq G_{o_{i+1}}$ we consider the following function:

$$\Gamma(\boldsymbol{x}) = \lambda \sum_{i=1}^{q} \left|\left|\boldsymbol{x}_{G_{o_i}}\right|\right|_2. \tag{2.33}$$

It can be shown (JMOB10) that:

$$\mathcal{P}_{\tau,\Gamma}\left(\boldsymbol{g}\right) = \bigcirc_{i=1}^{q}\left(\mathcal{T}_{G_{o_i},\tau\lambda,\boldsymbol{I}}\right)\left(\boldsymbol{g}\right), \tag{2.34}$$

where $\bigcirc$ is the composition operator and $\mathcal{T}_{s,\lambda,\boldsymbol{I}}\left(\boldsymbol{g}\right)$ is the MSTO defined on a subset s,

$$\mathcal{T}_{s,\lambda,\boldsymbol{I}}\left(\boldsymbol{g}\right) \;\; := \arg\min_{\boldsymbol{x}} \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} + \lambda\|\boldsymbol{x}_s\|_2, \tag{2.35}$$

where $s \subseteq \{1, \cdots, N\}$. It is clear that $\left[\mathcal{T}_{s,\lambda,\boldsymbol{I}}\left(\boldsymbol{g}\right)\right]_s = \mathcal{T}_{\lambda,\boldsymbol{I}}\left(\boldsymbol{g}_s\right)$ and $\left[\mathcal{T}_{s,\lambda,\boldsymbol{I}}\left(\boldsymbol{g}\right)\right]_i = \boldsymbol{g}_i$ for $i \notin s$.

2.4 Numerical Results

In this section we first illustrate the advantage of evaluating the MSTO using our theoretical results. Second, we will apply the MSTO to find predictive genes that separate two populations in a real gene expression time course study.

2.4.1 Evaluation of the MSTO

In this section we illustrate the advantage of evaluating the MSTO using our theoretical results. To this end, we compare the elapsed times to evaluate equation (2.2) using three different optimization methods. The first one, which we denote by MSTO in the figures, solves the dual problem in (2.5) using the projected Newton approach described in Sec. 2.2.1. The second method uses an accelerated first order method named FISTA[2] (BT09) and the third method uses the commercial state-of-the-art SOCP solver Mosek®. Our experiment consists of solving problem (2.27) for randomly generated $\boldsymbol{X}$ and $\boldsymbol{y}$ where we control the conditioning of the matrix $\boldsymbol{H} = 2\boldsymbol{X}^T\boldsymbol{X}$ through the ratio p/n (where p is the number of columns and n is the number of rows of $\boldsymbol{X}$).

We show in Figure 2.2 the average elapsed times to achieve the same value of the objective function, as a function of the number of variables and the ratio p/n. Our algorithm outperforms the other two over a large range of values of p when p/n is close to one, and offers comparable performances for larger values of p/n.

[2]FISTA is implemented using backtracking and (2.22) to compute its corresponding shrinkage/thresholding update.

(a) $p/n = 1$ fixed, varying p (b) $n = 100$ fixed, varying p/n

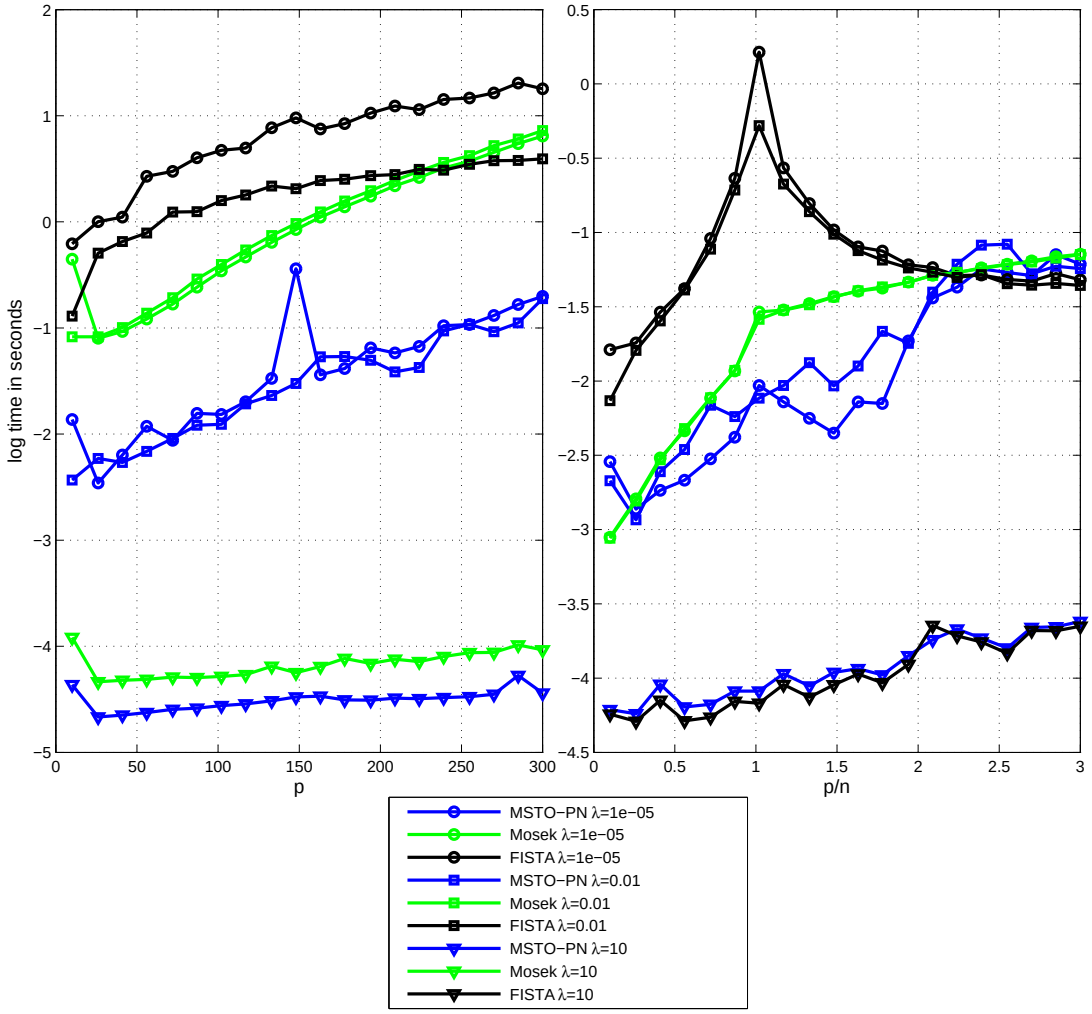


Figure 2.2: Comparison of Mosek®, MSTO and FISTA elapsed times for solving (2.27) while varying p (with $p/n = 1$ fixed, plot (a)) and varying p/n (with $n = 100$ fixed, plot (b)). For each algorithm, we compute the MSTO solution for three different values of the penalty parameter λ. MSTO is significantly faster than the other two when the conditioning of the problem is not too poor and offers comparable performance in the other regimes.

2.4.2 Finding discriminative genes in a time course gene expression study

In this section we will apply the MSTO to the discovery of discriminative genes in a time-course gene expression study. Our goal is to find a subset of genes that discriminate two populations consistently across different time points. Our work can be seen as an extension to time-course data of the popular "nearest shrunken centroid"

method (THNC03) for discovering genes whose expression level discriminates different biological conditions. We will first briefly describe the nearest shrunken centroid method before turning to the description of our extension and the application to the PHD dataset.

2.4.2.1 Class prediction via nearest shrunken centroid

The nearest shrunken centroid method can be understood as an ℓ_1 penalized estimation of a Gaussian mean under known, diagonal covariance. The data model is akin to the following. Consider a situation where we have p gene expression values from samples obtained from K different conditions. The gene expression level $x_i^k \in \mathbb{R}^p$ for the i-th gene for a sample from the k-th class is assumed to be a Gaussian random variable with mean $\mu_i^k + \bar{\mu}_i$ and variance σ_i^2:

$$x_i^k \sim \mathcal{N}\left(\mu_i^k + \bar{\mu}_i, \sigma_i^2\right). \tag{2.36}$$

Here $\bar{\mu}_i$ corresponds to the interclass mean, σ_i^2 is assumed to be the same for each class and the genes are assumed to be uncorrelated to each other [3]. Under this assumptions an for given $\bar{\mu}_i$ and σ_i^2, the maximum likelihood estimate of $\boldsymbol{\mu}_k$ is given by the intra-class average:

$$\hat{\mu_i}^k = \bar{x_i}^k - \bar{\mu}_i,$$

where $\bar{x_i}^k$ denotes the empirical mean of the training samples from class k. The estimates $\bar{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}}^k$ constitute the centroids. A new (test) sample is classified by assigning it to the class corresponding to the nearest centroid, as explained in (THNC03).

A negative characteristic of the estimate $\hat{\boldsymbol{\mu}}^k$ is that is very sensitive to noise, specially when the number of samples is small compared to p, the number of genes. To alleviate this problem, (THNC03) propose to estimate instead "shrunken centroids", which relate to the ℓ_1 penalized MLE estimate of $\boldsymbol{\mu}_k$ under the Gaussian model (2.36). Using equation (2.1), the i-th coordinate of this estimate is given by an application of the Scalar Shrinkage Thresholding Operator (2.1):

$$\tilde{\mu}_i^k = \mathcal{T}_{\lambda,\sigma_i}\left(\bar{x}_i^k - \bar{\mu}_i\right). \tag{2.37}$$

[3] These assumptions are far from being true for real gene expression data. However, estimation of the covariance between genes is impossible in practical scenarios, where there are tens of thousands of genes and only tens of samples. The approach taken in (THNC03) is to impose a very-low dimensional model on the covariance: a diagonal one. This obviously introduces bias, but mitigates the variance and yields good results in practice.

The shrunken centroids are then constructed as $\bar{\boldsymbol{\mu}} + \tilde{\boldsymbol{\mu}}^k$. The soft-thresholding operator has the effect of yielding sparse $\tilde{\boldsymbol{\mu}}^k$ that are more robust to noise. The genes corresponding to non-null coordinates of $\tilde{\boldsymbol{\mu}}^k$ are those that are more strongly associated with each particular class, giving insight into the relationship between the selected genes and each phenotype.

2.4.2.2 Class prediction via nearest group-shrunken centroid

In gene expression time course studies, we are often interested in finding genes whose expression values discriminate between two or more classes over time. In this setting, genes of a given class can not be expected to have the same mean over different time points, since this would imply that their gene expression value is not time-dependent. We propose instead the following model. Denote by $\boldsymbol{x}_i^k \in \mathbb{R}^T$ the vector gene expression values for the i-th gene over T time points, for a sample from class k. We characterize each T-dimensional vector of gene expression levels as:

$$\boldsymbol{x}_i^k \sim \mathcal{N}\left(\boldsymbol{\mu}_i^k + \bar{\boldsymbol{\mu}}_i, \boldsymbol{\Sigma}_i^2\right), \quad i = 1, \cdots, p, \ , k = 1, \cdots, K, \tag{2.38}$$

where $\boldsymbol{\Sigma}_i$ is the temporal covariance matrix for the i-th gene and we assume, as in (THNC03), that each gene is uncorrelated to the others, i.e. $E\left(\boldsymbol{x}_i^k \boldsymbol{x}_j^{k^T}\right) = \mathbf{0}$, $i \neq j$.

Similarly to the previous section, we now need to estimate $\boldsymbol{\mu}_i^k$ to construct the centroids that will allows us to discriminate samples from different classes. Since each gene's component is now multi-dimensional, we propose to replace the shrinkage thresholding operator in (2.37) by its multi-dimensional extension, the MSTO:

$$\tilde{\boldsymbol{\mu}}_i^k = \mathcal{T}_{\lambda, \boldsymbol{\Sigma}_i}\left(\bar{\boldsymbol{x}}_i^k - \bar{\boldsymbol{\mu}}_i\right). \tag{2.39}$$

Note that from the results in Section 2.3.2, this is precisely the penalized maximum likelihood estimate of $\boldsymbol{\mu}_i^k$ under known $\bar{\boldsymbol{\mu}}_i$ and $\boldsymbol{\Sigma}_i$, only that now, instead of an ℓ_1 penalty, we use a group-ℓ_2 penalty that enforces each gene to have all or none of its components activated.

In the next section we apply this methodology to the 2-class problem, and show the advantage of the MSTO-based approach over the original nearest shrunken centroid method of (THNC03).

2.4.2.3 Specialization to the 2-class prediction problem

When the number of classes is equal to two, we will consider the following variation. Given a test sample $\tilde{\boldsymbol{x}}$, the linear discriminant function of Section 2.3.2 specializes to the following,

$$\delta_{1,2}\left(\tilde{\boldsymbol{x}}\right) = \sum_{i=1}^{p} \tilde{\boldsymbol{x}}_i^T \boldsymbol{\Sigma}_i^{-1} \left(\boldsymbol{\mu}_i^1 - \boldsymbol{\mu}_i^2\right),$$

and we assign the label '1' if $\delta_{1,2}\left(\tilde{\boldsymbol{x}}\right) > 0$ and '2' otherwise. A common approach for the two class problem is to estimate $\Delta_i^{1-2} := \boldsymbol{\mu}_i^1 - \boldsymbol{\mu}_i^2$ instead of estimating $\boldsymbol{\mu}_i^1$ and $\boldsymbol{\mu}_i^2$ separately. Under the Gaussian model and with the assumption that we have the same number of training samples for each class, the ℓ_2-penalized MLE estimate of this quantity is given by:

$$\tilde{\Delta}_i^{1-2} = \mathcal{T}_{\lambda,\boldsymbol{\Sigma}_i}\left(\bar{\boldsymbol{x}}_i^1 - \bar{\boldsymbol{x}}_i^2\right). \tag{2.40}$$

Here $\bar{\boldsymbol{x}}_i^1$ and $\bar{\boldsymbol{x}}_i^2$ denote the average values for the '1' and for the '2' class, and each variable's covariance, $\boldsymbol{\Sigma}_i$, is estimated as follows:

$$\boldsymbol{\Sigma}_i = \hat{\boldsymbol{S}}_i + \delta \boldsymbol{I}$$

where $\hat{\boldsymbol{S}}_i$ is the pooled empirical covariance,

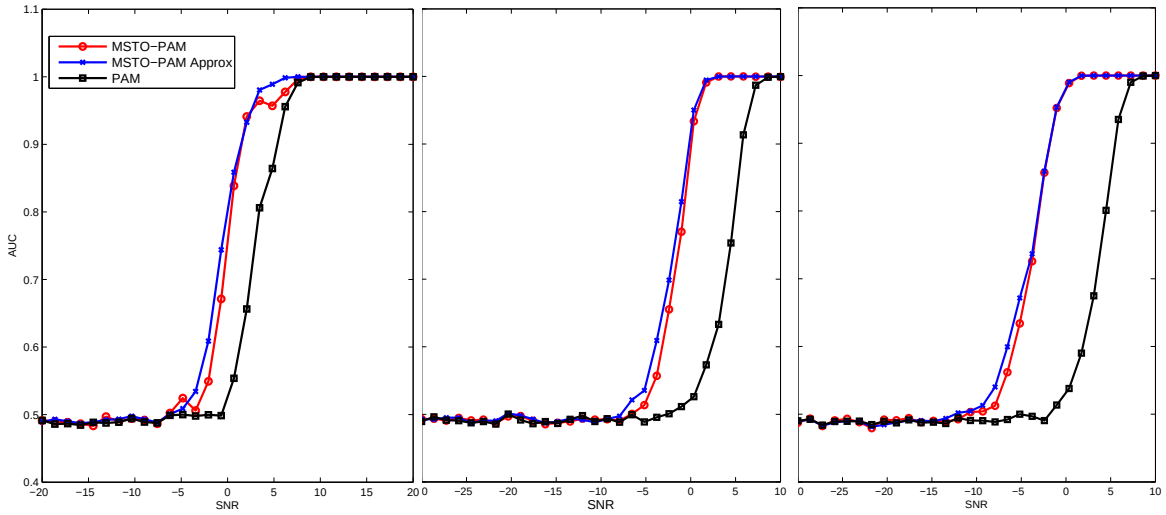
$$\hat{\boldsymbol{S}}_i = \frac{1}{n-1} \sum_{i=1}^{n} \boldsymbol{x}_i$$

and δ is a shrinkage parameter that guarantees $\boldsymbol{\Sigma}_i \succ 0$. The choice of a pooled covariance estimator is usually motivated by the scarcity of available samples, which hinders the estimation of a covariance matrix for each class.

We will now assess the performance of the Group-Shrunken Centroid approach in the 2-class prediction problem. For this purpose, we will generate data according to model (2.38) with $K = 2$ and $T = 6$, and we let:

$$\boldsymbol{\mu}_i^1 - \boldsymbol{\mu}_i^2 = \Delta \mathbf{1}, \quad \boldsymbol{\Sigma}_i = \left(\tfrac{1}{T} - 1\right) \mathbf{1}\mathbf{1}^T + \boldsymbol{I}.$$

with varying Δ, a parameter which controls the separability between the two means and is related to the Signal-to-Noise Ratio (SNR). As a measure of performance, we will consider the Area Under the Curve (AUC). We compare the Group-Shrunken



$p = 10^4$, $n = 10$ $p = 10^4$, $n = 10^2$ $p = 10^4$, $n = 10^3$

Figure 2.3: Comparison of prediction performances for the 2-class problem with p variables and n samples, for each of the methods discussed in Section 2.4.2: nearest shrunken centroid (PAM), nearest group-shrunken centroid (MSTO-PAM) and nearest group-shrunken centroid using the approximation from (TP07) (MSTO-PAM Approx). The measure of performance is the estimated Area Under the Curve (AUC). As the number of samples increases, the advantage of $MSTO - PAM$ and its approximation over PAM increases, possibly due to the incorporation of the covariance within groups of variables in the predictor estimator.

Centroids approach given by (2.40) and denoted by MSTO-PAM, to the MSTO-based approximation given in (2.31) from (TP07) and the classical Nearest Shrunken centroids described in Section 2.4.2.1 (also known as PAM, and thus labeled in our figure). The results, shown in Figure 2.3, reflect the advantage of taking into account the group structure of the generative model, as shown by the increased robustness with respect to noise. On the other hand, the approximation of (2.40) given by (2.31) performs remarkably well, and better than the MSTO-PAM when n is very small ($n = 10$). This can be explained by the fact that the performance of the covariance estimate given in (2.41) with such small sample size is very poor, thus degrading the predictive performance of the MSTO-PAM compared to the approximation in (2.31), which does not use a covariance estimate in order to construct the centroid.

2.4.2.4 Application to the PHD dataset

The PHD dataset for the H3N2 challenge study, described in Section 1.5, consists of $p = 11961$ gene expression levels of 17 different individuals at 16 different time points. Physicians have classified each individual into two classes, Symptomatic (Sx) and Asymptomatic (Asx), depending on the strength of the physical symptoms they show, with $n_1 = 9$ individuals in class Sx and $n_2 = 8$ individuals in class Asx. Our goal is to find a small subset of genes that consistently discriminate the two classes across samples from 6 different time points starting from inoculation time.

Using the results of the previous section for the 2-class problems, we denote or discriminant function by:

$$\delta_{\text{Sx,Asx}}\left(\tilde{\boldsymbol{x}}\right) = \sum_{i=1}^{p} \tilde{\boldsymbol{x}}_i^T \boldsymbol{\Sigma}_i^{-1} \tilde{\Delta}_i^{\text{Sx/Asx}},$$

where the shrunken centroid $\tilde{\Delta}_i^{\text{Sx/Asx}}$ for each gene is estimated as:

$$\tilde{\Delta}_i^{\text{Sx/Asx}} = \mathcal{T}_{\lambda,\boldsymbol{\Sigma}_i}\left(\bar{\boldsymbol{x}}_i^{\text{Sx}} - \bar{\boldsymbol{x}}_i^{\text{Asx}}\right),$$

and each gene's covariance is estimated as in (2.41).

In our study we consider three possibilities for the weight matrix $\boldsymbol{W}$: An identity matrix, a diagonal matrix with decreasing exponential weights (diagonal $\boldsymbol{W}_1$, $[\boldsymbol{W}_1]_{k,k} = e^{-(k-1)}$) and one with increasing exponential weights (diagonal $\boldsymbol{W}_2$, $[\boldsymbol{W}_2]_{k,k} = \frac{e^{(k-1)}}{e^{p-1}}$). The weight matrix $\boldsymbol{W}_2$ penalizes late time points more strongly, whereas $\boldsymbol{W}_1$ heavily penalizes early time points, allowing us to additionally constrain the candidate gene trajectories depending on the temporal features we believe are more discriminatory.

As it is common practice, we choose the regularization parameter λ by leave-one-out cross-validation (HTF05), fixing the false alarm rate and estimating the power for each value of λ over a 30-point grid.

Figure 2.4 plots the cross-validation results for the three choices of $\boldsymbol{W}$. The curves show the estimated predictive power at a false alarm rate of .05 and the average number of genes used to construct the classifier for each level of λ. Each choice of $\boldsymbol{W}$ correspond to the selection of different genes, which in turn have different prediction capabilities, as reflected by the different power maximums in the upper panel. The lower panel shows that the number of genes selected at the optimum power level is also dependent on our choice of $\boldsymbol{W}$. It is clear that the choice of decreasing exponential

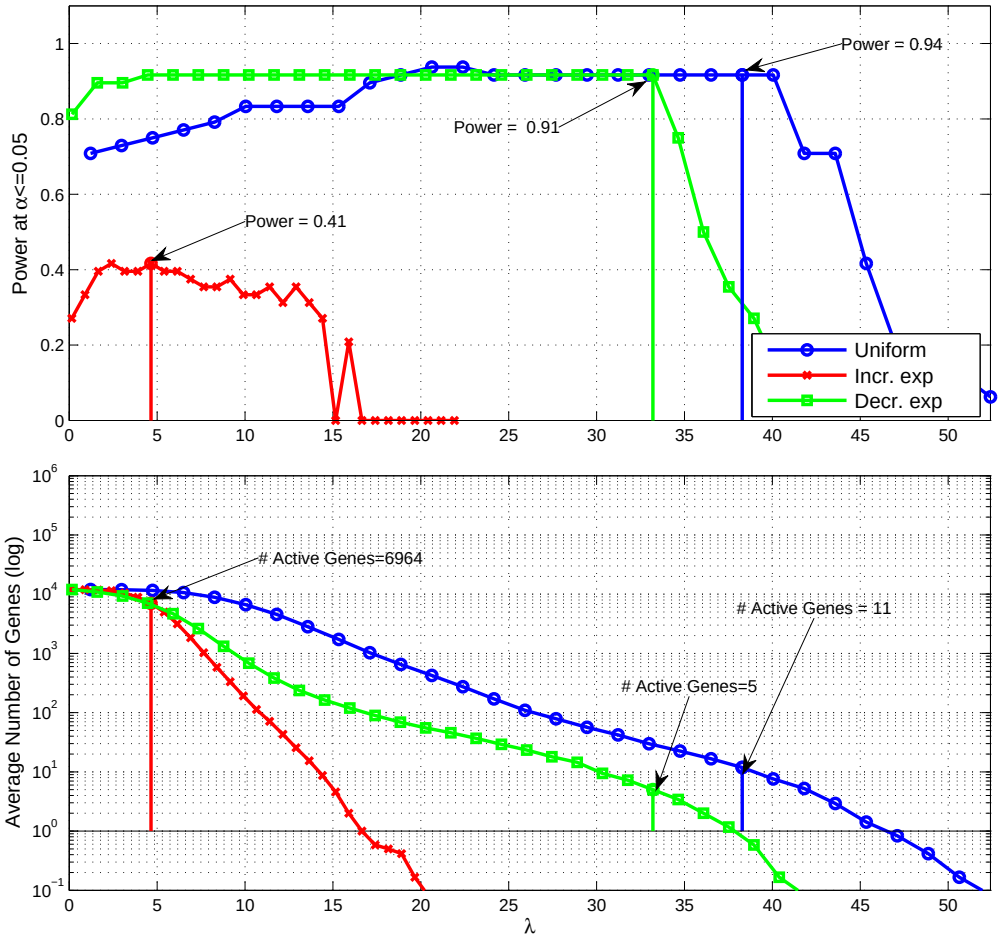


Figure 2.4: Cross-validation results for the three choices of $\boldsymbol{W}$. The top plot shows the estimated power of our classifier versus λ. The bottom plot shows the average number of genes used in the classifier versus λ. As λ increases, the penalty is more stringent and the number of genes included in the model decreases.

weights offer the best trade-off between number of genes selected and power of the classifier, achieving an estimated power of .91 with an average of only 5 genes.

In order to gain more insight on how the weight choice affects the gene selection, we plot the most significant genes for each classifier in Figure 2.5. The significance here is assessed through the number of times that each gene appears in the classifier, compared to the total number of Cross-Validation runs. A gene whose coordinate $\tilde{\Delta}_i^{\text{Sx/Asx}}$ is non-null in a large number of cross-validation runs is likely to be important in

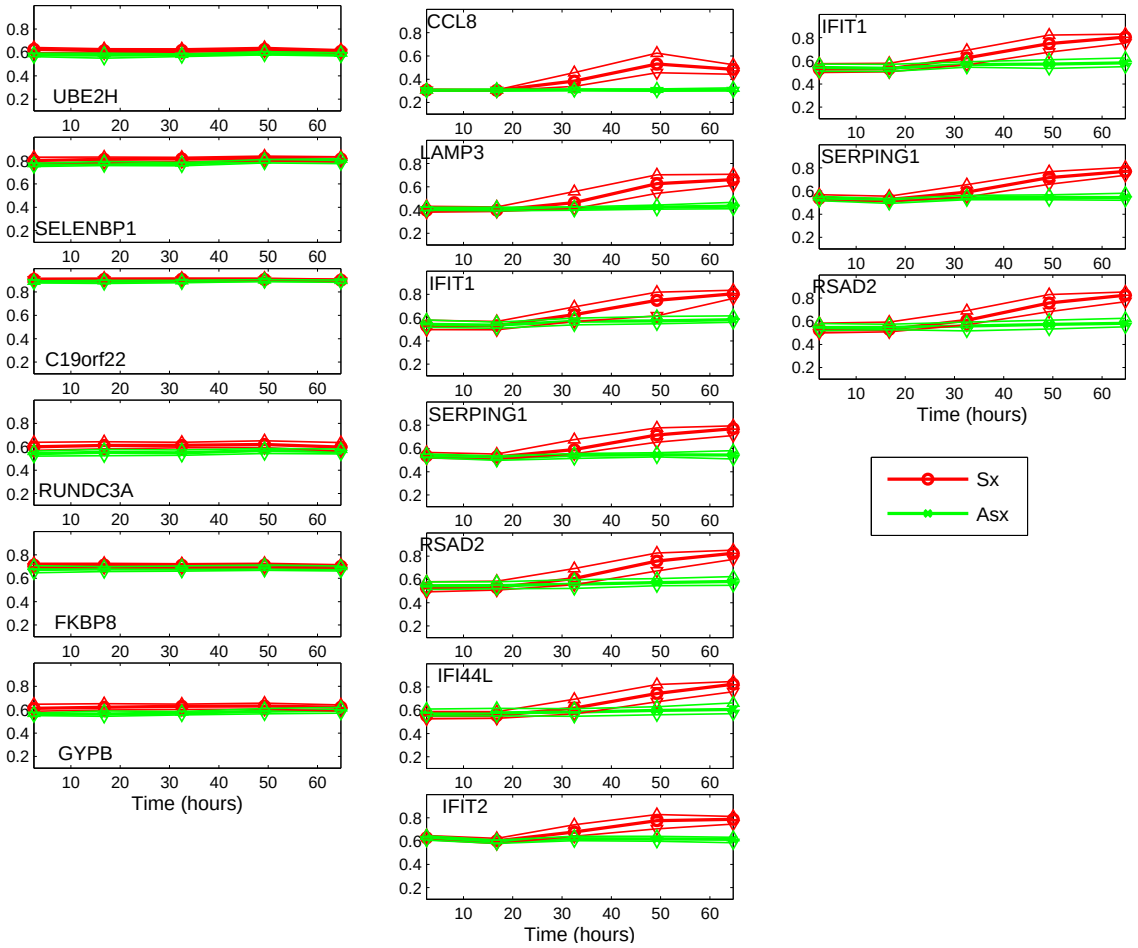
the discrimination task (Bac08). Here we select only genes that appear at least at 70% of the Cross-Validation runs. Figure 2.5 shows the average within-class expression response and the 95% confidence intervals for the significant genes obtained in each case. Since $\boldsymbol{W}_1$ favors genes that are highly discriminative in the early time points, this classifier selects genes whose trajectories remain separated over the 5 different time points, at the price of requiring a higher number of genes to perform at the same power level as the other two cases. On the other hand, $\boldsymbol{W}_2$ encourages a classifier that is highly discriminative at the late time points, as reflected by the average trajectory of gene LOC26010 (right panel).

Finally, we validate our results by constructing a classifier for the H1N1 challenge study. This study consists of samples from 21 individuals, divided into 10 Symptomatic and 14 Asymptomatic subjects. We train a simple LDA classifier with each group of genes that were declared highly discriminatory for H3N2, whose trajectories are shown in Figure 2.5. We estimate the resulting ROC curves by leave-one-out Cross Validation. The results, shown in 2.6, reflect the benefit of having a very sparse discriminator consisting of only one gene. In addition, the high-level of prediction accuracy suggests that genes discriminating sick individuals for the H3N2 virus are also good discriminators for the H1N1 virus, despite the biological differences between the two pathogens.

2.5 Conclusions

We have introduced the MSTO, which is a generalization of the Scalar Shrinkage Thresholding Operator. Our main theoretical result shows that the MSTO can be evaluated by solving a smooth low-dimensional problem and that they can be interpreted as an Shrinkage/Thresholding operation on the input vector.

The MSTO appears naturally in several l_2 penalized estimation problems. We have demonstrated the efficiency of the Projected Newton algorithm in evaluating the MSTO through its smooth reformulation, comparing it to other state of the art optimization methods. We have finally shown an example of its application for the discovery of predictive genes in a real time course gene expression study. Our methodology is capable of rapidly selecting genes that have good prediction power while allowing us to incorporate prior information on the type of time trajectories of interest.



Increasing exponential ($\boldsymbol{W}_1$) Uniform weights Decreasing exponential ($\boldsymbol{W}_2$)

Figure 2.5: Average within-class expression response and the bootstrapped 95% confidence intervals for the significant genes (appearing in more than 70% of the CV predictors) obtained for each choice of weight matrix $\boldsymbol{W}$. $\boldsymbol{W}_1$ favors genes that are discriminative in the early time points, which leads to poor prediction performances. On the contrary, $\boldsymbol{W}_2$ encourages a classifier that is highly discriminative at the late time points, which is where the difference between classes is stronger, leading to high prediction performance.

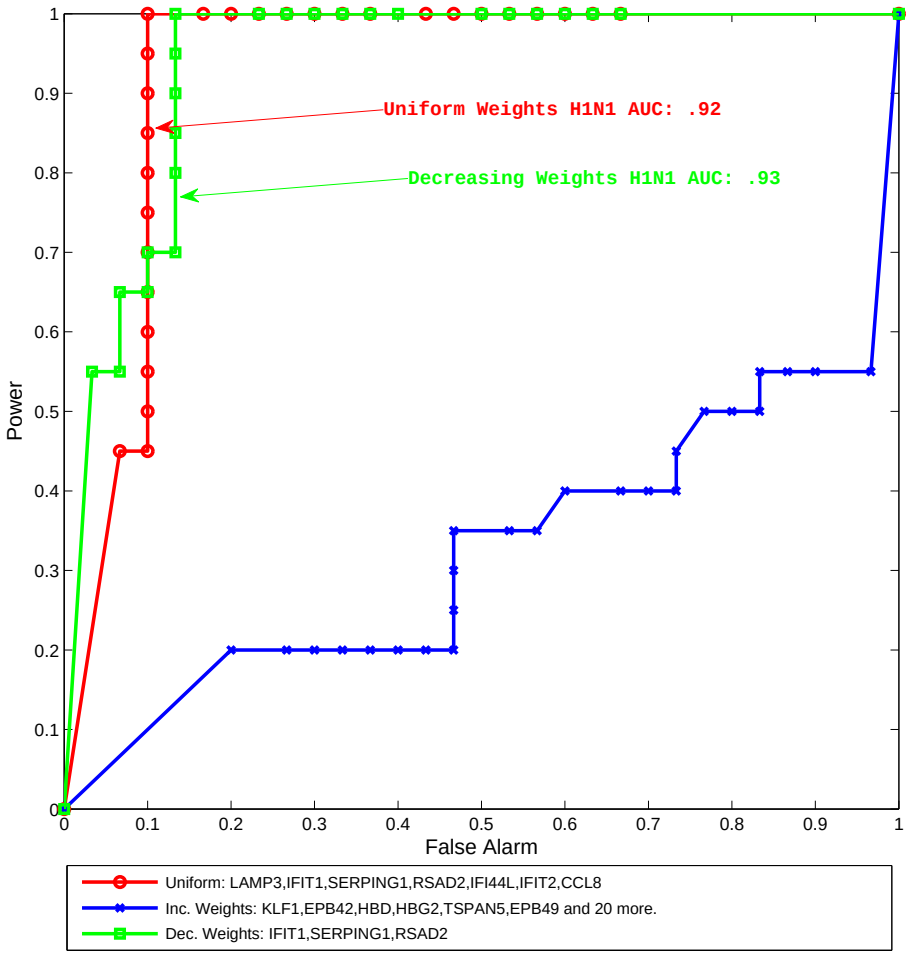


Figure 2.6: Estimated ROC for predictors of Symptomatic/Asymptomatic condition of H1N1-infected subjects, constructed from the sets of genes obtained in the H3N2 analyses.

CHAPTER III

A generalized shrinkage-thresholding operator

3.1 Introduction

In this chapter we extend the multidimensional shrinkage-thresholding operator (MSTO) of Chapter 2, defined as:

$$\mathcal{T}_{\lambda,\boldsymbol{H}}(\boldsymbol{g}) = \arg\min_{\boldsymbol{x}} \quad \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} + \lambda\,||\boldsymbol{x}||_2\,,$$

to penalties other than the ℓ_2 norm. In particular, we consider additive combinations of ℓ_2 norms applied to linear transformations of the optimization variables. Thus, similarly to the MSTO, we define the Generalized Shrinkage Thresholding Operator (GSTO) as the solution to the following convex optimization problem:

$$\mathcal{T}_{\lambda,\boldsymbol{H}}(\boldsymbol{g}) = \arg\min_{\boldsymbol{x}} \quad \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} + \lambda\sum_{i=1}^{m}\sqrt{c_i}\,||\boldsymbol{A}_{G_i,*}\boldsymbol{x}||_2\,,$$

where $\boldsymbol{H}$, $\boldsymbol{g}$, λ, $\boldsymbol{c}$, $\boldsymbol{A}$, and $\{G_i\}_{i=1}^m$ are the problem parameters, which we will define later. Since, the GSTO reduces to the MSTO when $\boldsymbol{A} = \boldsymbol{I}$ and $I_1 = \{1, \cdots, p\}$ we maintain the notation $\mathcal{T}_{\lambda,\boldsymbol{H}}(\boldsymbol{g})$ to refer to both. In our formulation, $\boldsymbol{A}$ is potentially a very large matrix and is not necessarily invertible, hence it will not always be possible or practical to evaluate $\mathcal{T}_{\lambda,\boldsymbol{H}}(\boldsymbol{g})$ by applying the change of variables $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ and solving with respect to $\boldsymbol{y}$.

In analogy to the MSTO, instances of the GSTO appear as a fundamental step in proximal methods for non-linear regression problems and allow us to explicitly characterize the solution of well-known penalized linear regression problems. Particularly, we will show that different choices of $\boldsymbol{A}$ and $\{G_i\}_{i=1}^m$ define a rich class of penalties, including some special cases that are popular in the machine learning and signal processing literature. For instance, choosing $\boldsymbol{A} = \boldsymbol{I}$ leads to the LASSO (Tib96)

or the Group LASSO (YL06b) penalties, depending on whether the disjoint subsets G_i are singletons or subsets of larger sizes. The Hierarchical (ZRY09; JMOB10) or Structured-Sparse Penalties (JAB09) can also be seen as special instances of the additive ℓ_2 penalty $\sum_{i=1}^{m} \sqrt{c_i} \left|\left|\boldsymbol{A}_{G_i,*}\boldsymbol{x}\right|\right|_2$ with $\boldsymbol{A}_{G_i,\cdot} = \boldsymbol{I}_{|G_i|}$ and overlapping subsets of variables G_i. This last type of structure is particularly important because it arises frequently in pathway-penalized gene expression regression problems such as the one described in Section 3.4. In this class of problems, one seeks to find a subset of pathways that are good linear predictors of the value of response variable. Pathways are groups of genes that have been experimentally verified to participate in certain biological processes, and are a priori known to be co-regulated. Unfortunately, the relationship between genes and pathways is not many-to-one, instead, as illustrated by the example in Figure 3.1, genes belong to several pathways leading to an overlapping group-sparse structure.

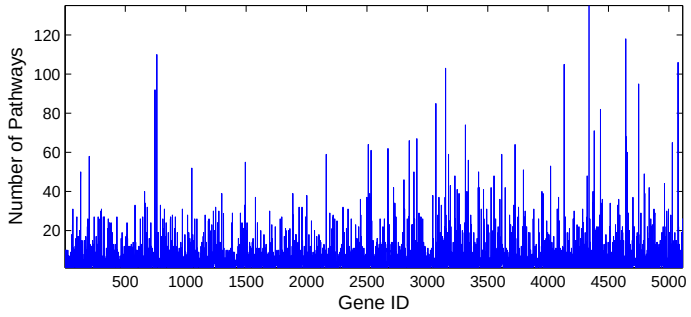


Figure 3.1: Number of pathways containing each gene, for the subset of 5125 genes and 826 pathways used in the analyses of Section 3.4. On average, each gene belongs to 6.2 pathways, and the number of pathways containing each gene ranges from 1 to 135.

All of the aforementioned penalties are important because they enforce a coordinate-wise sparse solution in penalized statistical learning problems. In the sequel, we will also show that the more general penalties we consider here extend this paradigm to the case where the sparsity is enforced in a space other than the ambient euclidean space. This kind of structure includes extensions of well known paradigms such as Total Variation penalized estimation, which enforce estimates with sparse discrete differences which do not necessarily have a sparse representation on the canonical coordinate system.

Interestingly, the analogy to the MSTO is not only related to the shrinkage-thresholding effect of the operator on the input vector $\boldsymbol{g}$. Using a transformation of the non-differentiable GSTO problem to a quadratic constraint problem, we show

that the GSTO can be evaluated by solving a problem of dimension equal to the number m of ℓ_2 norms appearing in the additive penalty. Thus, our reformulation is specially convenient for problems where the number of terms in the penalty, is much smaller than the dimension p of the ambient space. We will also show that the low-dimensional GSTO reformulation can be approximately solved with guaranteed accuracy using an efficient Projected Newton method that is specially convenient for m taking values in the hundreds. For the problem of computing the regularization path, that is, evaluating the GSTO over a grid of penalty parameters $\{\lambda_k\}_k$, we devise a path-following update that takes advantage of the smoothness of the GSTO reformulation over the active set of variables to reduce significantly the number of iterations necessary to evaluate the GSTO for each λ in the grid.

This chapter is organized as follows. In Section 3.2, we define the GSTO and introduce our first key theoretical result. Second, we apply this result to obtain a low-dimensional reformulation of the GSTO problem, and demonstrate that it indeed behaves as a shrinkage thresholding operator for the well known LASSO and Group LASSO cases. We give two algorithms to solve the low-dimensional reformulation in Section 3.3 and a path-following update that enhances the evaluation of the GSTO over a discrete grid of penalty parameters $\{\lambda_k\}_k$. In Section 3.4 we present numerical experiments and an application of the GSTO to the multi-task learning problem of finding symptom predictors from the gene expression data of a group of Symptomatic individuals.

3.2 The Generalized Shrinkage Thresholding Operator

Given $\boldsymbol{H} \in \mathbb{S}^p_+$, $\boldsymbol{g} \in \mathbb{R}^p$, $\lambda > 0$, $\boldsymbol{c} \in \mathbb{R}^m_{++}$, a collection of m subsets $G_i \subseteq \{1, \cdots, p\}$ of size n_i, and $\boldsymbol{A} \in \mathbb{R}^{\left(\sum_{i=1}^m n_i\right) \times p}$, we define the Generalized Shrinkage-Thresholding operator as the solution to the convex program:

$$\mathcal{T}_{\lambda,\boldsymbol{H}}(\boldsymbol{g}) = \arg\min_{\boldsymbol{x}} \quad \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} + \lambda \sum_{i=1}^m \sqrt{c_i}\, ||\boldsymbol{A}_{G_i,*}\boldsymbol{x}||_2 \,. \tag{3.1}$$

Notice that if $p = 1$, $m = 1$ and $\boldsymbol{A}_1 = 1$, then the GSTO reduces to the SSTO of Section 2.2. Also, if $p > 1$, $m = 1$ and $\boldsymbol{A}_1 = \boldsymbol{I}$, then the GSTO is equal to the MSTO (2.2).

Problem (3.1) can be cast as a Second Order Cone Program (SOCP), with m conic constraints. Efficient Interior Point solvers exist for this class of problems, however, they require storage of Hessian matrices of dimension $p \times p$ and computing solu-

tions of p-dimensional systems of equations. For large-scale problems, this approach necessitates exceedingly large memory requirements.

The following lemma shows that the solution of the convex p-dimensional, non-differentiable problem (3.1) is directly related to the solution of a smooth $(p + \sum_{i=1}^{m} n_i)$-dimensional problem, paving the way for an analog of MSTOs Theorem II.1 for the GSTO.

Lemma III.1. *Let $\boldsymbol{H} \in \mathbb{S}_{+}^{p\times p}$, $\boldsymbol{g} \in \mathcal{R}(\boldsymbol{H})$, $\boldsymbol{\lambda} \in \mathbb{R}_{++}^{m}$, a collection of m subsets $G_i \subseteq \{1, \cdots, \sum_{i=1}^{m} n_i\}$ of size n_i each, and $\boldsymbol{A} \in \mathbb{R}^{\left(\sum_{i=1}^{m} n_i\right)\times p}$. Then the GSTO problem (3.1),*

$$\min_{\boldsymbol{x}} \quad \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} + \textstyle\sum_{i=1}^{m} \lambda_i \left|\left|\boldsymbol{A}_{G_i,*}\boldsymbol{x}\right|\right|_2, \tag{3.2}$$

is equivalent to the following $(p + \sum_{i=1}^{m} n_i)$-dimensional, concave, differentiable dual problem:

$$\begin{aligned} \max \qquad & -\tfrac{1}{2}\boldsymbol{\beta}^T\boldsymbol{H}\boldsymbol{\beta} \\ & \left|\left|\boldsymbol{\nu}_{G_i}\right|\right|_2^2 \leq \lambda_i^2 \qquad i = 1, \cdots, m \\ & \boldsymbol{g} - \boldsymbol{A}^T\boldsymbol{\nu} = \boldsymbol{H}\boldsymbol{\beta} \end{aligned} \tag{3.3}$$

and strong duality holds, that is, at the optimum, the objective in (3.2) and (3.3) are the same. Furthermore, the solutions to (3.2) and (3.3) are related by:

$$\boldsymbol{H}\boldsymbol{x}^* = -\boldsymbol{H}\boldsymbol{\beta}^*, \tag{3.4}$$

where $\boldsymbol{x}^* = \mathcal{T}_{\lambda,\boldsymbol{H}}(\boldsymbol{g})$ denotes the optimum of (3.2) and is equal to the Generalized Shrinkage Thresholding Operator evaluated at $\boldsymbol{g}$.

Proof. Using a change of variables, we can rewrite problem (3.2) as follows:

$$\begin{aligned} \min \quad & \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} + \boldsymbol{\lambda}^T\boldsymbol{t} \\ & \begin{bmatrix} -\boldsymbol{A}_{G_i,*}\boldsymbol{x} \\ -t_i \end{bmatrix} \preceq_K 0 \qquad i = 1, \cdots, m \end{aligned}$$

where $\preceq_K$ denotes the generalized inequality corresponding to the second order cone and $\boldsymbol{t} = [t_1, \cdots, t_m]^T$. The dual function of this conic constrained problem is:

$$g(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_m) = \inf_{\boldsymbol{x},\boldsymbol{t}} \frac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} + \boldsymbol{\lambda}^T\boldsymbol{t} - \sum_i \boldsymbol{\mu}_i^T \begin{bmatrix} \boldsymbol{A}_{G_i,*}\boldsymbol{x} \\ t_i \end{bmatrix} \tag{3.5}$$

where $\boldsymbol{\mu}_i \in \mathbf{R}^{n_i+1}$, $i = 1, \cdots, m$. This infimum is given by:

$$\begin{cases} -\frac{1}{2}\left(\boldsymbol{g} - \sum_i \left[\boldsymbol{A}_{G_i,*}^T \; 0\right] \boldsymbol{\mu}_i\right)^T \boldsymbol{H}^\dagger \left(\boldsymbol{g} - \sum_i \left[\boldsymbol{A}_{G_i,*}^T \; 0\right] \boldsymbol{\mu}_i\right) & \text{if } \begin{cases} [\boldsymbol{\mu}_i]_{n+1} = \lambda_i, i = 1, \cdots, m \\ \boldsymbol{g} - \sum_i \left[\boldsymbol{A}_{G_i,*}^T \; 0\right] \boldsymbol{\mu}_i \in \mathcal{R}\left(\boldsymbol{H}\right) \end{cases} \\ -\infty & \text{otherwise.} \end{cases}$$

After a simple notation change, we can write the dual problem as:

$$\begin{aligned} \max \quad & -\frac{1}{2}\left(\boldsymbol{g} - \sum_i \boldsymbol{A}_{G_i,*}^T \tilde{\boldsymbol{\nu}}_i\right)^T \boldsymbol{H}^\dagger \left(\boldsymbol{g} - \sum_i \boldsymbol{A}_{G_i,*}^T \tilde{\boldsymbol{\nu}}_i\right) \\ & \begin{bmatrix} \tilde{\boldsymbol{\nu}}_i \\ \lambda_i \end{bmatrix} \succeq_{K^*} 0 \qquad\qquad i = 1, \cdots, m \\ & \boldsymbol{g} - \sum_i \boldsymbol{A}_{G_i,*}^T \tilde{\boldsymbol{\nu}}_i \in \mathcal{R}\left(\boldsymbol{H}\right) \end{aligned}$$

where $\tilde{\boldsymbol{\nu}}_i \in \mathbf{R}^{n_i}$, $i = 1, \cdots, m$. Equivalently, letting $\boldsymbol{\nu} = \begin{bmatrix} \tilde{\boldsymbol{\nu}}_1^T, \cdots, \tilde{\boldsymbol{\nu}}_m^T \end{bmatrix}^T$, we have:

$$\begin{aligned} \max \quad & -\frac{1}{2}\left(\boldsymbol{g} - \boldsymbol{A}^T\boldsymbol{\nu}\right)^T \boldsymbol{H}^\dagger \left(\boldsymbol{g} - \boldsymbol{A}^T\boldsymbol{\nu}\right) \\ & \begin{bmatrix} \tilde{\boldsymbol{\nu}}_i \\ \lambda_i \end{bmatrix} \succeq_{K^*} 0, \; i = 1, \cdots, m \\ & \boldsymbol{g} - \boldsymbol{A}^T\boldsymbol{\nu} \in \mathcal{R}\left(\boldsymbol{H}\right). \end{aligned} \tag{3.6}$$

Slaters constraint qualification is verified and strong duality holds. The relationship between the dual and primal variables is given by the set of points at which the infimum in (3.5) is attained, which are the points in the set:

$$\left\{\boldsymbol{x}^* \in \mathbb{R}^p : \boldsymbol{H}\boldsymbol{x}^* = -\left(\boldsymbol{g} - \boldsymbol{A}^T\boldsymbol{\nu}\right)\right\}. \tag{3.7}$$

The range constraint set in (3.6) can be expressed as:

$$\left\{\boldsymbol{\nu} \in \mathbb{R}^{\sum_{i=1}^m n_i} : \boldsymbol{g} - \boldsymbol{A}^T\boldsymbol{\nu} \in \mathcal{R}\left(\boldsymbol{H}\right)\right\} = \left\{\boldsymbol{\nu} \in \mathbb{R}^{\sum_{i=1}^m n_i} : \boldsymbol{g} - \boldsymbol{A}^T\boldsymbol{\nu} = \boldsymbol{H}\boldsymbol{\beta} \text{ for some } \boldsymbol{\beta}\right\}.$$

Enforcing the above constraint in the objective in (3.6), and taking into account that the second-order cone constraint in (3.6) is equivalent to a quadratic constraint, we obtain:

$$\begin{aligned} \max \quad & -\tfrac{1}{2}\boldsymbol{\beta}^T\boldsymbol{H}\boldsymbol{\beta}. \\ & ||\boldsymbol{\nu}_{G_i}||_2^2 \leq \lambda_i^2 \qquad i = 1, \cdots, m \\ & \boldsymbol{g} - \boldsymbol{A}^T\boldsymbol{\nu} = \boldsymbol{H}\boldsymbol{\beta} \end{aligned}$$

Finally, the relationship between the primal and dual optima is given by relation (3.7):

$$\boldsymbol{H}\boldsymbol{x}^* = -\left(\boldsymbol{g} - \boldsymbol{A}^T\boldsymbol{\nu}^*\right) = -\boldsymbol{H}\boldsymbol{\beta}^*, \tag{3.8}$$

which proves (3.4). In addition, we can conclude that if $\boldsymbol{\beta}^* \in \mathcal{R}(\boldsymbol{H})$, then $\boldsymbol{x}^* = -\boldsymbol{\beta}^*$. □

This result has the interesting property of transforming an unconstrained non-differentiable problem to a constrained, twice-differentiable one: both the objective and the constraint functions in (3.3) are smooth. The price to pay is that problem (3.3) is of dimension much larger than (3.2), rendering the approach of solving to (3.2) trough (3.3) unattractive. Note however that this transformation is reminiscent of the first part of the smooth, 1-dimensional reformulation of the MSTO (Theorem II.1). We will show next that in fact a low-dimensional reformulation is also possible here, at least in the following two cases of special interest:

- *Group-ℓ_2 Penalized Linear Regression*: The Group-ℓ_2 penalized Linear Regression problem is defined as:

$$\boldsymbol{\theta}^{\text{P-LS}} = \arg\min_{\boldsymbol{\theta}} \quad \tfrac{1}{2}\left|\left|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\right|\right|_2^2 + \lambda \sum_{i=1}^m \sqrt{c_i}\left|\left|\boldsymbol{A}_{G_i,*}\boldsymbol{\theta}\right|\right|_2, \tag{3.9}$$

  with $\boldsymbol{A}$, $\{G_i\}_{i=1}^m$ satisfying the properties in the statement of Lemma III.1. We will see next that applying Lemma III.1 yields an m-dimensional reformulation of this p-dimensional problem and expresses the penalized linear regression solution $\boldsymbol{x}^{\text{P-LS}}$ as a shrinkage thresholding operation on the input vector $\boldsymbol{X}^T\boldsymbol{y}$.

- *Proximity Operator for Group-ℓ_2 penalties*: As we discussed in Section 2.3.4, the proximity operator of a convex function is an essential piece of modern proximal algorithms for large scale optimization, and is defined as:

$$\mathcal{P}_{\tau,\Omega}(\boldsymbol{g}) := \arg\min_{\boldsymbol{\theta}} \frac{1}{2\tau}\left|\left|\boldsymbol{\theta} - \boldsymbol{g}\right|\right|_2^2 + \Omega(\boldsymbol{\theta}).$$

  For the Group-ℓ_2 penalties considered here, this operator specializes to:

$$\mathcal{P}_{\lambda}(\boldsymbol{y}) \quad = \arg\min_{\boldsymbol{\theta}} \quad \frac{1}{2}\left|\left|\boldsymbol{\theta} - \boldsymbol{y}\right|\right|_2^2 + \sum_{i=1}^m \lambda_i \left|\left|\boldsymbol{A}_{G_i,*}\boldsymbol{\theta}\right|\right|_2$$

  We will see later that this problem can be reformulated into an m-dimensional

problem. The efficient evaluation of the m-dimensional reformulation will be the subject of future work.

The following result shows that the Group-ℓ_2 Penalized Linear Regression problem (3.9) is equivalent to an m-dimensional convex problem and that its solution is forced to belong to the null-space of a submatrix obtained from a subset of rows of $\boldsymbol{A}$.

Theorem III.2 (Group-ℓ_2 penalized Linear Regression)**.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} \in \mathbb{R}^n$, $\lambda > 0$, $\boldsymbol{c} \in \mathbb{R}^m_{++}$, a collection of m subsets $G_i \subseteq \{1, \cdots, p\}$ of size n_i, and $\boldsymbol{A} \in \mathbb{R}^{\left(\sum_{i=1}^m n_i\right)\times p}$. Assume further that for any subset $S \subset \{1, \cdots, \sum_{i=1}^m n_i\}$ and its complementary $\bar{S} = \{1, \cdots, \sum_{i=1}^m n_i\} \setminus S$, $\boldsymbol{A}$ verifies:*

$$Ker\left(\boldsymbol{A}_{S,\cdot}\right) \cap Ker\left(\boldsymbol{A}_{\bar{S},\cdot}\right) = \{\boldsymbol{0}\} \quad \text{and} \quad rank\left(\boldsymbol{A}\right) = p. \tag{3.10}$$

Then, the solution to the Group-ℓ_2 penalized Linear Regression problem, defined as:

$$\min_{\boldsymbol{\theta}} \quad \left\{ f\left(\boldsymbol{\theta}\right) = \tfrac{1}{2}\left|\left|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\right|\right|_2^2 + \lambda \sum_{i=1}^m \sqrt{c_i}\left|\left|\boldsymbol{A}_{G_i,*}\boldsymbol{\theta}\right|\right|_2 \right\} \tag{3.11}$$

is given by:

$$\boldsymbol{\theta}^* = -\boldsymbol{B}_D \boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\eta}^*\right) \boldsymbol{B}_D^T \boldsymbol{X}^T \boldsymbol{y}, \tag{3.12}$$

where the matrix-valued function $\boldsymbol{\Gamma}\left(\boldsymbol{\eta}\right) \succ 0$ is defined as follows:

$$\boldsymbol{\Gamma}\left(\boldsymbol{\eta}\right) = \boldsymbol{B}_D^T \left(\sum_{i\in\bar{\Omega}} \frac{1}{2\eta_i} \boldsymbol{A}_{G_i,\cdot}^T \boldsymbol{A}_{G_i,\cdot} + \boldsymbol{X}^T\boldsymbol{X} \right) \boldsymbol{B}_D, \tag{3.13}$$

and the sets D, $\bar{D}$, $\bar{\Omega}$ are defined as:

$$\begin{aligned} \Omega = \{i \in \{1, \cdots, m\} : \eta_i^* = 0\} \qquad \bar{\Omega} = \{1, \cdots, m\} \setminus \Omega. \\ D = \cup_{i\in\Omega} G_i \quad \bar{D} = \{1, \cdots, \textstyle\sum_{i=1}^m n_i\} \setminus D, \end{aligned} \tag{3.14}$$

The matrix $\boldsymbol{B}_D$ is defined as follows:

$$\boldsymbol{B}_D = \begin{cases} \text{a basis for } Ker\left(\boldsymbol{A}_{D,\cdot}\right) & \text{if } |D| > 0 \\ \boldsymbol{I}_p & \text{otherwise.} \end{cases} \tag{3.15}$$

Finally, $\boldsymbol{\eta}^*$ is the solution to the m-dimensional convex problem:

$$\begin{aligned} \min \quad & \left\{ w(\boldsymbol{\eta}) = -\tfrac{1}{2}\boldsymbol{y}^T \boldsymbol{X} \boldsymbol{B}_D \boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta}) \boldsymbol{B}_D^T \boldsymbol{X}^T \boldsymbol{y} + \lambda^2 \boldsymbol{\eta}^T \boldsymbol{c} \right\}. \\ & \boldsymbol{\eta} \succeq 0 \end{aligned} \tag{3.16}$$

and, at the optimum, we have $f(\boldsymbol{x}^*) = w(\boldsymbol{\eta}^*) + \frac{1}{2}\|\boldsymbol{y}|_2^2$.

Proof. Applying Lemma (3.2) with $\boldsymbol{H} = \boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{g} = -\boldsymbol{X}^T\boldsymbol{y}$, $\boldsymbol{\lambda} = \lambda\sqrt{\boldsymbol{c}}$ and letting $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\psi}$ it follows that (3.11) is equivalent to:

$$\begin{aligned} \max \quad & -\tfrac{1}{2}\boldsymbol{\psi}^T\boldsymbol{\psi} \\ & \boldsymbol{\nu}_{G_i}^T \boldsymbol{\nu}_{G_i} \leq \lambda^2 c_i \qquad i = 1, \cdots, m \\ & -\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{A}^T\boldsymbol{\nu} = \boldsymbol{X}^T\boldsymbol{\psi} \end{aligned} \tag{3.17}$$

The Lagrangian of this problem is given by:

$$\begin{aligned} l_1(\boldsymbol{\eta}, \boldsymbol{\omega}) = \sup_{\boldsymbol{\psi},\boldsymbol{\nu}} \quad & \left\{ -\frac{1}{2}\left(\boldsymbol{\psi}^T\boldsymbol{\psi} + 2\boldsymbol{\nu}^T\boldsymbol{C}(\boldsymbol{\eta})\boldsymbol{\nu}\right) + \boldsymbol{\omega}^T\left(\boldsymbol{A}^T\boldsymbol{\nu} + \boldsymbol{X}^T(\boldsymbol{\psi} + \boldsymbol{y})\right)\right. \\ & \left. + \lambda^2 \boldsymbol{\eta}^T \boldsymbol{c} \right\} \end{aligned} \tag{3.18}$$

where we let $\boldsymbol{C}(\boldsymbol{\eta}) := \sum_i \eta_i \boldsymbol{B}_i$ and

$$[\boldsymbol{B}_i]_{k,l} = \begin{cases} 1 & \text{if } k = l \text{ and } \sum_{j=1}^{i-1} n_j + 1 \leq k \leq \sum_{j=1}^{i} n_j \\ 0 & \text{otherwise} \end{cases}.$$

The supremum is unbounded unless $(\boldsymbol{\omega}, \boldsymbol{\eta}) \in \mathcal{D}$, where $\mathcal{D}$ is the set:

$$\mathcal{D} := \left\{ (\boldsymbol{\omega}, \boldsymbol{\eta}) : \begin{cases} \boldsymbol{A}\boldsymbol{\omega} \in \mathcal{R}(\boldsymbol{C}(\boldsymbol{\eta})) & \text{if } \boldsymbol{\eta} \neq 0, \boldsymbol{\eta} \succeq 0 \\ \boldsymbol{\omega} = 0, & \boldsymbol{\eta} = 0 \end{cases} \right\}.$$

For any pair $(\boldsymbol{\omega}, \boldsymbol{\eta}) \in \mathcal{D}$, the supremum is attained at :

$$\begin{aligned} \boldsymbol{\nu}^* &= \frac{1}{2}\boldsymbol{C}(\boldsymbol{\eta})^\dagger \boldsymbol{A}\boldsymbol{\omega} \\ \boldsymbol{\psi}^* &= \boldsymbol{X}\boldsymbol{\omega}, \end{aligned} \tag{3.19}$$

and if $\boldsymbol{\eta} \succ 0$, for each $(\boldsymbol{\omega}, \boldsymbol{\eta}) \in \mathcal{D}$, both $\boldsymbol{\psi}^*$ and $\boldsymbol{\nu}^*$ are unique. This yields:

$$l_1(\boldsymbol{\eta}, \boldsymbol{\omega}) = \begin{cases} \frac{1}{2}\boldsymbol{\omega}^T\left(\frac{1}{2}\boldsymbol{A}^T\boldsymbol{C}(\boldsymbol{\eta})^\dagger \boldsymbol{A} + \boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{\omega} + \boldsymbol{\omega}^T\boldsymbol{X}^T\boldsymbol{y} + \lambda^2\boldsymbol{\eta}^T\boldsymbol{c} & \text{if } (\boldsymbol{\omega}, \boldsymbol{\eta}) \in \mathcal{D} \\ \infty & \text{o/w.} \end{cases}$$

To continue, we need to define the following notation. Given a dual feasible $\boldsymbol{\eta}$ and index sets $\{1,\cdots,m\}$, $\{1,\cdots,p\}$ and $\{1,\cdots,\sum_{i=1}^{m} n_i\}$, define:

$$\begin{array}{cc} \Omega = \{i \in \{1,\cdots,m\} : \eta_i = 0\} & \bar{\Omega} = \{1,\cdots,m\} \setminus \Omega. \\ D = \cup_{i\in\Omega} G_i & \bar{D} = \{1,\cdots,\sum_{i=1}^{m} n_i\} \setminus D. \end{array} \tag{3.20}$$

Thus the subspace $\mathcal{R}\left(\boldsymbol{C}\left(\boldsymbol{\eta}\right)\right)$ can be characterized as:

$$\mathcal{R}\left(\boldsymbol{C}\left(\boldsymbol{\eta}\right)\right) = \left\{\boldsymbol{x} \in \mathbb{R}^{\sum_{i=1}^{m} n_i} : \boldsymbol{x}_D = \boldsymbol{0}\right\}$$

from what follows that:

$$\begin{array}{rcl} \{\boldsymbol{\omega} : \boldsymbol{A}\boldsymbol{\omega} \in \mathcal{R}\left(\boldsymbol{C}\left(\boldsymbol{\eta}\right)\right)\} & = & \{\boldsymbol{\omega} : \boldsymbol{A}_{D,\cdot}\boldsymbol{\omega} = 0\} \\ & = & \{\boldsymbol{\omega} = \boldsymbol{B}_D\boldsymbol{\phi}\} \end{array}$$

with $\boldsymbol{B}_D$ a $p \times \dim\left(\mathrm{Ker}\left(\boldsymbol{A}_{D,\cdot}\right)\right)$ matrix defined as:

$$\boldsymbol{B}_D = \begin{cases} \text{ is a basis for } \mathrm{Ker}\left(\boldsymbol{A}_{D,\cdot}\right) & \text{if } |D| > 0 \\ \boldsymbol{I}_p & \text{otherwise.} \end{cases}$$

This set equality allows us to parameterize $l_1\left(\boldsymbol{\eta},\boldsymbol{\omega}\right)$ as a function of $\boldsymbol{\eta}$ and $\boldsymbol{\phi}$, which we will denote by $l_2\left(\boldsymbol{\eta},\boldsymbol{\phi}\right)$:

$$l_2\left(\boldsymbol{\eta},\boldsymbol{\phi}\right) = \begin{cases} \frac{1}{2}\boldsymbol{\phi}^T\boldsymbol{B}_D^T\left(\frac{1}{2}\boldsymbol{A}^T\boldsymbol{C}\left(\boldsymbol{\eta}\right)^\dagger\boldsymbol{A} + \boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{B}_D\boldsymbol{\phi} + & \\ \qquad + \boldsymbol{\phi}^T\boldsymbol{B}_D^T\boldsymbol{X}^T\boldsymbol{y} + \lambda^2\boldsymbol{\eta}^T\boldsymbol{c} & \text{if } \boldsymbol{\eta} \succeq 0 \\ \infty & \text{otherwise.} \end{cases} \tag{3.21}$$

where $l_2\left(\boldsymbol{\eta},\boldsymbol{\phi}\right) = l_1\left(\boldsymbol{\eta},\boldsymbol{B}_D\boldsymbol{\phi}\right)$ and:

$$\boldsymbol{\omega}^* = \boldsymbol{B}_D\boldsymbol{\phi}^* \tag{3.22}$$

Observe that, by construction, $\boldsymbol{A}\boldsymbol{B}_D = \left[\boldsymbol{0}^T \ \left(\boldsymbol{A}_{\bar{D},\cdot}\boldsymbol{B}_D\right)^T\right]^T$, thus:

$$\boldsymbol{B}_D^T\boldsymbol{A}^T\boldsymbol{C}\left(\boldsymbol{\eta}\right)^\dagger\boldsymbol{A}\boldsymbol{B}_D = \boldsymbol{B}_D^T\boldsymbol{A}_{\bar{D},\cdot}^T\boldsymbol{C}_{\bar{D},\bar{D}}\left(\boldsymbol{\eta}_{\bar{\Omega}}\right)^{-1}\boldsymbol{A}_{\bar{D},\cdot}\boldsymbol{B}_D.$$

For fixed $\boldsymbol{\eta}$, the dual function $l_2\left(\boldsymbol{\eta},\boldsymbol{\phi}\right)$ is convex (in fact, it is also strictly convex) and quadratic in $\boldsymbol{\phi}$ and can be minimized in closed form with respect to $\boldsymbol{\phi}$. The

optimality conditions are:

$$\boldsymbol{\Gamma}(\boldsymbol{\eta})\,\boldsymbol{\phi} = -\boldsymbol{B}_D^T \boldsymbol{X}^T \boldsymbol{y}. \tag{3.23}$$

where we define the matrix $\boldsymbol{\Gamma}(\boldsymbol{\eta})$ as:

$$\boldsymbol{\Gamma}(\boldsymbol{\eta}) = \boldsymbol{B}_D^T \left(\frac{1}{2} \boldsymbol{A}_{\bar{D},\cdot}^T \boldsymbol{C}_{\bar{D},\bar{D}} (\boldsymbol{\eta}_{\bar{\Omega}})^{-1} \boldsymbol{A}_{\bar{D},\cdot} + \boldsymbol{X}^T \boldsymbol{X} \right) \boldsymbol{B}_D \tag{3.24}$$

The matrix $\boldsymbol{\Gamma}(\boldsymbol{\eta})$ is non-singular and hence the above system has a unique solution $\boldsymbol{\phi}^* = -\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta})\,\boldsymbol{B}_D^T \boldsymbol{X}^T \boldsymbol{y}$. The positive definiteness of $\boldsymbol{\Gamma}(\boldsymbol{\eta})$ follows from the fact that

$$\boldsymbol{B}_D^T \boldsymbol{A}_{\bar{D},\cdot}^T \boldsymbol{C}_{\bar{D},\bar{D}} (\boldsymbol{\eta}_{\bar{\Omega}})^{-1} \boldsymbol{A}_{\bar{D},\cdot} \boldsymbol{B}_D \succ 0.$$

To prove this last assertion, notice that since $\boldsymbol{C}_{\bar{D},\bar{D}}^{-1}(\boldsymbol{\eta}_{\bar{\Omega}}) \succ 0$, a necessary and sufficient condition for this to hold is that $\mathrm{Ker}\left(\boldsymbol{A}_{\bar{D},\cdot}\right) \cap \mathcal{R}(\boldsymbol{B}_D) = \emptyset$. If $|D| = 0$, then $\mathrm{Ker}\left(\boldsymbol{A}_{\bar{D},\cdot}\right) = \emptyset$ by the second assumption in (3.10). If $|D| > 0$, $\boldsymbol{B}_D$ is a basis for $\mathrm{Ker}(\boldsymbol{A}_{D,\cdot})$, and the former condition is equivalent to requiring that $\mathrm{Ker}(\boldsymbol{A}_{D,\cdot}) \cap \mathrm{Ker}\left(\boldsymbol{A}_{\bar{D},\cdot}\right) = \emptyset$, which is guaranteed by the first assumption in (3.10).

Plugging in the optimal $\boldsymbol{\phi}^*$ verifying (3.23) in (3.21), leads us to the (second) dual function:

$$\begin{aligned} w(\boldsymbol{\eta}) &= \min_{\boldsymbol{\phi}} l_2(\boldsymbol{\eta}, \boldsymbol{\phi}) \\ &= -\frac{1}{2} \boldsymbol{y}^T \boldsymbol{X} \boldsymbol{B}_D \boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta})\,\boldsymbol{B}_D^T \boldsymbol{X}^T \boldsymbol{y} + \lambda^2 \boldsymbol{\eta}^T \boldsymbol{c} \end{aligned}$$

with domain $\mathrm{dom}\, w(\boldsymbol{\eta}) = \{\boldsymbol{\eta} \succeq 0\}$. Hence the dual problem of (3.11) is given by:

$$\begin{aligned} \min \quad & w(\boldsymbol{\eta}), \\ & \boldsymbol{\eta} \succeq 0 \end{aligned}$$

which proves (3.16). Slaters constraint qualifications are verified on (3.17) (take for instance $\boldsymbol{\psi} = -\boldsymbol{y}$ and $\boldsymbol{\nu} = \boldsymbol{0}$) and hence strong duality holds. On the other hand, the optimality conditions (3.7) from the proof of Lemma III.1, (3.22) and (3.19) imply

that:

$$\begin{aligned}\left(\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{x}^* &= -\left(\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{\beta}^* \\ &= -\boldsymbol{X}^T\boldsymbol{\psi}^* \\ &= -\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\omega}^* \\ &= -\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}_D\boldsymbol{\phi}^*\end{aligned}$$

which proves (3.12). □

The above result shows that our Group-ℓ_2 penalty extends the LASSO and Group LASSO penalties in that it generalizes coordinate-wise sparsity. Indeed, the solution $\boldsymbol{x}^*$ to (3.11) is given as a linear combination of elements of a basis $\boldsymbol{B}_D$ of the kernel of the matrix $\boldsymbol{A}_D$. This implies that the elements in $\boldsymbol{x}^*$ are not necessarily zero, however, their projection on $\boldsymbol{A}_D$ will be, where D is determined by the inactive set of shrinkage variables $\eta_i^* > 0$. It is worth observing that the assumption (3.10) could be relaxed but is loose enough for most practical purposes.

To illustrate the specialization of our result to two well known cases, we now consider the following Non-overlapping and Overlapping Group LASSO-penalized linear regression problems.

3.2.1 Non-overlapping Group LASSO

Corollary III.3 (Group LASSO-penalized Linear Regression)**.** *Given a collection of m groups of indices $G_i \subseteq \{1, \cdots, p\}$ of size $n_i = |G_i|$ satisfying $\cup_{i=1}^m G_i = \{1, \cdots, p\}$ and $G_i \cap G_j = \emptyset$ for $i \neq j$, the LASSO (Tib96) or separable Group-LASSO problem (YL06b), is defined as:*

$$\min_{\boldsymbol{\theta}} \quad \tfrac{1}{2}\left|\left|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\right|\right|_2^2 + \lambda \sum_{i=1}^m \sqrt{c_i}\left|\left|\boldsymbol{\theta}_{G_i}\right|\right|_2. \tag{3.25}$$

The solution to this problem is given by:

$$\boldsymbol{\theta}^*_{\bar{Z}} = -\left(\boldsymbol{S}\left(\boldsymbol{\eta}^*\right) + \boldsymbol{X}^T_{\bar{Z},\cdot}\boldsymbol{X}_{\bar{Z},\cdot}\right)^{-1}\boldsymbol{X}^T_{\bar{Z},\cdot}\boldsymbol{y}, \tag{3.26}$$

$$\boldsymbol{\theta}^*_{Z} = \boldsymbol{0}, \tag{3.27}$$

where the index subsets Z and $\bar{Z}$ are given by:

$$Z = \cup_{i:\eta_i^*=0} G_i \quad \bar{Z} = \{1, \cdots, p\} \setminus Z,$$

and $\boldsymbol{S}(\boldsymbol{\eta})$ *is a* $|\bar{Z}| \times |\bar{Z}|$ *diagonal shrinkage matrix, with elements:*

$$[\boldsymbol{S}(\boldsymbol{\eta})]_{k,k} = \frac{1}{2\eta_i}, \text{ with } i \text{ such that } \bar{Z}_k \in G_i, \tag{3.28}$$

and $\boldsymbol{\eta}^*$ *is the solution to the m-dimensional convex problem:*

$$\begin{aligned} \min \quad & \left\{ w(\boldsymbol{\eta}) = \tfrac{1}{2}\boldsymbol{\theta}^{*T}\boldsymbol{X}^T\boldsymbol{y} + \lambda^2 \boldsymbol{\eta}^T \boldsymbol{c} \right\}. \\ & \boldsymbol{\eta} \succeq 0 \end{aligned} \tag{3.29}$$

The proof follows by application of Theorem III.2 to this specific choice of $\boldsymbol{A}$. See Appendix B.1 for details.

3.2.2 Overlapping Group LASSO

Corollary III.4 (Overlapping Group LASSO penalized Linear Regression)**.** *Let* $G_i \subseteq \{1, \cdots, p\}$, $i = 1, \cdots, m$, *be a collection of groups of variables, with* $\cup_{i=1}^m G_i = \{1, \cdots, p\}$ *and* $n_i = |G_i|$. *Associate to each group* G_i *a positive vector of weights* $\boldsymbol{w}_i \in \mathbb{R}^{n_i}_{++}$. *The structured-sparse (JOB10) or hierarchical Group-LASSO problems (ZRY09), defined as:*

$$\min_{\boldsymbol{\theta}} \quad \tfrac{1}{2}\left|\left|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\right|\right|_2^2 + \lambda \sum_{i=1}^m \sqrt{c_i}\left|\left| diag(\boldsymbol{w}_i)\,\boldsymbol{\theta}_{G_i}\right|\right|_2, \tag{3.30}$$

has a unique solution given by:

$$\boldsymbol{\theta}^*_{\bar{Z}} = -\left(\boldsymbol{S}(\boldsymbol{\eta}^*) + \boldsymbol{X}^T_{\bar{Z},\cdot}\boldsymbol{X}_{\bar{Z},\cdot}\right)^{-1}\boldsymbol{X}^T_{\bar{Z},\cdot}\boldsymbol{y}, \tag{3.31}$$

$$\boldsymbol{\theta}^*_{Z} = \boldsymbol{0}, \tag{3.32}$$

where the index subsets Z *and* $\bar{Z}$ *are given by:*

$$Z = \cup_{i:\eta_i^*=0} G_i \quad \bar{Z} = \{1, \cdots, p\} \setminus Z,$$

and $\boldsymbol{S}(\boldsymbol{\eta})$ *is a* $|\bar{Z}| \times |\bar{Z}|$ *diagonal shrinkage matrix, with elements:*

$$[\boldsymbol{S}(\boldsymbol{\eta})]_{k,k} = \sum_{i:\bar{Z}_k \in G_i} \frac{1}{2\eta_i}[\boldsymbol{w}_i]^2_{\bar{Z}_k}, \tag{3.33}$$

and $\boldsymbol{\eta}^*$ *is the solution to the* m*-dimensional convex problem:*

$$\begin{aligned} \min \quad & \left\{ w\left(\boldsymbol{\eta}\right) = \tfrac{1}{2}\boldsymbol{\theta}^{*T}\boldsymbol{X}^T\boldsymbol{y} + \lambda^2\boldsymbol{\eta}^T\boldsymbol{c} \right\}. \\ & \boldsymbol{\eta} \succeq 0 \end{aligned} \tag{3.34}$$

The proof is an application of Theorem III.2 to this specific choice of $\boldsymbol{A}$. See Appendix B.2 for details.

3.2.3 Proximity operator for arbitrary Group-ℓ_2 penalties

We showed in Section 2.3.4 that the proximity operator of a (possibly non-differentiable) convex function $\Omega\left(\boldsymbol{x}\right)$ is defined as (Mor65), (CW06):

$$\mathcal{P}_\tau\left(\boldsymbol{g}\right) := \arg\min_{\boldsymbol{x}} \frac{1}{2\tau}\left|\left|\boldsymbol{x}-\boldsymbol{g}\right|\right|_2^2 + \Omega\left(\boldsymbol{x}\right),$$

and is omnipresent in efficient, large-scale algorithms for solving problems of the type:

$$\arg\min_{\boldsymbol{\theta}} f\left(\boldsymbol{\theta}\right) + \Omega\left(\boldsymbol{\theta}\right),$$

where $f\left(\boldsymbol{\theta}\right)$ is a convex, differentiable function with Lipschitz gradient. Example of such functions in statistical learning are the logistic or the poisson regression losses.

In previous work, various authors have derived efficient algorithms to evaluate $\mathcal{P}_\tau\left(\boldsymbol{g}\right)$ for the Group LASSO (LY10) and the Hierarchical LASSO penalties (JMOB10). The next result shows that evaluating the proximity operator for the general class of group-ℓ_2 penalties can be done through the solution of an m dimensional problem, where m is the number of ℓ_2 norm terms in the penalty. In addition, similarly to (LY10), we show that this problem is differentiable and hence efficient algorithms for smooth optimization can be applied to evaluate $\mathcal{P}_\tau\left(\boldsymbol{g}\right)$ for the general group-ℓ_2 penalties we introduced in Section 3.2.

Theorem III.5 (Proximity operator for arbitrary Group-ℓ_2 penalties)**.** *Let* $\boldsymbol{y} \neq \boldsymbol{0} \in \mathbb{R}^p$*,* $\boldsymbol{\lambda} \in \mathbb{R}^m_{++}$*, a collection of* m *subsets* $G_i \subseteq \{1, \cdots, p\}$ *of size* n_i*, and* $\boldsymbol{A} \in \mathbb{R}^{\sum_{i=1}^m n_i \times p}$ *such that* $\boldsymbol{A}\boldsymbol{A}^T \succ \boldsymbol{0}$*. Then the proximal operator for the overlapping group lasso penalty, defined as:*

$$\mathcal{P}_\lambda\left(\boldsymbol{y}\right) \quad = \arg\min_{\boldsymbol{\theta}} \quad \frac{1}{2}\left|\left|\boldsymbol{\theta}-\boldsymbol{y}\right|\right|_2^2 + \sum_{i=1}^m \lambda_i \left|\left|\boldsymbol{A}_{G_i,*}\boldsymbol{\theta}\right|\right|_2 \tag{3.35}$$

is given by:

$$\mathcal{P}_\lambda(\boldsymbol{y}) = \left(\boldsymbol{I} - \boldsymbol{A}^T\left(\boldsymbol{A}\boldsymbol{A}^T + 2\boldsymbol{C}(\boldsymbol{\eta}^*)\right)^{-1}\boldsymbol{A}\right)\boldsymbol{y} \tag{3.36}$$

where $\boldsymbol{C}(\boldsymbol{\eta}) := \sum_i \eta_i \boldsymbol{B}_i$*, with* $\boldsymbol{B}_i$ *defined in (3.44), and* $\boldsymbol{\eta}^*$ *is the solution to the* m*-dimensional convex problem:*

$$\begin{aligned} \min \quad & \left\{ l(\boldsymbol{\eta}) = \tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{A}^T\left(\boldsymbol{A}\boldsymbol{A}^T + 2\boldsymbol{C}(\boldsymbol{\eta})\right)^{-1}\boldsymbol{A}\boldsymbol{y} + \boldsymbol{\eta}^T\boldsymbol{\lambda}^2 - \tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{y}. \right\} \\ & \boldsymbol{\eta} \succeq 0 \end{aligned} \tag{3.37}$$

Furthermore, it holds that:

$$\left\{ \boldsymbol{v} \in \mathbb{R}^m : v_i = \boldsymbol{\nu}^{*T}\boldsymbol{B}_i\boldsymbol{\nu}^* - \lambda_i^2 \right\} \in \partial l(\boldsymbol{\eta}) \tag{3.38}$$

where $\partial l(\boldsymbol{\eta})$ *denotes the subdifferential of* $l(\boldsymbol{\eta})$ *and the* $(\sum_{i=1}^m n_i)$*-dimensional vector* $\boldsymbol{\nu}^*$ *is defined as:*

$$\boldsymbol{\nu}^* = -\left(\boldsymbol{A}\boldsymbol{A}^T + 2\boldsymbol{C}(\boldsymbol{\eta})\right)^{-1}\boldsymbol{A}\boldsymbol{y}. \tag{3.39}$$

The proof of this result is given in Appendix B.5.

3.3 Algorithms

By Theorem III.2, the evaluation of the GSTO when $\boldsymbol{H} = \boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{g} = \boldsymbol{X}^T\boldsymbol{y}$ requires the solution of the m-dimensional convex problem:

$$\begin{aligned} \min \quad & \left\{ w(\boldsymbol{\eta}) = -\tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{B}_D^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta})\boldsymbol{B}_D\boldsymbol{X}^T\boldsymbol{y} + \lambda^2\boldsymbol{\eta}^T\boldsymbol{c} \right\}. \\ & \boldsymbol{\eta} \succeq 0 \end{aligned} \tag{3.40}$$

The function $w(\boldsymbol{\eta})$ has domain $\operatorname{dom} w(\boldsymbol{\eta}) = \mathbb{R}_+^m$ and it is continuous but non-differentiable at the boundary of the feasible set. We can nonetheless show that it is actually differentiable in the interior of its domain, $\boldsymbol{\eta} \succ 0$.

Theorem III.6. *The function* $w(\boldsymbol{\eta})$*, defined in (3.40), with* $\boldsymbol{\Gamma}(\boldsymbol{\eta})$ *defined in the statement of Theorem III.2, is twice differentiable in the interior of its domain, and its gradient and Hessian are given by:*

$$\left[\nabla_{\boldsymbol{\eta}} w(\boldsymbol{\eta})\right]_i = \lambda^2 c_i - ||\boldsymbol{\nu}_{G_i}||_2^2, \tag{3.41}$$

and:

$$\left[\nabla^2_{\boldsymbol{\eta}} w\left(\boldsymbol{\eta}\right)\right]_{i,j} = \begin{cases} \frac{2}{\eta_i}\left|\left|\boldsymbol{\nu}_{G_i}\right|\right|_2^2 - \frac{1}{\eta_i^2}\boldsymbol{\nu}^T\boldsymbol{B}_i\boldsymbol{A}\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\eta}\right)\boldsymbol{A}^T\boldsymbol{B}_i\boldsymbol{\nu} & i = j \\ -\frac{1}{\eta_i\eta_j}\boldsymbol{\nu}^T\boldsymbol{B}_i\boldsymbol{A}\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\eta}\right)\boldsymbol{A}^T\boldsymbol{B}_j\boldsymbol{\nu} & i \neq j \end{cases} \tag{3.42}$$

where

$$\boldsymbol{\nu} = -\frac{1}{2}\boldsymbol{C}^{-1}\left(\boldsymbol{\eta}\right)\boldsymbol{A}\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\eta}\right)\boldsymbol{X}^T\boldsymbol{y} \tag{3.43}$$

and $\boldsymbol{C}\left(\boldsymbol{\eta}\right) := \sum_i \eta_i \boldsymbol{B}_i$*, with:*

$$\left[\boldsymbol{B}_i\right]_{k,l} = \begin{cases} 1 & \text{if } k = l \text{ and } \sum_{j=1}^{i-1} n_j + 1 \leq k \leq \sum_{j=1}^{i} n_j \\ 0 & \text{otherwise} \end{cases} . \tag{3.44}$$

See Appendix B.3 for a proof of this result.

An immediate consequence of this theorem is that computing a subgradient of $w\left(\boldsymbol{\eta}\right)$ in the interior of its domain only requires evaluating the primal candidate (3.43). Evaluating the Hessian seems a priori computationally demanding since we need to solve a system of the type

$$\boldsymbol{\Gamma}\left(\boldsymbol{\eta}\right)\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{B}_j\boldsymbol{\nu}$$

for $\boldsymbol{x}$ for each $j = 1, \cdots, m$. Unfortunately, it is not straightforward to generalize our theory to compute subgradients of $w\left(\boldsymbol{\eta}\right)$ for $\boldsymbol{\eta}$ lying at the boundary of its domain. To circumvent this problem we propose to approximate $\boldsymbol{\eta}^*$, the solution to (3.40), by solving instead two perturbed versions of the original problem (3.40).

3.3.1 Subgradient method on a restricted parameter space

In this section we propose to approximate $\boldsymbol{\eta}^*$ by solving instead the following problem:

$$\begin{aligned} \min \quad & \left\{w\left(\boldsymbol{\eta}\right) = -\tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{B}_D^T\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\eta}\right)\boldsymbol{B}_D\boldsymbol{X}^T\boldsymbol{y} + \lambda^2\boldsymbol{\eta}^T\boldsymbol{c}\right\}. \\ & \boldsymbol{\eta} \succeq \epsilon\boldsymbol{1} \end{aligned} \tag{3.45}$$

for a very small ϵ. We denote the solution to the perturbed problem (3.45) by $\boldsymbol{\eta}^*_\epsilon$. The solution to the primal will then be approximated by thresholding the elements in $\boldsymbol{\eta}^*_\epsilon$ with magnitudes smaller than or equal to ϵ and plugging the result into the

expression for $\boldsymbol{\theta}^*$ given in by Theorem III.2:

$$\boldsymbol{\theta}_\epsilon^* = -\boldsymbol{B}_D \boldsymbol{\Gamma}^{-1} \left(\mathcal{T}_{\boldsymbol{I},\epsilon} \left(\boldsymbol{\eta}_\epsilon^* \right) \right) \boldsymbol{B}_D^T \boldsymbol{X}^T \boldsymbol{y}, \tag{3.46}$$

Our strategy is to solve (3.45) using a subgradient method. From Theorem III.6, computing a subgradient only requires to evaluate (3.43) at each iteration, which can be done efficiently using the Sherman-Morrison-Woodbury matrix inversion formula when n (the number of rows of $\boldsymbol{X}$) is small (and hence $\boldsymbol{X}^T\boldsymbol{X}$ is of rank n). Since we also need to respect the constraint $\boldsymbol{\eta} \succeq \epsilon \mathbf{1}$, we will use a projected subgradient strategy (Ber99). Starting from a feasible $\boldsymbol{\eta}^1 \succ \epsilon \mathbf{1}$, the projected subgradient algorithm generates a sequence:

$$\boldsymbol{\eta}^{t+1} = P_\epsilon \left(\boldsymbol{\eta}^t - \alpha_t \nabla_{\boldsymbol{\eta}} w \left(\boldsymbol{\eta}^t \right) \right). \tag{3.47}$$

where $P_\epsilon(\cdot)$ denotes the projector operator onto the set $\boldsymbol{\eta} \succeq \epsilon \mathbf{1}$ and α_t is a step size chosen as follows:

$$\alpha_t = \frac{w\left(\boldsymbol{\eta}^t\right) - w_t^*}{||\nabla_{\boldsymbol{\eta}} w\left(\boldsymbol{\eta}^t\right)||_2^2} \tag{3.48}$$

where w_t^* is a lower bound of the optimal value $w\left(\boldsymbol{\eta}_\epsilon^*\right)$. This lower bound can be dynamically updated following the strategy described in (BNO+03), Section 8.2.1. Since $||\nabla_{\boldsymbol{\eta}} w\left(\boldsymbol{\eta}^t\right)||_2^2 < \infty$ for any $\boldsymbol{\eta} \leq \epsilon \mathbf{1}$, the sequence $\boldsymbol{\eta}^t$ can be shown to verify ((Ber99), Proposition 8.2.8.)

$$\inf \left\{ w\left(\boldsymbol{\eta}^t\right) \right\}_{t \geq 1} \leq w\left(\boldsymbol{\eta}^*\right) + \delta$$

for some small $\delta > 0$, that depends on the parameters used to estimate w_t^*. Notice that the subgradient method is not a descent method, and hence at the end of our iterative procedure we will set:

$$\hat{\boldsymbol{\eta}}_\epsilon^* = \arg\min \left\{ w\left(\boldsymbol{\eta}^t\right) \right\}_{t \geq 1}.$$

In practice, we will stop the algorithm after a fix number of iterations or whenever the duality gap between (3.17) and (3.16), given by

$$\left| w\left(\boldsymbol{\eta}^t\right) - \left(-\frac{1}{2} \boldsymbol{\theta}^t \boldsymbol{\theta}^t \right) \right|$$

for a feasible $\boldsymbol{\theta}^t$ computed through (3.19), (3.22) and (3.23), is smaller than a pre-specified tolerance.

Note finally that the sequence (3.47) is very similar in structure to a projected gradient descent strategy. In this particular case, the only subgradient of $w(\boldsymbol{\eta})$ over $\boldsymbol{\eta} \succ \mathbf{0}$ is given by $\nabla_{\boldsymbol{\eta}} w(\boldsymbol{\eta})$.

3.3.2 Projected Newton method on regularized dual problem

The subgradient approach is very simple and computationally light, but it is known to have slow convergence properties that hinder its application whenever good accuracy is necessary. In this section we propose a different approach to approximate $\boldsymbol{\eta}^*$ which consists of solving a perturbed version of the dual problem (3.17), namely:

$$\begin{aligned} \max \quad & -\tfrac{1}{2}\boldsymbol{\psi}^T\boldsymbol{\psi} + \epsilon\boldsymbol{\nu}^T\boldsymbol{\nu}, \\ & ||\boldsymbol{\nu}_{G_i}||_2^2 \leq \lambda^2 c_i \qquad i = 1, \cdots, m \\ & -\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{A}^T\boldsymbol{\nu} = \boldsymbol{X}^T\boldsymbol{\psi} \end{aligned} \tag{3.49}$$

where $\epsilon > 0$ is a small parameter that we set to 10^{-9} in practice. Notice that when $\epsilon = 0$, the above problem is exactly the same as the original dual, (3.17), and that their feasible set is the same for all $\epsilon \geq 0$. For $\epsilon > 0$, problem (3.49) is strictly convex, and following the exact same development of Theorem III.2, we can show that strong duality holds and that its Lagrange dual is exactly given by:

$$\begin{aligned} \min \quad & \left\{ w_\epsilon(\boldsymbol{\eta}) := -\tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta} + \epsilon\mathbf{1})\boldsymbol{X}^T\boldsymbol{y} + \lambda^2\boldsymbol{\eta}^T\boldsymbol{c} \right\}, \\ & \boldsymbol{\eta} \succeq \mathbf{0} \end{aligned} \tag{3.50}$$

Since $\epsilon > 0$, according to Theorem III.6 the function $w(\boldsymbol{\eta} + \epsilon\mathbf{1})$ is twice differentiable over the feasible set $\boldsymbol{\eta} \succeq \mathbf{0}$, and thus so is the objective $w_\epsilon(\boldsymbol{\eta})$ in (3.50). Differentiability allows Newton-Raphson type methods, which are known to enjoy fast convergence properties, to be applied to our problem.

The optimal value in (3.49) is relatively robust to the choice of ϵ, and, for small ϵ, it is bound to be close to the optimal value of the original problem (3.17), as we will show next. Observe that one can bound the objective in 3.49 for every feasible pair $(\boldsymbol{\psi}, \boldsymbol{\nu})$ as follows. First notice that, for any feasible pair $(\boldsymbol{\psi}, \boldsymbol{\nu})$,

$$-\frac{1}{2}\boldsymbol{\psi}^T\boldsymbol{\psi} \leq -\frac{1}{2}\boldsymbol{\psi}^{*T}\boldsymbol{\psi}^*$$

where $(\boldsymbol{\psi}^*, \boldsymbol{\nu}^*)$ denotes the solution to the original problem, with $\epsilon = 0$. Choose $(\tilde{\boldsymbol{\psi}}, \tilde{\boldsymbol{\nu}})$ to be the optimum of (3.49) for any $\epsilon > 0$. Then:

$$-\frac{1}{2}\tilde{\boldsymbol{\psi}}^T\tilde{\boldsymbol{\psi}} \leq -\frac{1}{2}\boldsymbol{\psi}^{*T}\boldsymbol{\psi}^*$$

and

$$-\frac{1}{2}\tilde{\boldsymbol{\psi}}^T\tilde{\boldsymbol{\psi}} + \epsilon\tilde{\boldsymbol{\nu}}^T\tilde{\boldsymbol{\nu}} \leq -\frac{1}{2}\boldsymbol{\psi}^{*T}\boldsymbol{\psi}^* + \epsilon\tilde{\boldsymbol{\nu}}^T\tilde{\boldsymbol{\nu}}$$

Now, since $\tilde{\boldsymbol{\nu}}$ is feasible, $\epsilon\tilde{\boldsymbol{\nu}}^T\tilde{\boldsymbol{\nu}}$ is upper bounded by $\epsilon\lambda^2\boldsymbol{c}^T\mathbf{1}$, which implies that:

$$-\frac{1}{2}\tilde{\boldsymbol{\psi}}^T\tilde{\boldsymbol{\psi}} + \epsilon\tilde{\boldsymbol{\nu}}^T\tilde{\boldsymbol{\nu}} \leq -\frac{1}{2}\boldsymbol{\psi}^{*T}\boldsymbol{\psi}^* + \epsilon\lambda^2\boldsymbol{c}^T\mathbf{1}$$

On the other hand, we also have:

$$-\frac{1}{2}\boldsymbol{\psi}^T\boldsymbol{\psi} + \epsilon\boldsymbol{\nu}^T\boldsymbol{\nu} \leq -\frac{1}{2}\tilde{\boldsymbol{\psi}}^T\tilde{\boldsymbol{\psi}} + \epsilon\tilde{\boldsymbol{\nu}}^T\tilde{\boldsymbol{\nu}},$$

for any feasible pair $(\boldsymbol{\psi}, \boldsymbol{\nu})$, which allows us to conclude that:

$$-\frac{1}{2}\boldsymbol{\psi}^{*T}\boldsymbol{\psi}^* \leq -\frac{1}{2}\boldsymbol{\psi}^{*T}\boldsymbol{\psi}^* + \epsilon\boldsymbol{\nu}^{*T}\boldsymbol{\nu}^* \leq -\frac{1}{2}\tilde{\boldsymbol{\psi}}^T\tilde{\boldsymbol{\psi}} + \epsilon\tilde{\boldsymbol{\nu}}^T\tilde{\boldsymbol{\nu}} \leq -\frac{1}{2}\boldsymbol{\psi}^{*T}\boldsymbol{\psi}^* + \epsilon\lambda^2\boldsymbol{c}^T\mathbf{1}.$$

Since strong duality holds, the above inequality allows us to conclude that, the optimum of (3.50) satisfies:

$$|w_\epsilon(\boldsymbol{\eta}^*) - w(\boldsymbol{\eta}_0^*)| \leq \epsilon\lambda^2\boldsymbol{c}^T\mathbf{1}$$

where $\boldsymbol{\eta}_0^*$ is the solution to (3.50) for $\epsilon = 0$, that is, the solution to the original problem (3.40). Thus, choosing a small ϵ will necessarily lead to a solution with objective arbitrarily close to the original problem optimum.

To solve (3.49), we will use Bertsekas Projected Newton method (Ber82), which takes advantage of the fact that the objective in (3.49) is smooth, and that the constraint set $\boldsymbol{\eta} \succeq \mathbf{0}$ is very simple. Essentially, at each iteration t, the projected newton method consists of three steps. First, a candidate descent direction is computed:

$$\Delta\boldsymbol{\eta} = -\boldsymbol{D}_t^{-1}\nabla w_\epsilon\left(\boldsymbol{\eta}^t\right). \tag{3.51}$$

where $\boldsymbol{D}_t$ is a matrix constructed from the Hessian of $w_\epsilon(\boldsymbol{\eta})$, given in (3.42), evaluated at $\boldsymbol{\eta}^t$ (though not the Hessian itself as in traditional Newton-Raphson descent). The

descent direction leads to a possibly non-feasible update:

$$\bar{\boldsymbol{\eta}}_\alpha = \boldsymbol{\eta}^t - \alpha \boldsymbol{D}_t^{-1} \nabla w_\epsilon \left(\boldsymbol{\eta}^t \right), \tag{3.52}$$

and finally, the step size α is selected so that the next feasible iterate, defined as:

$$\boldsymbol{\eta}^{t+1} = \left[\bar{\boldsymbol{\eta}}_\alpha \right]_+, \tag{3.53}$$

where $[\cdot]_+$ denotes the projection to the positive orthant, verifies an Armijo-type rule that guarantees the descent at each iteration, even when far from the optimum.

3.3.3 Homotopy: path-following strategy for the shrinkage variables

The solutions to the Group-ℓ_2 penalized regression problem (3.11), denoted by $\boldsymbol{x}_\lambda^*$, depend on the penalty parameter λ. In most statistical learning problems, we need to compute the shrinkage variables $\boldsymbol{\eta}^*$ and the primal solution $\boldsymbol{x}_\lambda^*$ for several values of this penalty parameter, to construct the *regularization path*, i.e. the trajectory of $\boldsymbol{x}_\lambda^*$ as a function of λ.

Usually, a discretization approach is taken, where the user specifies a grid $\{\lambda_k\}_{k=1}^K$ of candidate penalty parameters and then computes $\boldsymbol{x}_{\lambda_k}^*$ for each penalty parameter in the grid. In the special case of ℓ_1, or LASSO penalties, it has been shown that the regularization path is piecewise linear, hence an efficient algorithm is to compute $\boldsymbol{x}_{\lambda_k}^*$ only at the breaking points where variables enter or leave the active set (OPT00; EHJT04).

For general Group-ℓ_2 penalties, the path is no longer linear, as we will show below. In most current algorithms, to reduce the computation time needed to compute the regularization path $\left\{ \boldsymbol{\theta}_{\lambda_k}^* \right\}_k$, a warm start/continuation strategy is usually employed, where the algorithm to compute $\boldsymbol{\theta}_{\lambda_{k+1}}^*$ is initialized with $\boldsymbol{\theta}_{\lambda_k}^*$ (MVVR10; LY10).

Nonetheless, in our context, it is reasonable to ask whether for two different choices of λ that are close to each other, the values of $\boldsymbol{\eta}^*$, and hence those of $\boldsymbol{\theta}$, do not change too much. The following result shows that this is indeed the case for the active variables $\eta_i^* > 0$, so long as the Hessian of $w(\boldsymbol{\eta})$ restricted to the active variables is well-behaved.

Theorem III.7. *Consider the setting and assumptions of Theorem III.2. For any $\boldsymbol{\eta} \succeq \mathbf{0}$, let $\bar{\Omega}$ be the set of active variables of $\boldsymbol{\eta}$, that is $\bar{\Omega} = \{1, \cdots, m\} \setminus \Omega$, with $\Omega = \{i \in \{1, \cdots, m\} : \eta_i = 0\}$. Define the restriction of $w(\boldsymbol{\eta})$ to the active variables*

as follows:

$$w_{\bar{\Omega}}\left(\boldsymbol{\eta}_{\bar{\Omega}}\right) = -\frac{1}{2}\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{B}_D^T\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\eta}_{\bar{\Omega}}\right)\boldsymbol{B}_D\boldsymbol{X}^T\boldsymbol{y} + \lambda^2\boldsymbol{\eta}_{\bar{\Omega}}^T\boldsymbol{c} \tag{3.54}$$

where $\boldsymbol{B}_D$ *is a basis for* $Ker(\boldsymbol{A}_{D,\cdot})$ *and* $D = \cup_{i\in\Omega}G_i$*. Denote by* $\boldsymbol{\eta}^*$ *the solution to the m-dimensional convex problem (3.16). Then, the function* $w_{\bar{\Omega}}\left(\boldsymbol{\eta}_{\bar{\Omega}}\right)$ *is twice differentiable for any* $\boldsymbol{\eta}_{\bar{\Omega}} \succ \mathbf{0}$ *and, provided that* $\nabla^2 w_{\bar{\Omega}}\left(\boldsymbol{\eta}_{\bar{\Omega}}^*\right)$ *is non-singular, the following holds:*

$$\frac{d\boldsymbol{\eta}_{\bar{\Omega}}^*}{d\lambda} = -2\lambda\left(\nabla^2 w_{\bar{\Omega}}\left(\boldsymbol{\eta}_{\bar{\Omega}}^*\right)\right)^{-1}\boldsymbol{c}_{\bar{\Omega}}. \tag{3.55}$$

See Appendix B.4 for a proof of this theorem.

This result suggests a strategy to obtain a good initialization value for our iterative algorithms from the solution obtained for λ_{k-1}. Hence, we will predict $\boldsymbol{\eta}^k$ and $\boldsymbol{x}_{\lambda_k}$ at the k-th point of the regularization path, from the computed values at the $(k-1)$-th iteration, using an update rule based on relation (3.55). Specifically, we will set:

$$\begin{aligned}\boldsymbol{\eta}_{\bar{\Omega}}^k &= \left[\boldsymbol{\eta}_{\lambda_{k-1}}^*\right]_{\bar{\Omega}} - 2\left(\lambda_k - \lambda_{k-1}\right)\lambda_{k-1}\left(\nabla^2 w_{\bar{\Omega}}\left(\left[\boldsymbol{\eta}_{\lambda_{k-1}}^*\right]_{\bar{\Omega}}\right)\right)^{-1}\boldsymbol{c}_{\bar{\Omega}}, \\ \boldsymbol{\eta}_{\Omega}^k &= \mathbf{0}.\end{aligned} \tag{3.56}$$

Here $\bar{\Omega}$ denotes the indices of non-zero elements of $\boldsymbol{\eta}_{\lambda_{k-1}}^*$, computed at λ_{k-1}. These predictions will then be used to initialize our Projected Newton algorithm and obtain the optimal pair $\boldsymbol{\eta}_{\lambda_k}^*$ and $\boldsymbol{\theta}_{\lambda_k}^*$ at λ_k. The Projected Newton step converges in practice in very few iterations, as long as the difference between λ_k and λ_{k-1} is reasonably small.

3.4 Numerical Results

In this section we first evaluate the numerical perfomance of our algorithms to evaluate the GSTO for overlapping Group LASSO penalized linear regression problem. Second, we investigate the application of the GSTO to the multi-task learning problem of predicting symptom scores from the gene expression responses of symptomatic individuals.

3.4.1 Evaluation of the GSTO

We study here the numerical performance of our algorithm in solving the overlapping Group LASSO linear regression problem of Section 3.2.2. Our experiments consider different scenarios where we fix all of the problem parameters but one and generate random data following a group-sparse linear regression model,

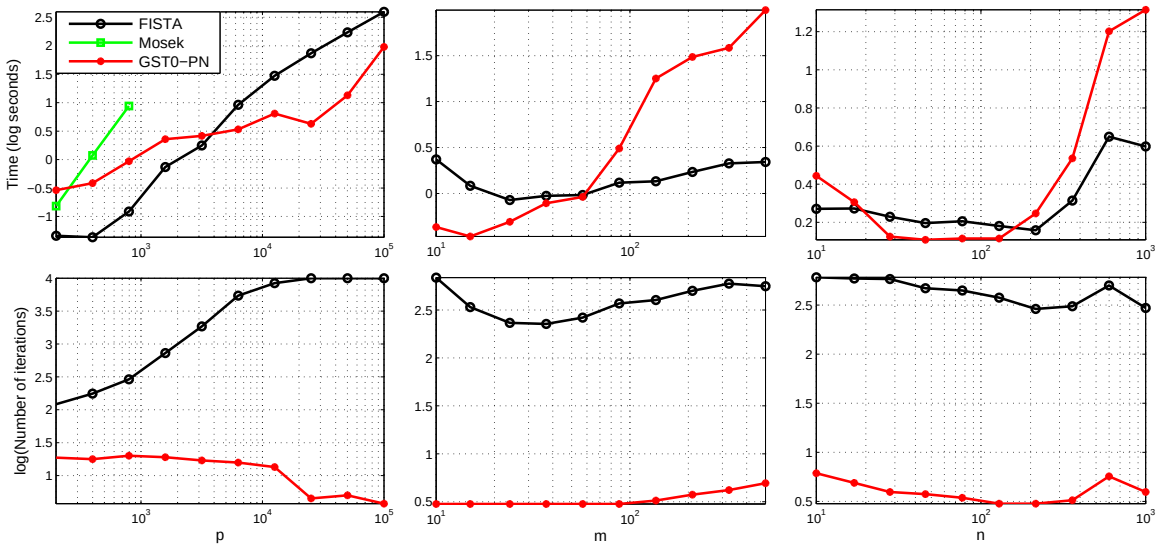
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{n}, \quad \boldsymbol{\theta} \text{ is group-sparse},$$

with $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and m overlapping groups. These groups are randomly generated with average overlap of 50% and the ratio of active groups is equal to 1%. As a measure of performance, we consider the elapsed time and the number of iterations needed for the algorithm to achieve a given target objective value. We compare our GSTO algorithms of Section 3.3 to a state-of-the-art first-order algorithm (LY10) based on the Fast Iterative Thresholding Algorithm (FISTA) paradigm of (BT09), and to the commercial interior-point solver Mosek®. The comparison to the Projection Subgradient method is omitted because our simulations show that this algorithm is not competitive in this setting as it fails to achieve the desired accuracy in reasonable time.

We consider three experiments, shown in Figure 3.2: (a) varying number of variables p with the number of groups m and the number of samples n fixed and equal to 100, (b) varying number of groups m with fixed $p = 10^4$ and $n = 100$ and (c) varying number of samples with fixed $p = 10^4$ and $m = 100$. Since the computation time for each algorithms depends on the sparsity of the solution, which is in turn controlled by the parameter λ, we set this parameter to a tenth of the parameter which yields all-zero solutions. Our results show that our algorithm outperforms the other two for moderately small m and n and potentially very large p. This is expected since the computation of the Hessian matrix required to implement the Projected Newton Step requires the solution of m systems of equations of size n and the evaluation of a number $m(m-1)/2$ of p-dimensional vector multiplications and additions.

3.4.2 Computation of the regularization path

We consider here the application of the GSTO Projected Newton algorithm combined with the continuation update of Section 3.3.3, to the problem of computing the regularization path for the overlapping Group LASSO linear regression problem of Section 3.2.2. We generate random data following the group-sparse linear regression model of the previous section, with varying p, m and n. For each realization, we com-



(a) $m = 100$, $n = 100$ (b) $p = 10^4$, $n = 100$ (c) $p = 10^4$, $m = 100$

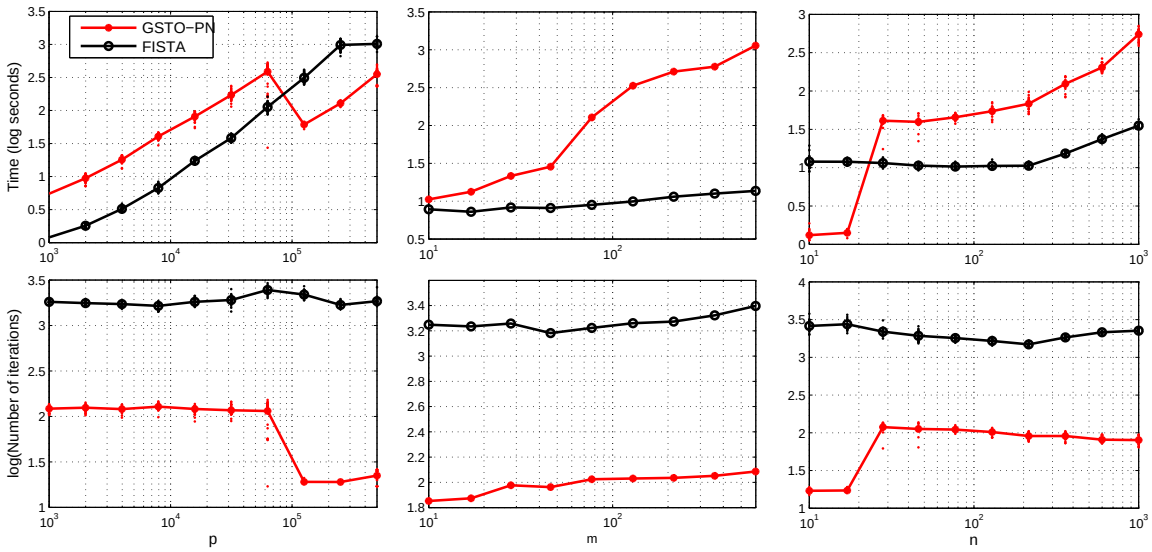
Figure 3.2: Comparison of GSTO and FISTA (LY10; BT09) elapsed times for solving (3.30) as a function of p (a), m (b) and n (c). GSTO is significantly faster than FISTA for p larger than 4000 and small m.

pute the regularization path over a 20 point grid of regularization parameters. The maximum value in the grid is chosen to be such that it yields the all-zero solution, and the minimum value is set to a millionth fraction of the maximum value.

We compare our method to the fast overlapping algorithm of (LY10), denoted as FISTA, with a warm start as an initialization procedure. We compute the elapsed times as a function of p, m and n with the other two parameters fixed, with the exact same choices as in the previous section. The results, shown in Figure 3.3, demonstrate again the competitiveness of our algorithm in the large p, small m and small n regime, corroborating the results obtained in the previous section for computing a single point in the regularization path.

3.4.3 Application to multi-task regression

In this section we consider the application of the GSTO to the multi-task learning problem of predicting symptom scores from the gene expression responses of symptomatic individuals. In our context, the multi-task learning model is defined as follows. For each individual $s = 1, \cdots, S$, we measure a collection of gene expression vectors obtained at T different time points, to form a $T \times p$ matrix of gene expression values denoted by $\boldsymbol{X}_s$. The symptom score of each individual, denoted by $\boldsymbol{y}_s$, is a T-dimensional vector containing a symptom severity index for each time point. We



(a) $m = 100$, $n = 100$ (b) $p = 10^4$, $n = 100$ (c) $p = 10^4$, $n = 100$

Figure 3.3: Comparison of GSTO and FISTA (LY10; BT09) elapsed times for computing the regularization path of (3.30) as a function of p (a), m (b) and n (c). GSTO is significantly faster than FISTA for p larger than 10^5 and small m, n.

assume that the symptom scores can be well modeled by a sparse linear combination of the gene expression responses, that is:

$$\boldsymbol{y}_s = \boldsymbol{X}_s \boldsymbol{\theta}_s + \boldsymbol{n}_s \tag{3.57}$$

where $\boldsymbol{\theta}_s$ are the linear predictors for subject s and $\boldsymbol{n}_s$ is the residual noise. All subjects are humans inoculated by the same virus, therefore, a similar immune system response is to be expected. To exploit this a priori information, we enforce that the supports of the predictors $\boldsymbol{\theta}_s$ be the same:

$$\operatorname{supp}(\boldsymbol{\theta}_k) = \operatorname{supp}(\boldsymbol{\theta}_l), k \neq l, \tag{3.58}$$

whereas the specific contribution of each gene to the individual symptom prediction is not necessarily equal. The estimates $\{\boldsymbol{\theta}_s\}_{s=1}^S$ are then obtained as the solution to the following Group-ℓ_2 penalized least squares problem:

$$\min_{\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_S} \quad \tfrac{1}{2}\sum_{s=1}^S ||\boldsymbol{X}_s\boldsymbol{\theta}_s - \boldsymbol{y}_s||_2^2 + \lambda \sum_{i=1}^m \sqrt{c_i} \left|\left| \boldsymbol{A}_{G_i,*} \left[\boldsymbol{\theta}_1^T, \cdots, \boldsymbol{\theta}_S^T\right]^T \right|\right|_2 \tag{3.59}$$

In this section we consider three choices for the matrices $\boldsymbol{A}_{G_i,*}$ and the groups G_i in the penalty above:

1. *LASSO*: The LASSO penalty corresponds to $G_i = i$, $i = 1, \cdots, pS$ and $\boldsymbol{A} = \boldsymbol{I}_p$. This penalty yields sparse $\boldsymbol{\theta}_s$ but does not enforce the support constraint (3.58), but we consider it here to illustrate the benefit of structured-sparse penalties such as the following two.

2. *Gene-wise Group-LASSO*: The gene-wise group-LASSO penalty enforces the support constraint by choosing $G_i = \{i, i+p, \cdots, i+p(S-1)\}$, $i = 1, \cdots, p$. In this case the group matrix is chosen to be $\boldsymbol{A}_{G_i,G_i} = \boldsymbol{I}_{|G_i|}$, which corresponds to the non-overlapping group-lasso case of Corollary III.3.

3. *Pathway-wise Group-LASSO*: The Pathway-wise group-LASSO penalty enforces the support constraint and, in addition, incorporates prior information on groups of genes that are known to be co-expressed. Given a set of m pathways (groups of genes), $P_i \subset \{1, \cdots, p\}$, $i = 1, \cdots, m$, we choose $G_i = \{P_i, P_i + p, \cdots, P_i + p(S-1)\}$, $i = 1, \cdots, m$ and, again, $\boldsymbol{A}_{G_i,G_i} = \boldsymbol{I}_{|G_i|}$. This corresponds to the overlapping group-lasso penalized linear regression of Corollary III.4.

It is worthwhile to notice that this multi-task model and estimate have in addition the advantage of begin invariant to possible subject-dependent shifts in the temporal axis: the objective in (3.59) is invariant to left-hand multiplications of $\boldsymbol{X}_s$ and $\boldsymbol{y}_s$ by a circular shift matrix.

Our experiments are set as follows. We obtain a set of 831 curated gene pathways from Broad Institutes MSigDB database[1] and restrict our interest to the 5115 genes from our microarray assay that appear at least in one of these pathways. (Future studies including other sets of pathways and/or more genes will be the subject of future research.)

For each of the H3N2-infected subjects, we restrict our attention to those 9 declared as symptomatic, and construct a 16×5115 design matrix $\boldsymbol{X}_s$ consisting of 16-time points of gene expression levels corresponding to each of the 5115 genes in the MSigDB pathways. We also compile the symptoms (Runny Nose, Stuffy Nose, Earache, Sneezing, SoreThroat, Cough, Headache, Muscle/Joint Pain) declared by each individual at 10 different time points into a 10-time point aggregated symptom score obtained as the sum of each symptom score at each time point. Since the times at which the symptoms where assessed and the times at which the blood samples for

[1] `http://www.broadinstitute.org/gsea/downloads.jsp`

Method	Train MSE	Test MSE	Test MSE (LS)	# Genes	# Pathways
LASSO	0.07	0.03	0.025	14	-
Gene-wise GL	0.06	0.012	0.014	40	-
Pathway-wise GL	0.05	0.018	0.018	83	8
Least Squares	5×10^{-11}	0.015	0.015	5125	-

Table 3.1: Comparison of symptom prediction performance for each of the Multi-task regression methods of Section 3.4.3.

gene expression assay were drawn are not exatcly the same, we interpolate the former to obtain a symptom score for each individual in sync to the gene expression time points. The interpolated (dashed) and the original (solid lines) aggregated symptom scores are shown in Figure 3.4.

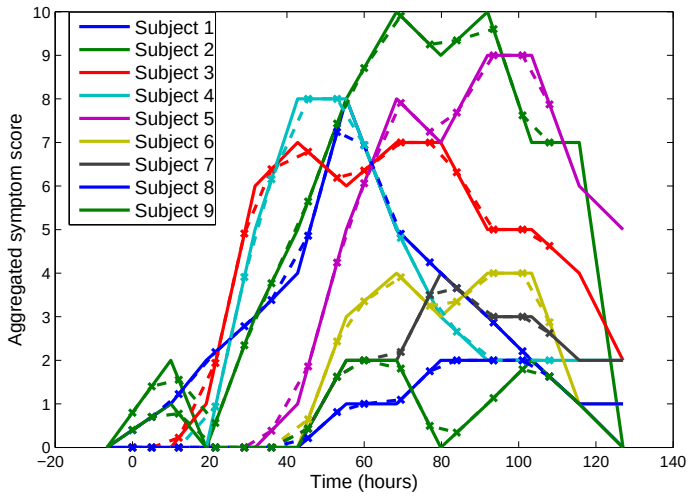


Figure 3.4: Interpolated (dashed) and original (solid lines) aggregated symptom scores for each of the H3N2 infected symptomatic individuals.

For each individual, we divide the data into 10 time points for training and 6 for testing each algorithm's performance. To choose the tuning parameter λ that controls the sparsity of the solutions, we perform 10-fold Cross Validation on the training data, and choose the λ that minimizes the prediction error, measured through the relative MSE. Since it is well-known that penalized estimators are biased for large λ, we also compute the Least Squares estimator restricted to the support obtained through (3.59). We also compare our results to the traditional Least Squares estimate, given by solving (3.59) with $\lambda = 0$, which is not sparse and ill-conditioned due to the small sample size. The results of our experiment are shown in Table 3.1.

For the gene-wise and pathway-wise group lasso estimates, we give in Tables 3.2 and 3.3 the top-10 genes selected to construct the predictor, ordered by their p-values, which are computed through an ANOVA procedure on the 11961 genes of our normalized dataset. The first table reflects the association of the genes in the gene-wise predictor with pathways linked to the immune system and the inflamatory response. Table 3.3, in contrast, shows a less obvious correspondance between some of the active pathways and known inflamatory-specific pathways. Some of the active pathways, such as GLYCAN BIOSYNTHESIS or SNARE INTERACTIONS IN VESICULAR TRANSPORT perform fundamental functions of cell homeostasis that may or may not be associated to the specific response to a viral infection. On the other hand, this might be a limitation of the pathway-wise sparsity penalty: if a certain gene is highly predictive of the symptomatic response, selection of this gene might entail the selection of one of the pathways to which it belong, since the support of the predictor has to be equal to the complement of the inactive groups, as we showed in Corollary III.4.

Finally, we show in Figure 3.5 the true and predicted aggregated symptom scores for each individual, for the optimal pathway-wise group lasso predictor, and a heatmap with the gene expression values associated to the active genes. It is clear that our algorithm is able to construct the predictor from genes whose response correlates well with the true symptom temporal course.

3.5 Conclusions and future work

We have introduced the GSTO, which extends the Scalar and the Multidimensional Shrinkage Thresholding Operators to penalties composed of sums of ℓ_2 norms of linear combinations of the optimization variables. Our theoretical results give insight to the behavior of a general class of structured-sparse penalties that includes, but is not restricted, to the well known LASSO (Tib96), Group LASSO (YL06b) and structured-sparsity Overlapping Group LASSO penalties (ZRY09; JAB09). In addition, we have shown that the GSTO can be reformulated into a lower-dimensional problem, and that an infinitesimaly close perturbation of this problem is smooth. This allows for second-order methods to approximately solve the original low-dimensional non-smooth problem, and we numerically demonstrate that this approach is efficient in certain regimes. We finally have demonstrated the applicability of the GSTO in a high-dimensional multi-task learning problem involving the prediction of symptoms from the gene expression levels of infected individuals. Future work includes the effi-

Gene Symbol	**Gene Name**	**Function** (Wikipedia & Genecards)
SIGLEC1	Sialic acid binding Ig-like lectin 1	Macrophage marker
SERPING1	Serpin peptidase inhibitor, clade G	Inhibition of the complement system (part of the immune system)
LAMP3	Lysosome-associated membrane glycoprotein 3	
ISG15	Interferon-induced 17 kDa	Its activity is regulated by specific signaling pathways that have a role in innate immunity
C1QB	Complement component 1, q subcomponent, B	Encodes a major constituent of the human complement system
ATF3	Cyclic AMP-dependent transcription factor ATF-3	
C1QA	Complement C1q subcomponent subunit A	Encodes a major constituent of the human complement system
CXCL10	C-X-C motif chemokine 10	Chemoattraction for monocytes/macrophages, T cells, NK cells, and dendritic cells.
LAP3	Leucine aminopeptidase 3	
AIM2	Absent in melanoma 2	Contributes to the defence against bacterial and viral DNA.

Table 3.2: Top-10 genes in the support of the Gene-wise Group-LASSO multi-task predictor, ordered by ANOVA p-value.

cient evaluation of the proximal algorithm associated to general Group-ℓ_2 penalties, which would allow the application of this rich class of penalties to general differentiable cost functions with Lipschitz gradient.

Pathway	**Ratio Active Genes**	**Active Genes**	**Function** (Kegg Database)
KEGG RIG I LIKE RECEPTOR SIGNALING PATHWAY	0.16	ISG15, IFIH1, NLRX1, AZI2, RNF125, CYLD, DDX3X, DDX3Y, TBKBP1, DHX58, TRIM25, DDX58	RIG-I-like receptors recruit specific intracellular adaptor proteins to initiate signaling pathways that lead to the synthesis of type I interferon and other inflammatory cytokines, which are important for eliminating viruses.
KEGG CYTOSOLIC DNA SENSING PATHWAY	0.07	IL33, ZBP1, AIM2, DDX58'	Responsible for detecting foreign DNA from invading microbes or host cells and generating innate immune responses.
KEGG O GLYCAN BIOSYNTHESIS	0.53	GALNT3, GCNT4, GALNT2, GALNT1, GCNT3, GALNT7, GALNT6, GALNT4, GALNT10, GALNT11, GCNT1, C1GALT1 B4GALT5, GALNT14, C1GALT1C1, GALNT8	
KEGG GLYCOSAMINOGLYCAN DEGRADATION	0.38	HYAL2, HYAL3, HS3ST3A1, HYAL4, HPSE2, HPSE, SPAM1, HS3ST3B1	
KEGG GLYCOSAMINOGLYCAN BIOSYNTHESIS HEPARAN SULFATE	0.57	EXTL3, NDST3, NDST4, NDST1, NDST2, HS3ST2, HS3ST1,EXTL1, EXTL2, HS2ST1, HS3ST3A1, GLCE, EXT1, EXT2, HS3ST3B1	
KEGG FOLATE BIOSYNTHESIS	0.72	ALPL, ALPPL2, ALPI, GGH, SPR, PTS, ALPP, GCH1	
KEGG SNARE INTERACTIONS IN VESICULAR TRANSPORT	0.39	SNAP29, BET1, USE1, STX18, STX17, SEC22B, VTI1B, BNIP1, VAMP5, VAMP4, BET1L, VAMP3, GOSR2, GOSR1, YKT6	
KEGG CIRCADIAN RHYTHM MAMMAL	0.6	ARNTL, NPAS2, CRY2, PER2, PER1, PER3, CRY1, CLOCK	

Table 3.3: Active pathways and their active genes (ordered by ANOVA p-value) in the support of the Pathway-wise Group-LASSO multi-task predictor. Highlighted in red are the genes that also appeared in the Gene-wise Group LASSO predictor.

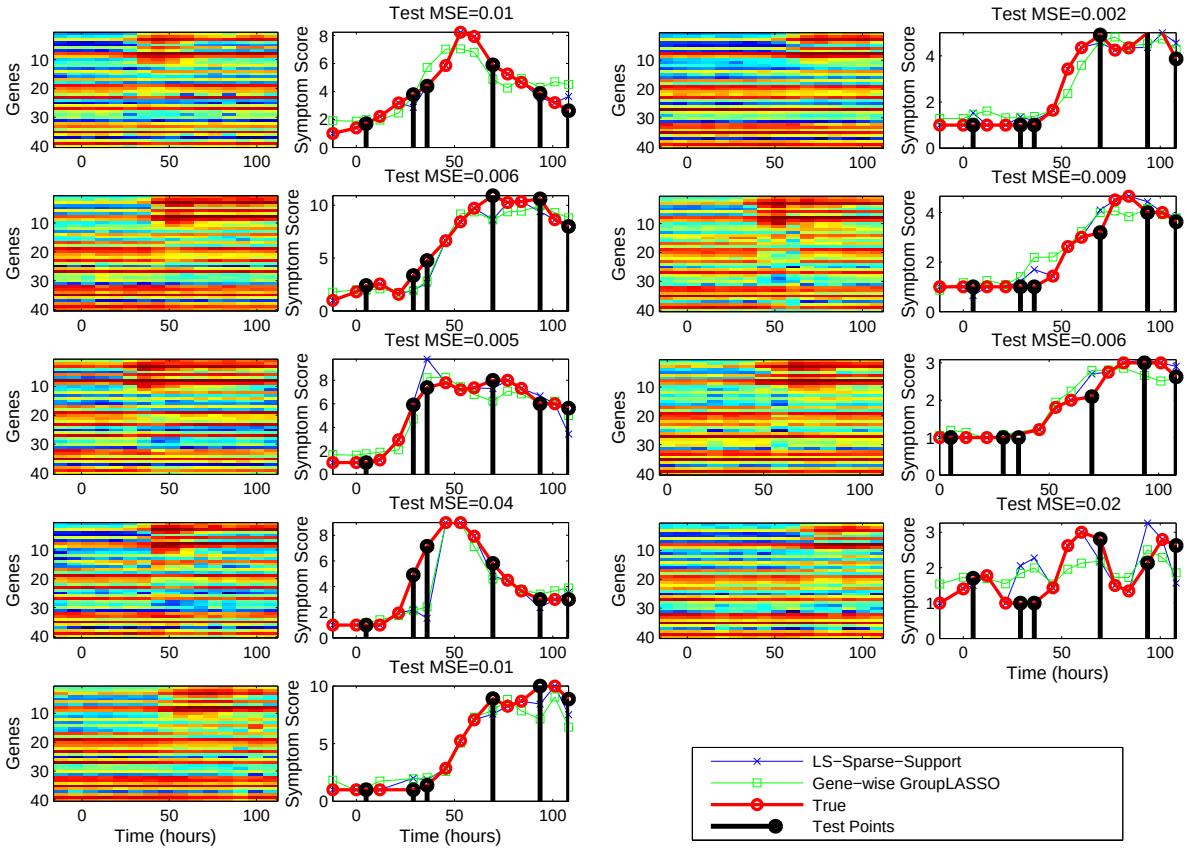


Figure 3.5: Left columns: Heatmap of the gene expression values associated to the active genes used in the predictor. Right columns: True and predicted aggregated symptom scores for each individual. The predictors considered here are (i) the Gene-wise Group LASSO multi-task estimate (labeled as "Gene-wise GroupLASSO") and (ii) the Least Squares predictor restricted to the support of the Gene-wise Group LASSO multi-task estimate (labeled as "LS-Sparse-Support"). The average relative MSE over the 9 subjects is 0.012.

CHAPTER IV

Order-Preserving Factor Analysis

4.1 Introduction

With the advent of high-throughput data collection techniques, low-dimensional matrix factorizations have become an essential tool for pre-processing, interpreting or compressing high-dimensional data. They are widely used in a variety of signal processing domains including electrocardiogram (JL09), image (JOB10), or sound (BD06) processing. These methods can take advantage of a large range of a priori knowledge on the form of the factors, enforcing it through constraints on sparsity or patterns in the factors. However, these methods do not work well when there are unknown misalignments between subjects in the population, e.g., unknown subject-specific time shifts. In such cases, one cannot apply standard patterning constraints without first aligning the data; a difficult task. An alternative approach, explored in this chapter, is to impose a factorization constraint that is invariant to factor misalignments but preserves the relative ordering of the factors over the population. This order-preserving factor analysis is accomplished using a penalized least squares formulation using shift-invariant yet order-preserving model selection (group lasso) penalties on the factorization. As a byproduct the factorization produces estimates of the factor ordering and the order-preserving time shifts.

In traditional matrix factorization, the data is modeled as a linear combination of a number of factors. Thus, given an $n \times p$ data matrix $\boldsymbol{X}$, the Linear Factor model is defined as:

$$\boldsymbol{X} = \boldsymbol{M}\boldsymbol{A} + \boldsymbol{\epsilon}, \tag{4.1}$$

where $\boldsymbol{M}$ is a $n \times f$ matrix of factor loadings or dictionary elements, $\boldsymbol{A}$ is a $f \times p$ matrix of scores (also called coordinates) and ϵ is a small residual. For example, in a gene

expression time course analysis, n is the number of time points and p is the number of genes in the study, the columns of $\boldsymbol{M}$ contain the features summarizing the genes' temporal trajectories and the columns of $\boldsymbol{A}$ represent the coordinates of each gene on the space spanned by $\boldsymbol{M}$. Given this model, the problem is to find a parsimonious factorization that fits the data well according to selected criteria, e.g. minimizing the reconstruction error or maximizing the explained variance. There are two main approaches to such a parsimonious factorization. One, called Factor Analysis, assumes that the number of factors is small and yields a low-rank matrix factorization (Pea01), (CC70). The other, called Dictionary Learning (AEB06), (KDMR$^+$03) or Sparse Coding (OF97), assumes that the loading matrix $\boldsymbol{M}$ comes from an overcomplete dictionary of functions and results in a sparse score matrix $\boldsymbol{A}$. There are also hybrid approaches such as Sparse Factor Analysis (JL09), (WTH09), (JOB10) that try to enforce low rank and sparsity simultaneously.

In many situations, we observe not one but several matrices $\boldsymbol{X}_s$, $s = 1, \cdots, S$ and there are physical grounds for believing that the $\boldsymbol{X}_s$'s share an underlying model. This happens, for instance, when the observations consist of different time-blocks of sound from the same music piece (BD06), (MLG$^+$08), when they consist of time samples of gene expression microarray data from different individuals inoculated with the same virus (ZCV$^+$09), or when they arise from the reception of digital data with code, spatial and temporal diversity (SGB00). In these situations, the fixed factor model (4.1) is overly simplistic.

An example, which is the main motivation for this work is shown in Figure 4.1, which shows the effect of temporal misalignment across subjects in a viral challenge study reported in (ZCV$^+$09). Figure 4.1 shows the expression trajectory for a particular gene that undergoes an increase (up-regulation) after viral inoculation at time 0, where the moment when up-regulation occurs differs over the population. Training the model (4.1) on this data will produce poor fit due to misalignment of gene expression onset times.

A more sensible approach for the data in Figure 4.1 would be to separately fit each subject with a translated version of a common up-regulation factor. This motivates the following extension of model (4.1), where the factor matrices $\boldsymbol{M}_s$, $\boldsymbol{A}_s$ are allowed to vary across observations. Given a number S of $n \times p$ data matrices $\boldsymbol{X}_s$, we let:

$$\boldsymbol{X}_s = \boldsymbol{M}_s \boldsymbol{A}_s + \boldsymbol{\epsilon}_s \quad s = 1, \cdots, S. \tag{4.2}$$

Following the gene expression example, here n is the number of time points, p is

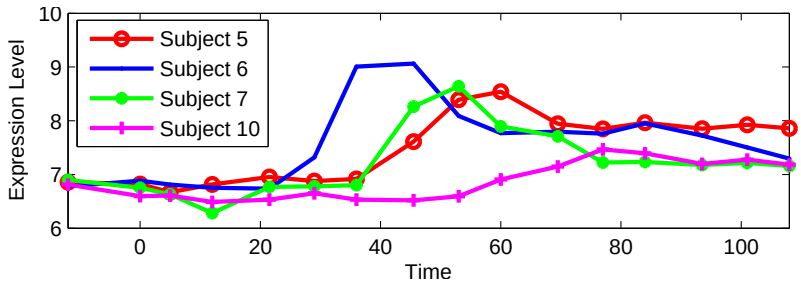


Figure 4.1: Example of temporal misalignment across subjects of upregulated gene *CCRL2*. Subject 6 and subject 10 show the earliest and the latest up-regulation responses, respectively.

the number of genes in the study, and S is the number of subjects participating in the study. Hence, the $n \times f$ matrices $\boldsymbol{M}_s$ contain the translated temporal features corresponding to the s-th subject and the $f \times p$ matrices $\boldsymbol{A}_s$ accommodate the possibility of subjects having different mixing weights. For different constraints on $\boldsymbol{M}_s$, $\boldsymbol{A}_s$, this model specializes to several well-known paradigms such as Principal Components Analysis (PCA) (Pea01), sparse PCA (JL09), k-SVD (AEB06), structured PCA (JOB10), Non-Negative Matrix Factorization (NNMF) (LS99a), Maximum-Margin Matrix Factorization (MMMF) (SRJ05), Sparse Shift-invariant models (BD06), Parallel Factor Analysis (PARAFAC) (CC70), (KB09) or Higher-Order SVD (HOSVD) (BL10). Table 4.1 summarizes the characteristics of these decomposition models when seen as different instances of the general model (4.2).

Model	**Structure of $\boldsymbol{M}_s$**	**Structure of $\boldsymbol{A}_s$**	**Reference**
PCA	Orthogonal $\boldsymbol{M}_s = \boldsymbol{F}$	Orthogonal $\boldsymbol{A}_s$	SVD
Sparse-PCA	Sparse $\boldsymbol{M}_s = \boldsymbol{F}$	Sparse $\boldsymbol{A}_s$	Sparse PCA (JL09), (dEGJL07), k-SVD (AEB06), PMD (WTH09)
Structured-PCA	$\boldsymbol{M}_s = \boldsymbol{F}$	Structured Sparse $\boldsymbol{A}_s$	(JOB10)
NNMF	Non-negative $\boldsymbol{M}_s = \boldsymbol{F}$	Non-negative $\boldsymbol{A}_s$	(LS99a)
Sparse Shift-invariant models	$\boldsymbol{M}_s = [\boldsymbol{M}(\boldsymbol{F}, d_1) \cdots \boldsymbol{M}(\boldsymbol{F}, d_D)]$ where $\{d_j\}_{j=1}^{D}$ are all possible translations of the n-dimensional vectors in $\boldsymbol{F}$.	Sparse $\boldsymbol{A}_s$	(LS99b), (BD06), (MLG$^+$08)
PARAFAC/CP	$\boldsymbol{M}_s = \boldsymbol{F}$	$\boldsymbol{A}_s = \text{diag}(\boldsymbol{C}_{\cdot,s})\, \boldsymbol{B}'$	(KB09)
HOSVD	Orthogonal $\boldsymbol{M}_s = \boldsymbol{F}$	$\boldsymbol{A}_s = (\mathcal{G} \times_3 \boldsymbol{C}_{\cdot,s})\, \boldsymbol{B}'$ where slices of $\mathcal{G}$ are orthogonal	(BL10)
OPFA	$\boldsymbol{M}_s = \boldsymbol{M}(\boldsymbol{F}, \boldsymbol{d}^s)$, $\boldsymbol{d}^s \in \mathcal{K}$ where $\boldsymbol{F}$ is smooth and non-negative and $\mathcal{K}$ enforces consistent precedence order	Non-negative, sparse $\boldsymbol{A}_s$	This work.

Table 4.1: Special cases of the general model (4.2).

In this chapter, we will restrict the columns of $\boldsymbol{M}_s$ to be translated versions of a common set of factors, where these factors have onsets that occur in some relative

order that is consistent across all subjects. Our model differs from previous shift-invariant models considered in (LS99b), (BD06), (MLG$^+$08) in that it restricts the possible shifts to those which preserve the relative order of the factors among different subjects. We call the problem of finding a decomposition (4.2) under this assumption the Order Preserving Factor Analysis (OPFA) problem.

The contributions of this chapter are the following. First, we propose a non-negatively constrained linear model that accounts for temporally misaligned factors and order restrictions. Second, we give a computational algorithm that allows us to fit this model in reasonable time. Finally, we demonstrate that our methodology is able to succesfully extract the principal features in a simulated dataset and in a real gene expression dataset. In addition, we show that the application of OPFA produces factors that can be used to significantly reduce the variability in clustering of gene expression responses.

This chapter is organized as follows. In Section 4.2 we present the biological problem that motivates OPFA and introduce our mathematical model. In Section 4.3, we formulate the non-convex optimization problem associated with the fitting of our model and give a simple local optimization algorithm. In Section 4.4 we apply our methodology to both synthetic data and real gene expression data. Finally we conclude in Section 4.5.

4.2 Motivation: gene expression time-course data

In this section we motivate the OPFA mathematical model in the context of gene expression time-course analysis. Temporal profiles of gene expression often exhibit motifs that correspond to cascades of up-regulation/down-regulation patterns. For example, in a study of a person's host immune response after inoculation with a certain pathogen, one would expect genes related to immune response to exhibit consistent patterns of activation across pathogens, persons, and environmental conditions.

A simple approach to characterize the response patterns is to encode them as sequences of a few basic motifs such as (see, for instance, (SLMB07)):

- *Up-regulation*: Gene expression changes from low to high.
- *Down-regulation*: Gene expression changes from a high to a low level.
- *Steady*: Gene expression does not vary.

If gene expression is coherent over the population of several individuals, e.g., in response to a common viral insult, the response patterns can be expected to show some

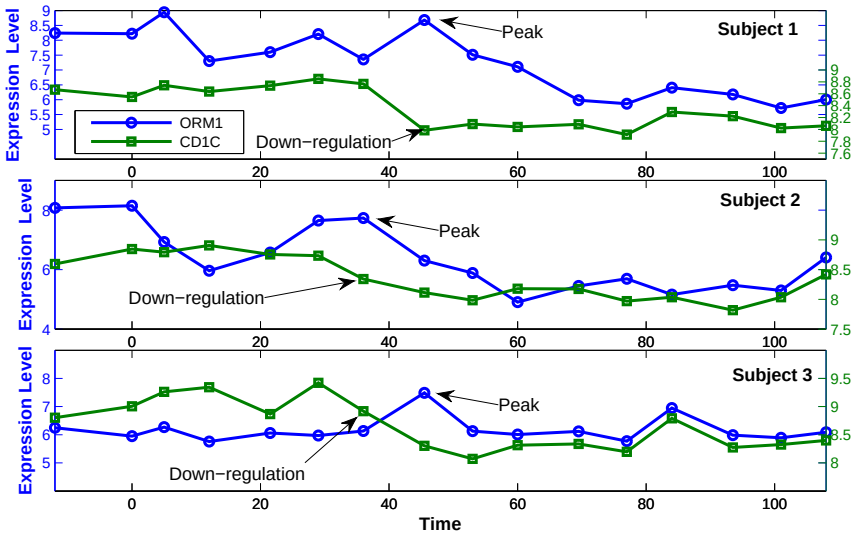


Figure 4.2: Example of gene patterns with a consistent precedence-order across 3 subjects. The down-regulation motif of gene *CD1C* precedes the peak motif of gene *ORM1* across these three subjects.

degree of consistency across subjects. Human immune system response is a highly evolved system in which several biological pathways are recruited and organized over time. Some of these pathways will be composed of genes whose expressions obey a precedence-ordering, e.g., virally induced ribosomal protein production may precede toll-like receptor activation and antigen presentation (AU00). This consistency exists despite temporal misalignment: even though the order is preserved, the specific timing of these events can vary across the individuals. For instance, two different persons can have different inflammatory response times, perhaps due to a slower immune system in one of the subjects. This precedence-ordering of motifs in the sequence of immune system response events is invariant to time shifts that preserve the ordering. Thus if a motif in one gene precedes another motif in another gene for a few subjects, we might expect the same precedence relationship to hold for all other subjects. Figure 4.2 shows two genes from (ZCV$^+$09) whose motif precedence-order is conserved across 3 different subjects. This conservation of order allows one to impose ordering constraints on (4.2) without actually knowing the particular order or the particular factors that obey the order-preserving property.

Often genes are co-regulated or co-expressed and have highly correlated expression profiles. This can happen, for example, when the genes belong to the same signaling pathway. Figure 4.3 shows a set of different genes that exhibit a similar expression pattern (up-regulation motif). The existence of high correlation between large groups

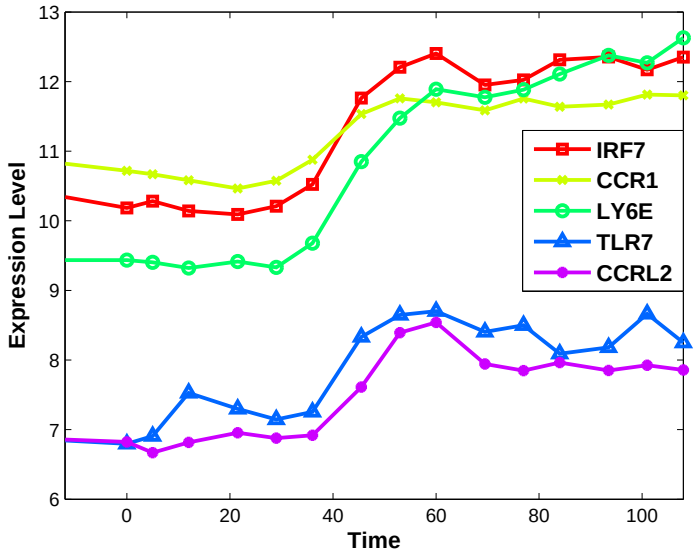


Figure 4.3: Example of gene patterns exhibiting co-expression for a particular subject in the viral challenge study in (ZCV$^+$09).

of genes allows one to impose a low rank property on the factorization in (4.2).

In summary, our OPFA model is based on the following assumptions:

- *A1: Motif consistency across subjects*: Gene expression patterns have consistent (though not-necessarily time aligned) motifs across subjects undergoing a similar treatment.

- *A2: Motif sequence consistency across subjects*: If motif X precedes motif Y for subject s, the same precedence must hold for subject $t \neq s$.

- *A3: Motif consistency across groups of genes*: There are (not necessarily known) groups of genes that exhibit the same temporal expression patterns for a given subject.

- *A4: Gene Expression data is non-negative*: Gene expression on a microarray is measured as an abundance and standard normalization procedures, such as RMA (IHC$^+$03), preserve the non-negativity of this measurement.

A few microarray normalization software packages produce gene expression scores that do not satisfy the non-negativity assumption A4. In such cases, the non-negativity constraint in the algorithm implementing (4.9) can be disabled. Note that in general, only a subset of genes may satisfy assumptions *A1-A3*.

4.3 OPFA mathematical model

In the OPFA model, each of the S observations is represented by a linear combination of *temporally aligned* factors. Each observation is of dimension $n \times p$, where n is the number of time points and p is the number of genes under consideration. Let $\boldsymbol{F}$ be an $n \times f$ matrix whose columns are the f common *alignable* factors, and let $\boldsymbol{M}(\boldsymbol{F}, \boldsymbol{d})$ be a matrix valued function that applies a circular shift to each column of $\boldsymbol{F}$ according to the vector of shift parameters $\boldsymbol{d}$, as depicted in Figure 4.4. Then, we can refine model (4.2) by restricting $\mathbf{M}_s$ to have the form:

$$\boldsymbol{M}_s = \boldsymbol{M}(\boldsymbol{F}, \boldsymbol{d}^s). \tag{4.3}$$

where $\boldsymbol{d}^s \in \{0, \cdots, d_{\max}\}^f$ and $d_{\max} \leq n$ is the maximum shift allowed in our model. This model is a generalization of a simpler one that restricts all factors to be aligned but with a common delay:

$$\boldsymbol{M}_s = \boldsymbol{U}_s \boldsymbol{F}, \tag{4.4}$$

where $\boldsymbol{U}_s$ is a circular shift operator. Specifically, the fundamental characteristic of our model (4.3) is that each column can have a different delay, whereas (4.4) is a restriction of (4.3) with $d_i^s = d_j^s$ for all s and all i, j.

The circular shift is not restrictive. By embedding the observation into a larger time window it can accommodate transient gene expression profiles in addition to periodic ones, e.g., circadian rhythms (TPWH10). There are several ways to do this embedding. One way is to simply extrapolate the windowed, transient data to a larger number of time points $n_F = n + d_{\max}$. This is the strategy we follow in the numerical experiments of Section IV-B.

This alignable factor model parameterizes each observation's intrinsic temporal dynamics through the f-dimensional vector $\boldsymbol{d}^s$. The precedence-ordering constraint *A2* is enforced by imposing the condition

$$d_{j_1}^{s_1} \leq d_{j_2}^{s_1} \Leftrightarrow d_{j_1}^{s_2} \leq d_{j_2}^{s_2} \; \forall s_2 \neq s_1, \tag{4.5}$$

that is, if factor j_1 precedes factor j_2 in subject s_1, then the same ordering will hold in all other subjects. Since the indexing of the factors is arbitrary, we can assume without loss of generality that $d_i^s \leq d_{i+1}^s$ for all i and all s. This characterization constrains each observation's delays $\boldsymbol{d}^s$ independently, allowing for a computationally

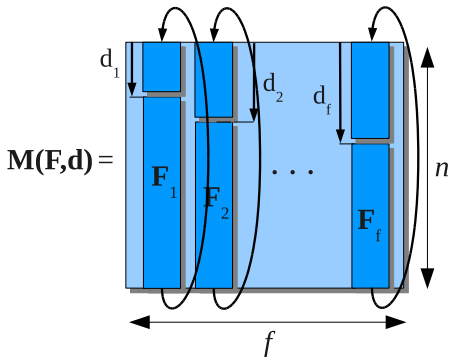


Figure 4.4: Each subject's factor matrix $\boldsymbol{M}_s$ is obtained by applying a circular shift to a common set of factors $\boldsymbol{F}$ parameterized by a vector $\boldsymbol{d}$.

efficient algorithm for fitting model (4.3).

4.3.1 Relationship to 3-way factor models.

Our proposed OPFA framework is significantly different from other factor analysis methods and these differences are illustrated in the simulated performance comparisons below. However, there are some similarities, especially to 3-way factor models (KB09), (Com02) that are worth pointing out.

An n-th order tensor or n-way array is a data structure whose elements are indexed by an n-tuple of indices (Com02). n-way arrays can be seen as multidimensional generalizations of vectors and matrices: an 1-way array is a vector and a 2-way array is a matrix. Thus, we can view our observations $\boldsymbol{X}_s$ as the slices of a third order tensor $\mathcal{X}$ of dimension $p \times n \times S$: $\boldsymbol{X}_s = \mathcal{X}_{\cdot,\cdot,i}$. Tensor decompositions aim at extending the ideas of matrix (second order arrays) factorizations to higher order arrays (KB09), (Com02) and have found many applications in signal processing and elsewhere (Com02), (KB09), (BL10), (SGB00), (DLDMV00). Since our data tensor is of order 3, we will only consider here 3-way decompositions, which typically take the following general form:

$$\mathcal{X}_{i,j,k} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} \mathcal{G}_{pqr} \boldsymbol{F}_{ip} \boldsymbol{B}_{jq} \boldsymbol{C}_{kr} \tag{4.6}$$

where P, Q, R are the number of columns in each of the factor matrices $\boldsymbol{F}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\mathcal{G}$ is a $P \times Q \times R$ tensor. This class of decompositions is known as the Tucker model. When orthogonality is enforced among $\boldsymbol{F}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and different matrix slices

of $\mathcal{G}$, one obtains the Higher Order SVD (BL10). When $\mathcal{G}$ is a superdiagonal tensor[1] and $P = Q = R$, this model amounts to the PARAFAC/Canonical Decomposition (CP) model (CC70), (SGB00), (DLDMV00). The PARAFAC model is the closest to OPFA. Under this model, the slices of $\mathcal{X}_{i,j,k}$ can be written as:

$$\boldsymbol{X}_s^{\mathrm{CP}} = \boldsymbol{F}\mathrm{diag}\left(\boldsymbol{C}_{\cdot,s}\right)\boldsymbol{B}'. \tag{4.7}$$

This expression is to be compared with our OPFA model, which we state again here for convenience:

$$\boldsymbol{X}_s^{\mathrm{OPFA}} = \boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\boldsymbol{A}_s. \tag{4.8}$$

Essentially, (4.7) shows that the PARAFAC decomposition is a special case of the OPFA model (4.8) where the factors are fixed ($\boldsymbol{M}_s = \boldsymbol{F}$) and the scores only vary in magnitude across observations ($\boldsymbol{A}_s = \mathrm{diag}\left(\boldsymbol{C}_{\cdot,s}\right)\boldsymbol{B}'$). This structure enhances uniqueness (under some conditions concerning the linear independence of the vectors in $\boldsymbol{F}$, $\boldsymbol{B}$, $\boldsymbol{C}$, see (KB09)) but lacks the additional flexibility necessary to model possible translations in the columns of the factor matrix $\boldsymbol{F}$. If $\boldsymbol{d}^s = \boldsymbol{0}$ for all s, then the OPFA (4.8) and the Linear Factor model (4.1) also coincide. The OPFA model can be therefore seen as an extension of the Linear Factor and PARAFAC models where the factors are allowed to experiment order-preserving circular translations across different individuals.

4.3.2 OPFA as an optimization problem

OPFA tries to fit the model (4.2)-(4.5) to the data $\{\boldsymbol{X}_s\}_{s=1}^S$. For this purpose, we define the following penalized and constrained least squares problem:

$$\begin{array}{ll} \min & \sum_{s=1}^S \left|\left|\boldsymbol{X}_s - \boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\boldsymbol{A}_s\right|\right|_F^2 + \lambda P_1\left(\boldsymbol{A}_1, \cdots, \boldsymbol{A}_S\right) + \beta P_2\left(\boldsymbol{F}\right) \\ \text{s.t.} & \left\{\boldsymbol{d}^s\right\}_s \in \mathcal{K}, \boldsymbol{F} \in \mathcal{F}, \boldsymbol{A}_s \in \mathcal{A}_s \end{array} \tag{4.9}$$

where $||\cdot||_F$ is the Frobenius norm, λ and β are regularization parameters, and the set $\mathcal{K}$ constrains the delays $\boldsymbol{d}^s$ to be order-preserving:

$$\mathcal{K} = \left\{\boldsymbol{d} \in \{0, \cdots, d_{\max}\}^f : d_{i+1} \geq d_i, \forall i\right\}. \tag{4.10}$$

where $d_{\max} \leq n$. The other soft and hard constraints are briefly described as follows.

[1] $\mathcal{G}$ is superdiagonal tensor when $\mathcal{G}_{ijk} = 0$ except for $i = j = k$.

For the gene expression application we wish to extract factors $\boldsymbol{F}$ that are smooth over time and non-negative. Smoothness will be captured by the constraint that $P_2(\boldsymbol{F})$ is small where $P_2(\boldsymbol{F})$ is the squared total variation operator

$$P_2(\boldsymbol{F}) = \sum_{i=1}^{f} ||\boldsymbol{W}\boldsymbol{F}_{\cdot,i}||_2^2 \tag{4.11}$$

where $\boldsymbol{W}$ is an appropriate weighting matrix and $\boldsymbol{F}_{\cdot,i}$ denotes the i-th column of matrix $\boldsymbol{F}$. From *A4*, the data is non-negative and hence non-negativity is enforced on $\boldsymbol{F}$ and the loadings $\boldsymbol{A}_s$ to avoid masking of positive and negative valued factors whose overall contribution sums to zero. To avoid numerical instability associated with the scale invariance $\boldsymbol{M}\boldsymbol{A} = \frac{1}{\alpha}\boldsymbol{M}\alpha\boldsymbol{A}$ for any $\alpha > 0$, we constrain the Frobenius norm of $\boldsymbol{F}$. This leads to the following constraint sets:

$$\begin{aligned} \mathcal{F} &= \left\{ \boldsymbol{F} \in \mathbb{R}_+^{n\times f} : ||\boldsymbol{F}||_F \leq \delta \right\} \\ \mathcal{A}_s &= \mathbb{R}_+^{f\times p}, s = 1, \cdots, S \end{aligned} \tag{4.12}$$

The parameter δ above will be fixed to a positive value as its purpose is purely computational and has little practical impact. Since the factors $\boldsymbol{F}$ are common to all subjects, assumption *A3* requires that the number of columns of $\boldsymbol{F}$ (and therefore, its rank) is small compared to the number of genes p. In order to enforce *A1* we consider two different models. In the first model, which we shall name OPFA, we constrain the columns of $\boldsymbol{A}_s$ to be sparse and the sparsity pattern to be consistent across different subjects. Notice that *A1* does not imply that the mixing weights $\boldsymbol{A}_s$ are the same for all subjects as this would not accommodate magnitude variability across subjects. We also consider a more restrictive model where we constrain $\boldsymbol{A}_1 = \cdots = \boldsymbol{A}_S = \boldsymbol{A}$ with sparse $\boldsymbol{A}$ and we call this model OPFA-C, the C standing for the additional constraint that the subjects share the same sequence $\mathbf{A}$ of mixing weights. The OPFA-C model has a smaller number of parameters than OPFA, possibly at the expense of introducing bias with respect to the unconstrained model. A similar constraint has been succesfully adopted in (JSD10) in a factor model for multi-view learning.

Similarly to the approach taken in (MBP$^+$10) in the context of simultaneous sparse coding, the common sparsity pattern for OPFA is enforced by constraining $P_1(\boldsymbol{A}_1, \cdots, \boldsymbol{A}_S)$ to be small, where P_1 is a mixed-norm group-Lasso type penalty function (YL06b). For each of the $p \times f$ score variables, we create a group containing

its S different values across subjects:

$$P_1\left(\boldsymbol{A}_1, \cdots, \boldsymbol{A}_S\right) = \sum_{i=1}^{p}\sum_{j=1}^{f} \| \left[\boldsymbol{A}_1\right]_{j,i} \cdots \left[\boldsymbol{A}_S\right]_{j,i} \|_2. \tag{4.13}$$

Table 4.2 summarizes the constraints of each of the models considered in this chapter.

Following common practice in factor analysis, the non-convex problem (4.9) is addressed using Block Coordinate Descent, which iteratively minimizes (4.9) with respect to the shift parameters $\{\boldsymbol{d}^s\}_{s=1}^S$, the scores $\{\boldsymbol{A}_s\}_{s=1}^S$ and the factors $\boldsymbol{F}$ while keeping the other variables fixed. Starting from an initial estimate of $\boldsymbol{F}$ and $\{\boldsymbol{A}_s\}_{s=1}^S$, and given ϵ, λ and β, at iteration t we compute:

$$\begin{aligned}
\{\boldsymbol{d}^s\}_{s=1}^S &\leftarrow \text{EstimateDelays}\left(\boldsymbol{F}, \{\boldsymbol{A}_s\}_{s=1}^S\right) \\
\{\boldsymbol{A}_s\}_{s=1}^S &\leftarrow \text{EstimateScores}\left(\boldsymbol{F}, \{\boldsymbol{d}^s\}_{s=1}^S\right) \\
\boldsymbol{F} &\leftarrow \text{EstimateFactors}\left(\{\boldsymbol{A}_s\}_{s=1}^S, \{\boldsymbol{d}^s\}_{s=1}^S\right) \\
t &\leftarrow t+1
\end{aligned}$$

and stop the algorithm whenever $c^{t-1} - c^t \geq \epsilon$. This algorithm is guaranteed to monotonically decrease the objective function at each iteration. Since both the Frobenius norm and $P_1(\cdot)$, $P_2(\cdot)$ are non-negative functions, this ensures that the algorithm converges to a (possibly local) minima or a saddle point of (4.9).

The subroutines EstimateFactors and EstimateScores solve the following penalized regression problems:

$$\begin{aligned}
&\min_{\boldsymbol{F}} \quad \textstyle\sum_{s=1}^S ||\boldsymbol{X}_s - \boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\boldsymbol{A}_s||_F^2 + \beta \sum_{i=1}^f ||\boldsymbol{W}\boldsymbol{F}_{\cdot,i}||_2^2 \\
&\text{s.t.} \qquad \begin{cases} ||\boldsymbol{F}||_{\mathcal{F}}^2 \leq \delta & \\ \boldsymbol{F}_{i,j} \geq 0 & i = 1, \cdots, n, \\ & j = 1, \cdots, f \end{cases}
\end{aligned} \tag{4.14}$$

and

$$\begin{aligned}
&\min_{\{\boldsymbol{A}_s\}_{s=1}^S} \quad \textstyle\sum_{s=1}^S ||\boldsymbol{X}_s - \boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\boldsymbol{A}_s||_F^2 + \lambda \sum_{i=1}^p \sum_{j=1}^f \| \left[\boldsymbol{A}_1\right]_{j,i} \cdots \left[\boldsymbol{A}_S\right]_{j,i} \|_2 \quad (4.15) \\
&\text{s.t.} \qquad \begin{cases} \left[\boldsymbol{A}_s\right]_{j,i} \geq 0 & i = 1, \cdots, n, \\ & j = 1, \cdots, f, \\ & s = 1, \cdots, S \end{cases} \quad (4.16)
\end{aligned}$$

Notice that in OPFA-C, we also incorporate the constraint $\boldsymbol{A}_1 = \cdots = \boldsymbol{A}_S$ in the optimization problem above. The first is a convex quadratic problem with a quadratic and a linear constraint over a domain of dimension fn. In the applications considered here, both n and f are small and hence this problem can be solved using any standard convex optimization solver. EstimateScores is trickier because it involves a non-differentiable convex penalty and the dimension of its domain is equal to[2] Sfp, where p can be very large. In our implementation, we use an efficient first-order method (PCP08) designed for convex problems involving a quadratic term, a non-smooth penalty and a separable constraint set. These procedures are described in more detail in Appendix C.3 and therefore we focus on the EstimateDelays subroutine. EstimateDelays is a discrete optimization that is solved using a branch-and-bound (BB) approach (LW66). In this approach a binary tree is created by recursively dividing the feasible set into subsets ("branch"). On each of the nodes of the tree lower and upper bounds ("bound") are computed. When a candidate subset is found whose upper bound is less than the smallest lower bound of previously considered subsets these latter subsets can be eliminated ("prune") as candidate minimizers. Whenever a leaf (singleton subset) is obtained, the objective is evaluated at the corresponding point. If its value exceeds the current optimal value, the leaf is rejected as a candidate minimizer, otherwise the optimal value is updated and the leaf included in the list of candidate minimizers. Details on the application of BB to OPFA are given below.

The subroutine EstimateDelays solves S uncoupled problems of the form:

$$\min_{\boldsymbol{d}\in\mathcal{K}} ||\boldsymbol{X}_s - \boldsymbol{M}(\boldsymbol{F},\boldsymbol{d})\,\boldsymbol{A}_s||_F^2\,, \tag{4.17}$$

where the set $\mathcal{K}$ is defined in (4.10). The "branch" part of the optimization is accomplished by recursive splitting of the set $\mathcal{K}$ to form a binary tree. The recursion is initialized by setting $\mathcal{S}_o = \{0,\cdots,d_{\max}\}^f$, $\mathcal{I}_o = \{\boldsymbol{d}\in\mathcal{K}\cap\mathcal{S}_o\}$. The splitting of the set $\mathcal{I}_o$ into two subsets is done as follows

$$\begin{aligned} \mathcal{I}_1 &= \{\boldsymbol{d}\in\mathcal{K}\cap\mathcal{S}_o : d_{\omega_1}\le\gamma_1\} \\ \mathcal{I}_2 &= \{\boldsymbol{d}\in\mathcal{K}\cap\mathcal{S}_o : d_{\omega_1}>\gamma_1\}\,, \end{aligned} \tag{4.18}$$

and we update $\mathcal{S}_1 = \{\boldsymbol{d}\in\mathcal{S}_o : d_{\omega_1}\le\gamma_1\}$, $\mathcal{S}_2 = \{\boldsymbol{d}\in\mathcal{S}_o : d_{\omega_1}>\gamma_1\}$. Here γ_1 is an integer $0\le\gamma_1\le d_{\max}$, and $\omega_1\in\{1,\cdots,f\}$. $\mathcal{I}_1$ contains the elements $\boldsymbol{d}\in\mathcal{K}$ whose ω_1-th component is strictly larger than γ_1 and $\mathcal{I}_2$ contains the elements whose ω_1-th

[2]This refers to the OPFA model. In the OPFA-C model, the additional constraint $\boldsymbol{A}_1 = \cdots = \boldsymbol{A}_S = \boldsymbol{A}$ reduces the dimension to fp.

component is smaller than γ_1. The same kind of splitting procedure is then subsequently applied to $\mathcal{I}_1$, $\mathcal{I}_2$ and its resulting subsets. After $k-1$ successive applications of this decomposition there will be 2^{k-1} subsets and the k-th split will be :

$$\begin{aligned} \mathcal{I}_t &:= \quad \{\boldsymbol{d} \in \mathcal{K} \cap \mathcal{S}_t\} \\ \mathcal{I}_{t+1} &:= \quad \{\boldsymbol{d} \in \mathcal{K} \cap \mathcal{S}_{t+1}\} \end{aligned} \tag{4.19}$$

where

$$\begin{aligned} \mathcal{S}_t &= \{\boldsymbol{d} \in \mathcal{S}_{\pi_k} : d_{\omega_k} \leq \gamma_k\} \\ \mathcal{S}_{t+1} &= \{\boldsymbol{d} \in \mathcal{S}_{\pi_k} : d_{\omega_k} > \gamma_k\}. \end{aligned}$$

and $\pi_k \in \left\{1, \cdots, 2^{k-1}\right\}$ denotes the parent set of the two new sets t and $t+1$, i.e. $\mathrm{pa}(t) = \pi_k$ and $\mathrm{pa}(t+1) = \pi_k$. In our implementation the splitting coordinate ω_k is the one corresponding to the coordinate in the set $\mathcal{I}_{\pi_k}$ with largest interval. The decision point γ_k is taken to be the middle point of this interval.

The "bound" part of the optimization is as follows. Denote $g(\boldsymbol{d})$ the objective function in (4.17) and define its minimum over the set $\mathcal{I}_t \subset \mathcal{K}$:

$$g_{\min}(\mathcal{I}_t) = \min_{\boldsymbol{d} \in \mathcal{I}_t} g(\boldsymbol{d}). \tag{4.20}$$

A lower bound for this value can be obtained by relaxing the constraint $\boldsymbol{d} \in \mathcal{K}$ in (4.19):

$$\min_{\boldsymbol{d} \in \mathcal{S}_t} g(\boldsymbol{d}) \leq g_{\min}(\mathcal{I}_t) \tag{4.21}$$

Letting $\boldsymbol{X}_s = \boldsymbol{X}_s^{\perp} + \boldsymbol{X}_s^{\parallel}$ where $\boldsymbol{X}_s^{\parallel} = \boldsymbol{X}_s \boldsymbol{A}_s^{\dagger} \boldsymbol{A}_s$ and $\boldsymbol{X}_s^{\perp} = \boldsymbol{X}_s \left(\boldsymbol{I} - \boldsymbol{A}_s^{\dagger} \boldsymbol{A}_s\right)$, we have:

$$\begin{aligned} ||\boldsymbol{X}_s - \boldsymbol{M}(\boldsymbol{F}, \boldsymbol{d}) \boldsymbol{A}_s||_F^2 &= \left|\left|\left(\boldsymbol{X}_s \boldsymbol{A}_s^{\dagger} - \boldsymbol{M}(\boldsymbol{F}, \boldsymbol{d})\right) \boldsymbol{A}_s\right|\right|_F^2 \\ &\quad + \left|\left|{\boldsymbol{X}_s}^{\perp}\right|\right|_F^2, \end{aligned}$$

where $\boldsymbol{A}_s^{\dagger}$ denotes the pseudoinverse of $\boldsymbol{A}_s$. This leads to:

$$\underline{\lambda}\left(\boldsymbol{A}_s \boldsymbol{A}_s^T\right) \left|\left|\boldsymbol{X}_s \boldsymbol{A}_s^{\dagger} - \boldsymbol{M}(\boldsymbol{F}, \boldsymbol{d})\right|\right|_F^2 + \left|\left|{\boldsymbol{X}_s}^{\perp}\right|\right|_F^2 \leq g(\boldsymbol{d}), \tag{4.22}$$

where $\underline{\lambda}\left(\boldsymbol{A}_s \boldsymbol{A}_s^T\right)$ denotes the smallest eigenvalue of the symmetric matrix $\boldsymbol{A}_s \boldsymbol{A}_s^T$. Combining the relaxation in (4.21) with inequality (4.22), we obtain a lower bound

on $g_{\min}(\mathcal{I}_t)$:

$$\begin{aligned}\Phi_{lb}(\mathcal{I}_t) &= \min_{\boldsymbol{d}\in\mathcal{S}_t} \underline{\lambda}\left(\boldsymbol{A}_s\boldsymbol{A}_s^T\right)\left|\left|\boldsymbol{X}_s\boldsymbol{A}_s^\dagger - \boldsymbol{M}(\boldsymbol{F},\boldsymbol{d})\right|\right|_F^2 \\ &\quad + \left|\left|\boldsymbol{X}_s{}^\perp\right|\right|_F^2 \\ &\leq g_{\min}(\mathcal{I}_t), \end{aligned} \tag{4.23}$$

which can be evaluated by performing f *decoupled* discrete grid searches. At the k-th step, the splitting node π_k will be chosen as the one with smallest $\Phi_{lb}(\mathcal{I}_t)$. Finally, this lower bound is complemented by the upper bound

$$g_{\min}(\mathcal{I}_t) \leq \Phi_{ub}(\mathcal{I}_t) = g(\boldsymbol{d}) \text{ for } \forall \boldsymbol{d}\in\mathcal{I}_t. \tag{4.24}$$

These bounds enable the branch-and-bound optimization of (4.17).

4.3.3 Selection of the tuning parameters f, λ and β

From (4.9), it is clear that the OPFA factorization depends on the choice of f, λ and β. This is a paramount problem in unsupervised learning, and several heuristic approaches have been devised for simpler factorization models (OP09; Wol78; WTH09). These approaches are based on training the factorization model on a subset of the elements of the data matrix (training set) to subsequently validate it on the excluded elements (test set).

The variational characterization of the OPFA decomposition allows for the presence of missing variables, i.e. missing elements in the observed matrices $\{\boldsymbol{X}_s\}_{s=1}^S$. In such case, the Least Squares fitting term in (4.9) is only applied to the observed set of indices[3]. We will hence follow the approach in (WTH09) and train the OPFA model over a fraction $1-\delta$ of the entries in the observations $\boldsymbol{X}_s$. Let Ω_s denote the set of $\delta(n\times p)$ excluded entries for the s-th observation. These entries will constitute our test set, and thus our Cross-Validation error measure is:

$$\mathrm{CV}(f,\lambda,\beta) = \frac{1}{S}\sum_{s=1}^{S}\left|\left|\left[\boldsymbol{X}_s - \boldsymbol{M}\left(\hat{\boldsymbol{F}},\hat{\boldsymbol{d}}^s\right)\hat{\boldsymbol{A}}_s\right]_{\Omega_s}\right|\right|_F^2$$

where $\hat{\boldsymbol{F}}, \left\{\hat{\boldsymbol{d}}^s\right\}_{s=1}^S, \left\{\hat{\boldsymbol{A}}_s\right\}_{s=1}^S$ are the OPFA estimates obtained on the training set excluding the entries in $\{\Omega_s\}_{s=1}^S$, for a given choice of f, λ, and β.

[3]See the Appendix C.3 and C.4 for the extension of the EstimateFactors, EstimateScores and Estimatedelays procedures to the case where there exist missing observations.

Model	$\boldsymbol{M}_s$	$\boldsymbol{A}_s$
OPFA	$\boldsymbol{M}_s = \boldsymbol{M}(\boldsymbol{F}, \boldsymbol{d}^s)$ $\boldsymbol{d}^s \in \mathcal{K}$, $\boldsymbol{F}$ smooth and non-negative	Non-negative sparse $\boldsymbol{A}_s$
OPFA-C	$\boldsymbol{M}_s = \boldsymbol{M}(\boldsymbol{F}, \boldsymbol{d}^s)$ $\boldsymbol{d}^s \in \mathcal{K}$, $\boldsymbol{F}$ smooth and non-negative	Non-negative sparse $\boldsymbol{A}_1 = \cdots = \boldsymbol{A}_S$
SFA	$\boldsymbol{M}_s = \boldsymbol{M}(\boldsymbol{F}, \boldsymbol{d}^s)$, $\boldsymbol{d}^s = \boldsymbol{0}$,$\boldsymbol{F}$ smooth and non-negative	Non-negative sparse $\boldsymbol{A}_s$

Table 4.2: Models considered in Section IV-A.

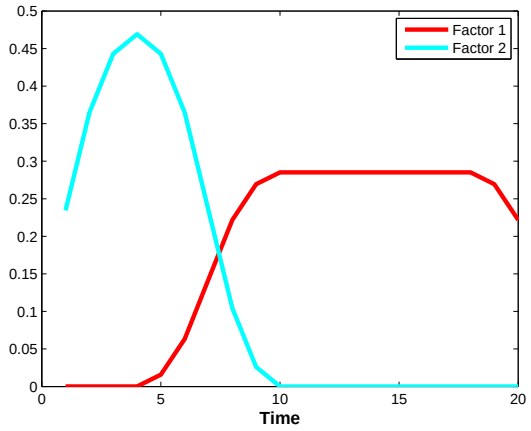


Figure 4.5: Dictionary used to generated the 2-factor synthetic data of Section 4.4.

4.4 Numerical results

4.4.1 Synthetic data: Periodic model

First we evaluate the performance of the OPFA algorithm for a periodic model observed in additive Gaussian white noise:

$$\boldsymbol{X}_s = \boldsymbol{M}(\boldsymbol{F}, \boldsymbol{d}^s)\boldsymbol{A}_s + \epsilon_s \quad s = 1, \ldots, S. \tag{4.25}$$

Here $\epsilon_s \sim \mathcal{N}_{n\times p}(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I})$, $\boldsymbol{d}^s = \text{sort}(\boldsymbol{t}^s)$ where σ_ϵ^2 is the variance of ϵ_s and $\boldsymbol{t}^s \sim \mathcal{U}\left(0, \sqrt{12\sigma_d^2 + 1}\right)$ are i.i.d. The $f = 2$ columns of $\boldsymbol{F}$ are non-random smooth signals from the predefined dictionary shown in Figure 4.5. The scores $\boldsymbol{A}_s$ are generated according to a consistent sparsity pattern across all subjects and its non zero elements are i.i.d. normal truncated to the non-negative orthant.

Here the number of subjects is $S = 10$, the number of variables is $p = 100$, and the number of time points is $n = 20$. In these experiments experiments we choose to initialize the factors $\boldsymbol{F}$ with temporal profiles obtained by hierarchical clustering of the data. Hierarchical clustering (HTF05) is a standard unsupervised learning technique that groups the p variables into increasingly finer partitions according to the normalized euclidean distance of their temporal profiles. The average expression patterns of the clusters found are used as initial estimates for $\boldsymbol{F}$. The loadings $\{\boldsymbol{A}_s\}_{s=1}^S$ are initialized by regressing the obtained factors onto the data.

We compare OPFA and OPFA-C to a standard Sparse Factor Analysis (SFA) solution, obtained by imposing $d_{\max} = 0$ in the original OPFA model. Table 4.2 summarizes the characteristics of the three models considered in the simulations. We fix $f = 2$ and choose the tuning parameters (λ, β) using the Cross-Validation procedure of Section 4.3.3 with a 5×3 grid and $\delta = .1$.

In these experiments, we consider two measures of performance, the Mean Square Error (MSE) with respect to the generated data:

$$MSE := \frac{1}{S} \sum_{s=1}^{S} E \left|\left| \boldsymbol{D}_s - \hat{\boldsymbol{D}}_s \right|\right|_F^2,$$

where E is the expectation operator, $\boldsymbol{D}_s = \boldsymbol{M}(\boldsymbol{F}, \boldsymbol{d}^s)\boldsymbol{A}_s$ is the generated noiseless data and $\hat{\boldsymbol{D}}_s = \boldsymbol{M}\left(\hat{\boldsymbol{F}}, \hat{\boldsymbol{d}}^s\right)\hat{\boldsymbol{A}}_s$ is the estimated data, and the Distance to the True Factors (DTF), defined as:

$$DTF := 1 - \frac{1}{f} \sum_{i=1}^{f} E \frac{\boldsymbol{F}_{\cdot,i}^T \hat{\boldsymbol{F}}_{\cdot,i}}{||\boldsymbol{F}_{\cdot,i}||_2 \left|\left| \hat{\boldsymbol{F}}_{\cdot,i} \right|\right|_2},$$

where $\boldsymbol{F}$, $\hat{\boldsymbol{F}}$ are the generated and the estimated factor matrices, respectively.

Figure 4.6 shows the estimated MSE and DTF performance curves as a function of the delay variance σ_d^2 for fixed SNR= 15dB (which is defined as $SNR = 10 \log\left(\frac{1}{S}\sum_s \frac{E\left(||\boldsymbol{M}(\boldsymbol{F},\boldsymbol{d}^s)\boldsymbol{A}_s||_F^2\right)}{np\sigma_\epsilon^2}\right)$). OPFA and OPFA-C perform at least as well as SFA for zero delay ($\sigma_d = 0$) and significantly better for $\sigma_d > 0$ in terms of DTF. OPFA-C outperforms OPFA for high delay variances σ_d^2 at the price of a larger MSE due to the bias introduced by the constraint $\boldsymbol{A}_1 = \cdots = \boldsymbol{A}_S$. In Figure 4.7 the performance curves are plotted as a function of SNR, for fixed $\sigma_d^2 = 5$. Note that OPFA and OPFA-C outperform SFA in terms of DTF and that OPFA is better than the others in terms of MSE for SNR$>$ 0db. Again, OPFA-C shows increased robustness

	DTF [mean (standard deviation)] $\times 10^{-3}$		
SNR	$\rho = 0.002$	$\rho = 1.08$	$\rho = 53.94$
22.8	0.0 (0.0)	3.4 (9.4)	1.9 (3.2)
-2	1.3 (0.5)	1 (9.4)	1.25 (1.5)
-27.1	46 (20)	58 (17)	63 (8)
	MSE [mean (standard deviation $\times 10^{-3}$)]		
SNR	$\rho = 0.002$	$\rho = 1.08$	$\rho = 53.94$
22.8	0.02 (1.5)	0.05 (69)	0.11 (99)
-2	0.35 (7.9)	0.36(22)	0.38 (32)
-27.1	0.96 (19)	0.99 (18)	1.00 (24)

Table 4.3: Sensitivity of the OPFA estimates to the initialization choice with respect to the relative norm of the perturbation (ρ).

to noise in terms of DTF.

We also performed simulations to demonstrate the value of imposing the order-preserving constraint in (4.17). This was accomplished by comparing OPFA to a version of OPFA for which the constraints in (4.17) are not enforced. Data was generated according to the model (4.25) with $S = 4$, $n = 20$, $f = 2$, and $\sigma_d^2 = 5$. The results of our simulations (not shown) were that, while the order-preserving constraints never degrade OPFA performance, the constraints improve performance when the SNR is small (below 3dB for this example).

Finally, we conclude this sub-section by studying the sensitivity of the final OPFA estimates with respect to the initialization choice. To this end, we initialize the OPFA algorithm with the correct model perturbed with a random gaussian vector of increasing variance. We analyze the performance of the estimates in terms of MSE and DTF as a function of the norm of the model perturbation relative to the norm of the noiseless data, which we denote by ρ. Notice that larger ρ corresponds to increasingly random initialization. The results in Table 4.3 show that the MSE and DTF of the OPFA estimates are very similar for a large range of values of ρ, and therefore are robust to the initialization.

4.4.2 Experimental data: Predictive Health and Disease (PHD)

The PHD data set was collected as part of a viral challenge study that is described in (ZCV$^+$09). In this study 20 human subjects were inoculated with live H3N2 virus and Genechip mRNA gene expression in peripheral blood of each subject was measured over 16 time points. The raw Genechip array data was pre-processed using

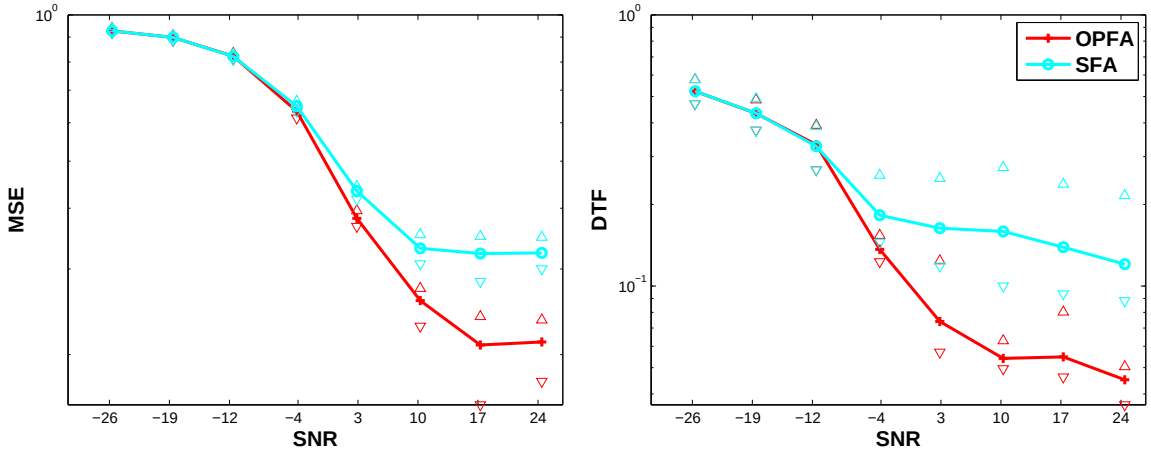


Figure 4.6: MSE (top) and DTF (bottom) as a function of delay variance σ_d^2 for OPFA and Sparse Factor Analysis (SFA). These curves are plotted with 95% confidence intervals. For $\sigma_d^2 > 0$, OPFA outperforms SFA both in MSE and DTF, maintaining its advantage as σ_d increases. For large σ_d, OPFA-C outperforms the other two.

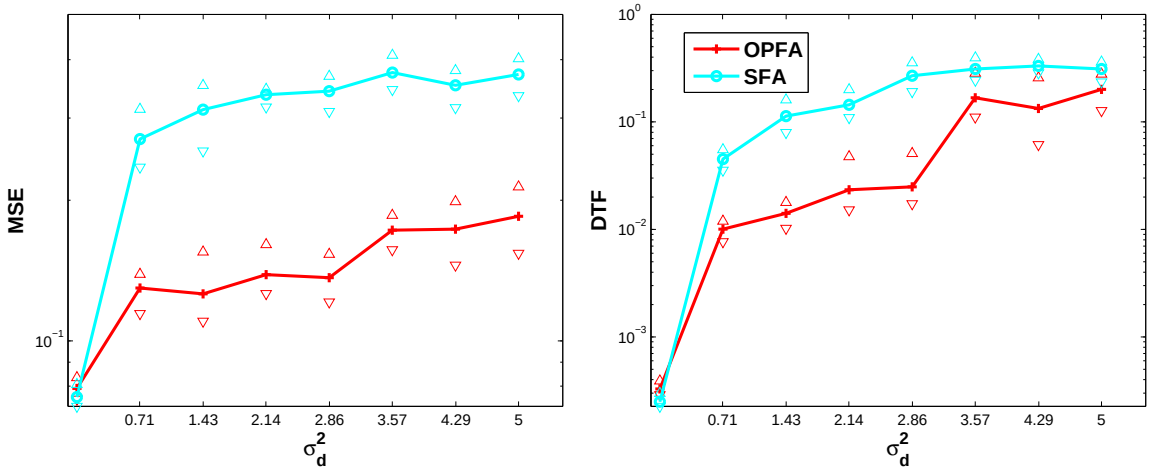


Figure 4.7: Same as Figure 4.6 except that the performance curves are plotted with respect to SNR for fixed $\sigma_d^2 = 5$.

robust multi-array analysis (IHC$^+$03) with quantile normalization (BIAS03). In this section we show results for the constrained OPFA model (OPFA-C). While not shown here, we have observed that OPFA-C gives very similar results to unconstrained OPFA but with reduced computation time.

Specifically, we use OPFA-C to perform the following tasks:

1. *Subject Alignment:* Determine the alignment of the factors to fit each subject's response, therefore revealing each subject's intrinsic response delays.

2. *Gene Clustering:* Discover groups of genes with common expression signature by clustering in the low-dimensional space spanned by the aligned factors. Since we are using the OPFA-C model, the projection of each subject's data on this lower dimensional space is given by the scores $\boldsymbol{A} := \boldsymbol{A}_1 = \cdots = \boldsymbol{A}_S$. Genes with similar scores will have similar expression signatures.

3. *Symptomatic Gene Signature discovery:* Using the gene clusters obtained in step 2 we construct temporal signatures common to subjects who became sick.

The data was normalized by dividing each element of each data matrix by the sum of the elements in its column. Since the data is non-negative valued, this will ensure that the mixing weights of different subjects are within the same order of magnitude, which is necessary to respect the assumption that $\boldsymbol{A}_1 = \cdots = \boldsymbol{A}_S$ in OPFA-C. In order to select a subset of strongly varying genes, we applied one-way Analysis-Of-Variance (NWK$^+$96) to test for the equality of the mean of each gene at 4 different groups of time points, and selected the first $p = 300$ genes ranked according to the resulting F-statistic. To these gene trajectories we applied OPFA-C to the $S = 9$ symptomatic subjects in the study. In this context, the columns in $\boldsymbol{F}$ are the set of signals emitted by the common immune system response and the vector $\boldsymbol{d}^s$ parameterizes each subject's characteristic onset times for the factors contained in the columns of $\boldsymbol{F}$. To avoid wrap-around effects, we worked with a factor model of dimension $n = 24$ in the temporal axis.

The OPFA-C algorithm was run with 4 random initializations and retained the solution yielding the minimum of the objective function (6). For each $f = 1, \cdots, 5$ (number of factors), we estimated the tuning parameters (λ, β) following the Cross-Validation approach described in 4.3.3 over a 10×3 grid. The resulting results, shown in Table 4.4 resulted in selecting $\beta = 1 \times 10^{-8}$, $\lambda = 1 \times 10^{-8}$ and $f = 3$. The choice of three factors is also consistent with an expectation that the principal gene trajectories over the period of time studied are a linear combination of increasing, decreasing or constant expression patterns (ZCV$^+$09).

To illustrate the goodness-of-fit of our model, we plot in Figure 4.8 the observed gene expression patterns of 13 strongly varying genes and compare them to the OPFA-C fitted response for three of the subjects, together with the relative residual error. The average relative residual error is below 10% and the plots demonstrate the agreement between the observed and the fitted patterns. Figure 4.9 shows the trajectories

	$f=1$	$f=2$	$\mathbf{f=3}$	$f=4$	$f=5$
$\min \mathrm{CV}(f,\lambda,\beta)$	20.25	13.66	**12.66**	12.75	12.72
Relative residual error ($\times 10^{-3}$)	7.2	4.8	**4.5**	4.5	4.4
$\hat{\lambda}(\times 10^{-8})$	5.99	1	**1**	1	35.9
$\hat{\beta}$ ($\times 10^{-6}$)	3.16	3.16	**0.01**	0.01	100

Table 4.4: Cross Validation Results for Section 4.4.2.

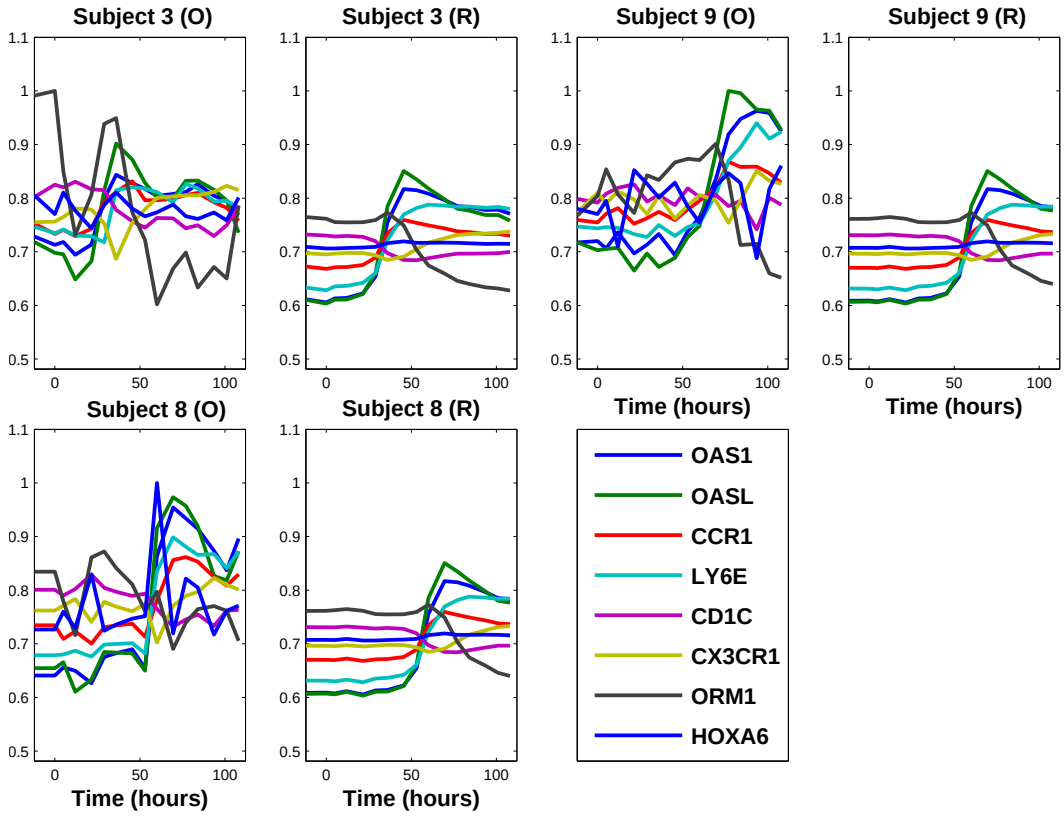


Figure 4.8: Comparison of observed (O) and fitted responses (R) for three of the subjects and a subset of genes in the PHD data set. Gene expression profiles for all subjects were reconstructed with a relative residual error below 10%. The trajectories are smoothed while respecting each subject's intrinsic delay.

for each subject for four genes having different regulation motifs: up-regulation and down-regulation. It is clear that the gene trajectories have been smoothed while conserving their temporal pattern and their precedence-order, e.g. the up-regulation of gene *OAS1* consistently follows the down-regulation of gene *ORM1*.

In Figure 4.10 we show the 3 factors along with the factor delays and factor loading

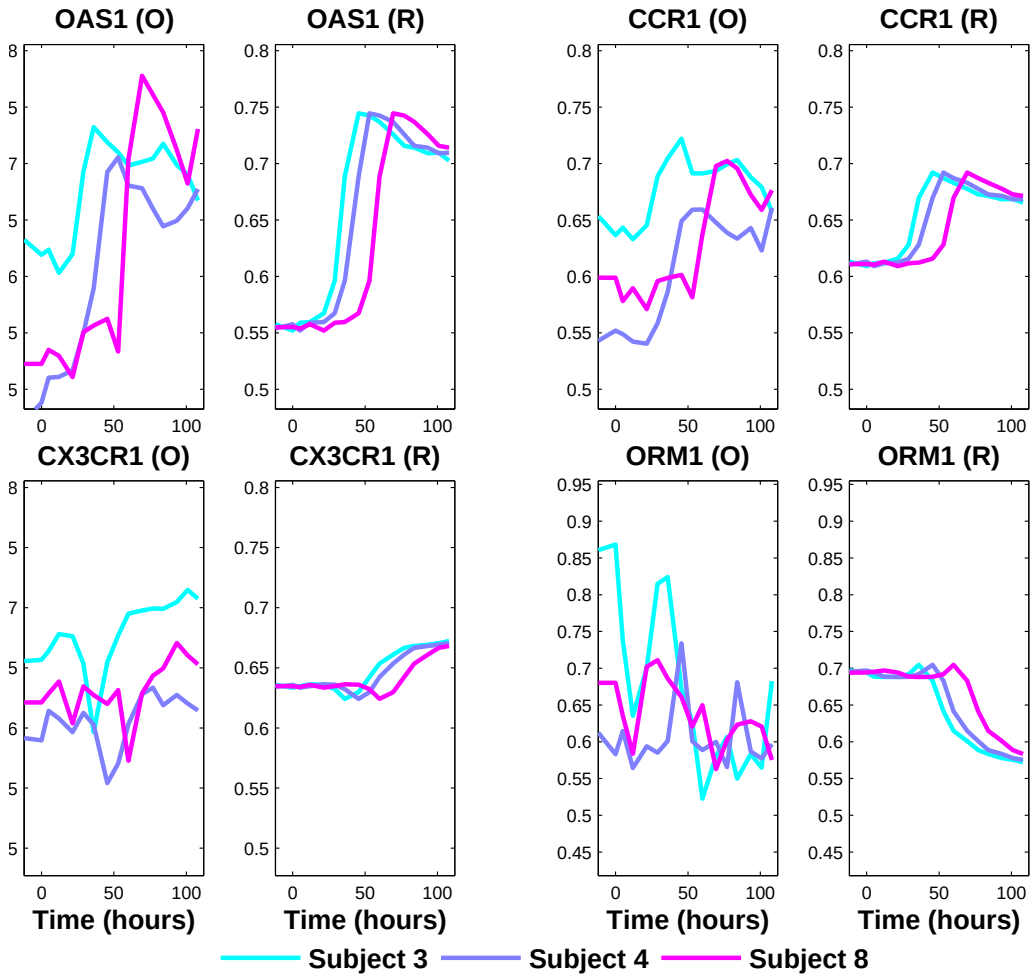


Figure 4.9: Comparison of observed (O) and fitted responses (R) for four genes (*OAS1*, *CCR1*, *CX3CR1*, *ORM1*) showing up-regulation and down-regulation motifs and three subjects in the PHD dataset. The gene trajectories have been smoothed while conserving their temporal pattern and their precedence-order. The OPFA-C model revealed that *OAS1* up-regulation occurs consistently after *ORM1* down-regulation among all subjects.

discovered by OPFA-C. The three factors, shown in the three bottom panels of the figure, exhibit features of three different motifs: factor 1 and factor 3 correspond to up-regulation motifs occurring at different times; and factor 2 is a strong down-regulation motif. The three top panels show the onset times of each motif as compared to the clinically determined peak symptom onset time. Note, for example, that the strong up-regulation pattern of the first factor coincides closely with the onset peak time. Genes strongly associated to this factor have been closely associated to acute anti-

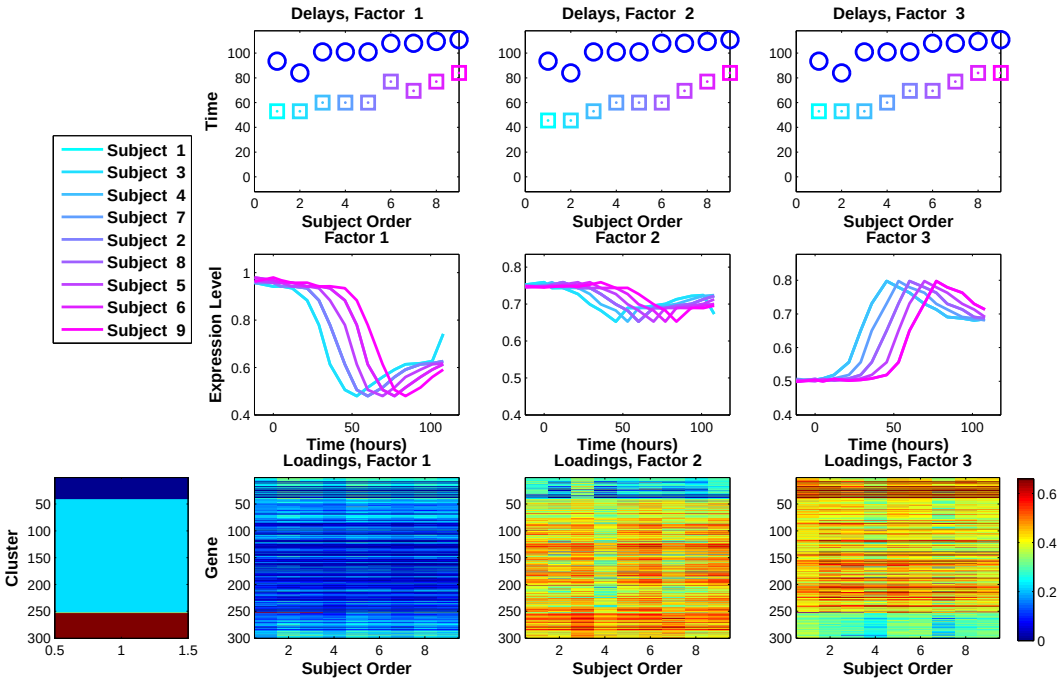


Figure 4.10: Top plots: Motif onset time for each factor (□) and peak symptom time reported by expert clinicians (O). Bottom plots: Aligned factors for each subject. Factor 1 and 3 can be interpreted as up-regulation motifs and factor 2 is a strong down-regulation pattern. The arrows show each factor's motif onset time.

viral and inflammatory host response (ZCV+09). Interestingly, the down-regulation motifs associated with factor 2 consistently precedes this up-regulation motif.

Finally, we consider the application of OPFA as a pre-processing step preceding a clustering analysis. Here the goal is to find groups of genes that share similar expression patterns and determine their characteristic expression patterns. In order to obtain gene clusters, we perform hierarchical clustering on the raw data ($\{\boldsymbol{X}_s\}_{s=1}^S$) and on the lower dimensional space of the estimated factor scores ($\{\boldsymbol{A}_s\}_{s=1}^S$), obtaining two different sets of 4 well-differentiated clusters. We then compute the average expression signatures of the genes in each cluster by averaging over the observed data ($\{\boldsymbol{X}_s\}_{s=1}^S$) and averaging over the data after OPFA correction for the temporal misalignments. Figure 4.11 illustrates the results. Clustering using the OPFA-C factor scores produces a very significant improvement in cluster concentration as compared to clustering using the raw data $\{\boldsymbol{X}_s\}_{s=1}^S$. The first two columns in Figure compare the variation of the gene profiles over each cluster for the temporally realigned data (labeled `A´) as compared to to the profile variation of these same genes for the

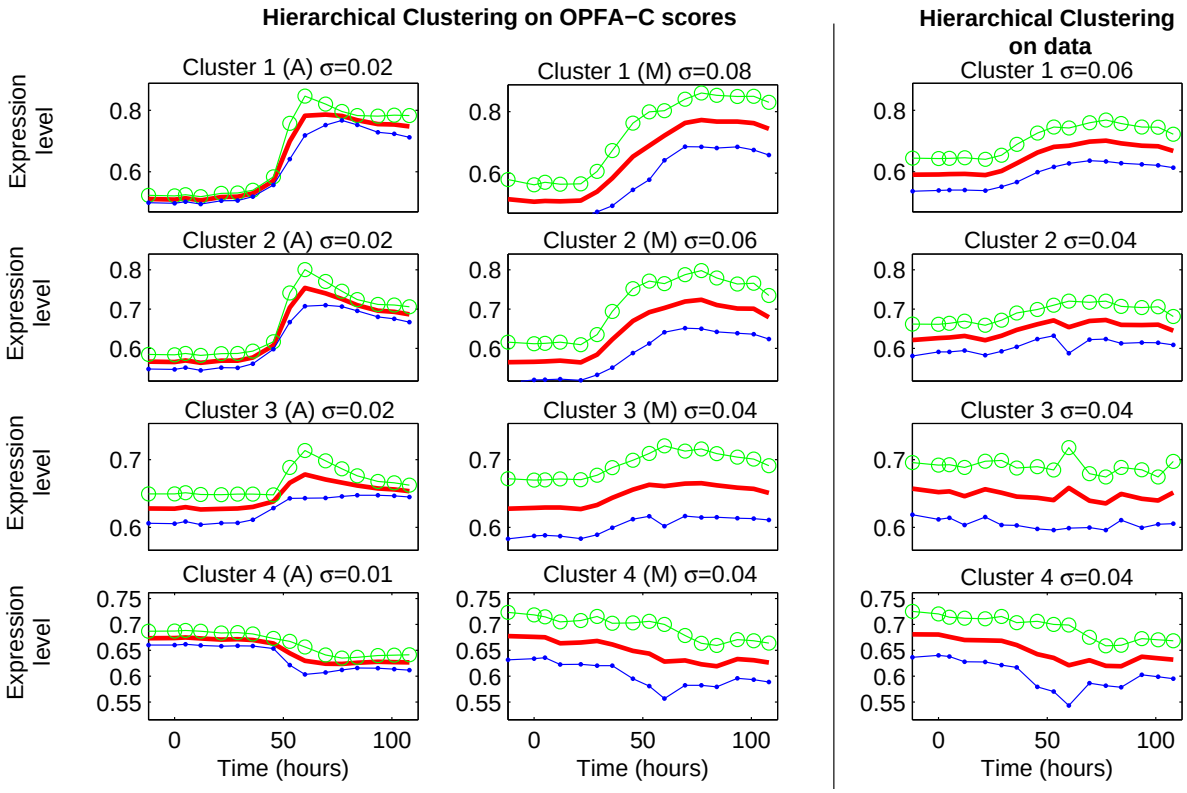


Figure 4.11: The first two columns show the average expression signatures and their estimated upper/lower confidence intervals for each cluster of genes obtained by: averaging the *estimated Aligned* expression patterns over the $S = 9$ subjects (A) and directly averaging the misaligned observed data for each of the gene clusters obtained from the OPFA-C scores (M). The confidence intervals are computed according to $+/-$ the estimated standard deviation at each time point. The cluster average standard deviation (σ) is computed as the average of the standard deviations at each time point. The last column shows the results of applying hierarchical clustering directly to the original misaligned dataset $\{\boldsymbol{X}_s\}_{s=1}^{S}$. In the first column, each gene expression pattern is obtained by mixing the estimated aligned factors $\boldsymbol{F}$ according to the estimated scores $\boldsymbol{A}$. The alignment effect is clear, and interesting motifs become more evident.

misaligned observed data (labeled `M´). For comparison, the last column shows the results of applying hierarchical clustering directly to the original misaligned dataset $\{\boldsymbol{X}_s\}_{s=1}^{S}$. It is clear that clustering on the low-dimensional space of the OPFA-C scores unveils interesting motifs from the original noisy temporal expression trajectories.

4.5 Conclusions

We have proposed a general method of order-preserving factor analysis that accounts for possible temporal misalignments in a population of subjects undergoing a common treatment. We have described a simple model based on circular-shift translations of prototype motifs and have shown how to embed transient gene expression time courses into this periodic model. The OPFA model can significantly improve interpretability of complex misaligned data. The method is applicable to other signal processing areas beyond gene expression time course analysis.

A Matlab package implementing OPFA and OPFA-C is available at the Hero Group Reproducible Research page (http://tbayes.eecs.umich.edu).

CHAPTER V

Misaligned Principal Components Analysis

5.1 Introduction

Principal Component Analysis (PCA) (Hot33) is a widely used technique for dimensionality-reduction of high dimensional data, with applications in pattern recognition (PK93), blind channel estimation (MDCM02) and network-traffic anomaly detection (LCD04). In all these applications, PCA can be used to separate the latent features corresponding to signal from the random fluctuations of noise. The extracted features can then be utilized for interpretation, classification or prediction purposes. The fundamental assumption underlying this approach is that the signal lies in a lower dimensional subspace, while the noise is random and isotropic; spreading its power across all directions in the observation space.

Unfortunately, in many cases, despite the appropriateness of the low-dimensional subspace model, measurement limitations can lead to different observations revealing very different signal subspaces. One important situation where this occurs arises when the sampling times of each observation batch cannot be synchronized appropriately. These synchronizations problems can be viewed as due to (i) technical limitations in the sampling procedure or (ii) different temporal latencies of the phenomena under study. Examples of the first appear in music signal processing (BD05) or uncalibrated arrays of antennas (NM96; SK00),

The second situation occurs for instance in multi-path communications (VdVVP02), sensor network-based geolocalization systems (PHIP$^+$03), speech (CB83), image (PK93), genomic (TPWZ$^+$11; BJ04), proteomic (FGR$^+$06) or Electro-cardiogram (ECG) (SL06) signal processing. In the previous chapter, we considered an Order Preserving Factor Analysis (OPFA) model that accounted for order-preserving circular shifts in each factor and we demonstrated its effectiveness for extracting order-preserving factors

from misaligned data. Here, we propose an alternative approach to OPFA that applies to misaligned data without order restrictions and is applicable to larger sample sizes.

In this chapter, we first consider the limitations of PCA for the problem of estimating a rank-F, $F \geq 1$, signal subspace from high-dimensional misaligned data. We introduce a modified version of PCA, called Misaligned PCA (MisPCA), which simultaneously aligns the data and estimates the aligned signal subspace.

The chapter is divided into two parts. First, we propose a simple approximation of the combinatorial MisPCA estimation problem that considerably improves the PCA estimate whenever misalignments are present. Second, building on recent results in random matrix theory (Pau07; BGN11), we derive high-dimensional asymptotic results that characterize the minimum SNR necessary to detect and estimate the signal from the sample covariance under a Gaussian observation model. (The Gaussian model assumption is common in our setting but may not be necessary for the derivation of the theoretical results, as illustrated by the more general setting of (BGN11).)

This chapter is organized as follows. Section 5.2 introduces the misaligned signal model. We give algorithms for Misaligned PCA in Section 5.3. Section 5.4 studies the statistical effects of misalignments on the sample covariance. We present numerical results and a gene expression data analysis application in Section 5.5 and we conclude the chapter in Section 4.5.

5.2 Problem Formulation

We consider the following discrete-time, circularly misaligned, rank-F signal model,

$$x_i[k] \quad = \sum_{f=1}^{F} a_f^i h_f[k - d_i] + \epsilon_i[k], \quad i = 1, \cdots, n. \tag{5.1}$$

Here $h_f[k]$ are unknown orthogonal real sequence of length equal to p and indexed by k, and the integer valued elements of the vector $\boldsymbol{d} \in \{0, \cdots, d_{\max}\}^n$ parameterize the amount of circular shift in each observation, with $d_{\max} < p$. For each $i = 1, \cdots, n$, the random variables a_i are i.i.d, zero-mean Gaussian and the p-length sequences $\epsilon_i[k]$ are i.i.d., zero-mean Gaussian white processes. To simplify the notation, we will further assume that $E[\epsilon_i^2[k]] = 1$ and $\sum_{k=1}^{p} h_f^2[k] = 1$. We denote each component's power by:

$$\sigma_f = E\left[\left(a_f^i\right)^2\right], \quad f = 1, \cdots, F. \tag{5.2}$$

The signal-to-noise ratio (SNR) of model (5.1) is defined as:

$$\text{SNR} = \max_{f=1,\cdots,F} \frac{\sigma_f}{E\left[\epsilon_i^2\left[k\right]\right]}, \tag{5.3}$$

and we define each component's normalized power as:

$$\bar{\sigma}_f = \frac{\sigma_f}{\max_{f=1,\cdots,F}\sigma_f}, \quad f = 1, \cdots, F. \tag{5.4}$$

The problem considered in this paper is that of estimating the signal sequences $h_f\left[k\right]$, $f = 1, \cdots, F$, from a collection of observations obeying model (5.1). For convenience, we will write (5.1) in vector form:

$$\boldsymbol{x}_i \quad = \boldsymbol{C}_{d_i}\boldsymbol{H}\boldsymbol{a}^i + \boldsymbol{\epsilon}_i, \quad i = 1, \cdots, n,$$

where $\boldsymbol{x}_i$ and $\boldsymbol{\epsilon}_i$ are p-dimensional real vectors, $\boldsymbol{a}_i$ is a real vector of dimension F, $\boldsymbol{H} = [\boldsymbol{H}_1, \cdots, \boldsymbol{H}_F]$ is a $p \times F$ matrix of factors such that $\boldsymbol{H}^T\boldsymbol{H} = \boldsymbol{I}_F$, and $\boldsymbol{C}_{d_i}$ is a $p \times p$ circular shift matrix with shift equal to d_i:

$$\left[\boldsymbol{C}_{d_i}\right]_{k,l} = \begin{cases} 1 & \text{if } k = (d_i + l) \bmod p \\ 0 & \text{otherwise.} \end{cases}$$

Using the properties of $\boldsymbol{a}^i$ and $\boldsymbol{\epsilon}_i$ we can conclude that $\boldsymbol{x}_i$ follows a multivariate Gaussian distribution with zero mean and covariance:

$$\boldsymbol{\Sigma}_i = E\left[\boldsymbol{x}_i\boldsymbol{x}_i^T\right] = \text{SNR}\boldsymbol{C}_{d_i}\boldsymbol{H}\text{diag}\,\bar{\boldsymbol{\sigma}}\boldsymbol{H}^T\boldsymbol{C}_{d_i}^T + \boldsymbol{I}_p. \tag{5.5}$$

5.3 Algorithms

In general, the covariance matrix of each observation is not the same for all $i = 1, \cdots, n$. However, equation (5.5) reflects an underlying rank-F structure corresponding to the signals $\boldsymbol{H}$. In this section we propose to exploit this fact by estimating $\boldsymbol{H}$ from the joint likelihood of the misaligned data $\{\boldsymbol{x}_i\}_{i=1}^n$. The log-likelihood function is:

$$l\left(\boldsymbol{H}, \boldsymbol{d}, \boldsymbol{\sigma}\right) = c - \sum_{i=1}^{n} \text{tr}\left(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\right) - \sum_{i=1}^{n} \log\det\boldsymbol{\Sigma}_i$$

where c denotes a constant independent of the relevant parameters. Using the Sherman-Morrison-Woodbury matrix inversion formula,

$$l\left(\boldsymbol{H}, \boldsymbol{d}, \boldsymbol{\sigma}\right) = c + n \sum_{f=1}^{F} \frac{\sigma_f}{1+\sigma_f} \boldsymbol{H}_f^T \boldsymbol{S}\left(\boldsymbol{d}\right) \boldsymbol{H}_f - n \sum_{f=1}^{F} \log\left(\sigma_f + 1\right),$$

where, for any $\boldsymbol{\tau} \in \{0, \cdots, d_{\max}\}^n$, possibly different from $\boldsymbol{d}$, we define the $p \times p$ matrix:

$$\boldsymbol{S}\left(\boldsymbol{\tau}\right) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{C}_{\tau_i}^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{C}_{\tau_i}. \tag{5.6}$$

This quantity can be interpreted as an *aligned sample covariance matrix*, with alignment parameter equal to $\boldsymbol{\tau}$. When $\boldsymbol{\tau} = \boldsymbol{0}$, this coincides with the sample covariance.

Maximizing $l\left(\boldsymbol{H}, \boldsymbol{d}, \boldsymbol{\sigma}\right)$ under the constraints $\boldsymbol{H}^T \boldsymbol{I} = \boldsymbol{I}_F$, for fixed $\boldsymbol{\sigma}$ leads to the Misaligned Principal Component Analysis (MisPCA) solution:

$$\begin{aligned} \lambda^{\text{MisPCA}} &= \max \quad \sum_{f=1}^{F} \lambda_f\left(\boldsymbol{S}\left(\boldsymbol{\tau}\right)\right) \\ &\text{s.t.} \quad \boldsymbol{\tau} \in \{0, \cdots, d_{\max}\}^n, \end{aligned} \tag{5.7}$$

which consists of finding the alignment vector $\boldsymbol{\tau}$ that maximizes the first F eigenvalues of the aligned covariance $\boldsymbol{S}\left(\boldsymbol{\tau}\right)$. The optimal alignment is denoted by $\boldsymbol{d}^{\text{MisPCA}}$, and the corresponding MisPCA signal estimates are given by:

$$\boldsymbol{H}^{\text{MisPCA}} = \mathcal{V}_F\left(\boldsymbol{S}\left(\boldsymbol{d}^{\text{MisPCA}}\right)\right),$$

where $\mathcal{V}_F\left(\boldsymbol{X}\right)$ denotes the matrix constructed from the first F leading eigenvectors of a matrix $\boldsymbol{X}$. To estimate the $\boldsymbol{\sigma}$, it suffices to maximize $l\left(\boldsymbol{H}^{\text{MisPCA}}, \boldsymbol{d}^{\text{MisPCA}}, \boldsymbol{\sigma}\right)$ under the constraint $\sigma_f \geq 0$. The optimum occurs at (see Appendix D.1):

$$\sigma_f^{\text{MisPCA}} = \begin{cases} 0 & \text{if } \lambda_f^{\text{MisPCA}} < 1 \\ \lambda_f^{\text{MisPCA}} - 1 & \text{otherwise.} \end{cases} \tag{5.8}$$

5.3.1 PCA and Alternate MisPCA (A-MisPCA) approximations

Unfortunately, the MisPCA problem (5.7) is combinatorial, and exhaustive search is prohibitive even for small n. Here we consider two simple approximate solutions to (5.7). The first approximation ignores the misalignments altogether, i.e. solving

(5.7) with $\boldsymbol{d} = \mathbf{0}$. This leads to the usual PCA estimate of $\boldsymbol{H}$:

$$\boldsymbol{H}^{\text{PCA}} = \mathcal{V}_F\left(\boldsymbol{S}\left(\mathbf{0}\right)\right). \tag{5.9}$$

The second approximation, alternatively estimates $\boldsymbol{d}$ and $\boldsymbol{H}$. At each iteration $t > 1$, we compute:

$$\begin{aligned} \boldsymbol{d}_t^{\text{A-MisPCA}} &= \arg\max_{\boldsymbol{\tau}\in\{0,\cdots,d_{\max}\}^n} \operatorname{tr}\left(\boldsymbol{S}\left(\boldsymbol{\tau}\right)\boldsymbol{H}_{t-1}^{\text{A-MisPCA}}\right) \\ \boldsymbol{H}_t^{\text{A-MisPCA}} &= \mathcal{V}_F\left(\boldsymbol{S}\left(\boldsymbol{d}_t^{\text{A-MisPCA}}\right)\right) \end{aligned} \tag{5.10}$$

where we set $\boldsymbol{H}_0^{\text{A-MisPCA}}$ to an initial estimate of $\boldsymbol{H}$ and stop the algorithm when the change in likelihood is sufficiently small. We call this procedure Alternating MisPCA (A-MisPCA).

5.3.2 Sequential MisPCA (S-MisPCA)

The Alternating MisPCA described in the previous section updates the estimates $\boldsymbol{d}_t^{\text{A-MisPCA}}$ and $\boldsymbol{H}_t^{\text{A-MisPCA}}$ at each iteration based on knowledge of the entire batch of observations, with which we need to compute the misaligned covariance for each element in $\{0,\cdots,d_{\max}\}^n$. This might be computationally restrictive for very large n, and is not adapted to situations where we may receive the data sequentially such as real-time applications. In this section we propose a simple algorithm that sequentially aligns each new observation to the previous estimates and updates the estimates for $\boldsymbol{d}$ and $\boldsymbol{H}$ accordingly. At iteration t, the Sequential MisPCA (S-MisPCA) algorithm computes:

$$\begin{aligned} d_t^{\text{S-MisPCA}} &= \arg\max_{\tau\in\{0,\cdots,d_{\max}\}} \operatorname{tr}\left(\frac{t-1}{t}\left(\boldsymbol{S}^{t-1}\left(\boldsymbol{d}_{t-1}^{\text{S-MisPCA}}\right) + \frac{1}{t}\boldsymbol{C}_\tau^T\boldsymbol{x}_t\boldsymbol{x}_t^T\boldsymbol{C}_\tau\right)\boldsymbol{H}_{t-1}^{\text{S-MisPCA}}\right) \\ \boldsymbol{H}_t^{\text{S-MisPCA}} &= \mathcal{V}_F\left(\boldsymbol{S}^t\left(\boldsymbol{d}_t^{\text{S-MisPCA}}\right)\right) \end{aligned} \tag{5.11}$$

where $\boldsymbol{S}^{t-1}\left(\boldsymbol{d}_{t-1}^{\text{S-MisPCA}}\right)$ denotes the aligned covariance of the preceding $t-1$ observations:

$$\boldsymbol{S}^{t-1}\left(\boldsymbol{\tau}\right) = \frac{1}{t-1}\sum_{i=1}^{t-1}\boldsymbol{C}_{\tau_i}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{C}_{\tau_i}.$$

This algorithm has lower complexity than A-MisPCA, at the price of lower performance, as we will numerically study in Section 5.5.3.

5.4 Statistics of the misaligned covariance

The performance of the algorithms presented in the last section depend on the statistics of the leading eigenvalue and eigenvector of the random matrix $\boldsymbol{S}(\boldsymbol{\tau})$, for a fixed, deterministic $\boldsymbol{\tau}$. In this section, we use recent asymptotic results on the spectrum of large random matrices (Pau07; BGN11) to characterize the asymptotic behavior of $\lambda_1(\boldsymbol{S}(\boldsymbol{\tau}))$ and $\boldsymbol{v}_1(\boldsymbol{S}(\boldsymbol{\tau}))$.

Before we proceed to state our main result, we will need to define the following quantities. For any $\boldsymbol{t} \in \{0, \cdots, p-1\}^n$, define the function $\boldsymbol{s}(\boldsymbol{t}) : \{0, \cdots, p-1\}^n \to \{0, \cdots, n\}^p$, with coordinates given by:

$$s_i(\boldsymbol{t}) = \frac{|\{j \in \{1, \cdots, n\} \ : \ t_j = i-1\}|}{n} \tag{5.12}$$

where $|S|$ denotes the cardinality of a set S. (One can interpret $\boldsymbol{s}(\boldsymbol{t}) = [s_1(\boldsymbol{t}), \cdots, s_p(\boldsymbol{t})]$ as a histogram of the values in $\boldsymbol{t}$.) In addition, for any $\boldsymbol{H} \in \mathbb{R}^{p \times F}$, we define the following $Fp \times Fp$ symmetric, block-Toeplitz matrix $\boldsymbol{R}_H$, with elements:

$$[\boldsymbol{R}_H]_{Fk+i,Fl+j} = \boldsymbol{H}_k^T \boldsymbol{C}_{i-j} \boldsymbol{H}_l \quad \begin{cases} 1 \leq i,j \leq F \\ 0 \leq k,l \leq p \end{cases} \tag{5.13}$$

This matrix specializes to the autocorrelation matrix of $\boldsymbol{H}$, $\boldsymbol{R}_h$, when $F = 1$. For $F > 1$, $\boldsymbol{R}_H$ can be interpreted as a multi-dimensional autocorrelation matrix.

Under the assumptions of Section 5.2, is is easy to show that the expected value of the matrix $\boldsymbol{S}(\boldsymbol{\tau})$ is given by:

$$\boldsymbol{\Sigma}(\boldsymbol{\tau}) := E[\boldsymbol{S}(\boldsymbol{\tau})] = \mathrm{SNR}\boldsymbol{\mathcal{H}}\mathrm{diag}(\boldsymbol{s}(\boldsymbol{d}-_p\boldsymbol{\tau}) \otimes \bar{\boldsymbol{\sigma}})\boldsymbol{\mathcal{H}}^T + \boldsymbol{I}_p,$$

and we call this matrix the *aligned population covariance*, where

$$\boldsymbol{\mathcal{H}} = \begin{bmatrix} \boldsymbol{H} & \boldsymbol{C}_1\boldsymbol{H} & \cdots & \boldsymbol{C}_{p-1}\boldsymbol{H} \end{bmatrix},$$

$\boldsymbol{d}$ denotes the true alignment parameter with which the data was generated, and $-_p$ indicates a modulo p subtraction.

In the classical fixed p, large n setting, it is known that $\boldsymbol{S}(\boldsymbol{\tau})$ converges to $\boldsymbol{\Sigma}(\boldsymbol{\tau})$, and hence so does its corresponding eigenstructure. In this section, we consider the following high-dimensional setting, where p is allowed to grow with n:

(A1) The number of variables $p = p_n$ grows linearly with the number of samples n,

and as n tends to infinity,

$$\lim_{n\to\infty} \frac{p_n}{n} = c > 0. \tag{5.14}$$

Note that this includes the possibility of p_n being larger than the number of observations n.

(A2) The observations $\boldsymbol{x}_i$, $i = 1, \cdots, n$, are Gaussian and obey model (5.5), with parameters SNR, $\bar{\boldsymbol{\sigma}}$, $\boldsymbol{H}$ and $\boldsymbol{d}$, defined in Section 5.2.

(A3) The first F eigenvalues of the aligned population covariance, $\boldsymbol{\Sigma}(\boldsymbol{\tau})$, defined in (5.14), have multiplicity one.

The following result shows that the leading eigenpairs of $\boldsymbol{S}(\boldsymbol{\tau})$ matches those of $\boldsymbol{\Sigma}(\boldsymbol{\tau})$ only if the SNR is higher than a phase transition SNR which depends on the unknown parameters of the model, $\boldsymbol{H}$ and $\boldsymbol{d}$.

Theorem V.1. *Let $\boldsymbol{\tau} \in \{0, \cdots, d_{\max}\}^n$, and $\boldsymbol{S}(\boldsymbol{\tau})$, $\boldsymbol{\Sigma}(\boldsymbol{\tau})$ be the $p_n \times p_n$ aligned sample and population covariance matrices evaluated at $\boldsymbol{\tau}$, defined in (5.6) and (5.14), respectively. Then, under assumptions (A1)-(A3), as $n \to \infty$ we have:*

$$\lambda_f\left(\boldsymbol{S}(\boldsymbol{\tau})\right) \overset{a.s.}{\to} \begin{cases} (SNR\gamma_f + 1)\left(1 + \frac{c}{SNR\gamma_f}\right) & SNR > \frac{\sqrt{c}}{\gamma_f} \\ (1+\sqrt{c})^2 & otherwise, \end{cases}$$

and:

$$\left|\left\langle \boldsymbol{v}_f\left(\boldsymbol{S}(\boldsymbol{\tau})\right), \boldsymbol{v}_f\left(\boldsymbol{\Sigma}(\boldsymbol{\tau})\right)\right\rangle\right|^2 \overset{a.s.}{\to} \begin{cases} \frac{(SNR\gamma_f)^2 - c}{(SNR\gamma_f)^2 + cSNR\gamma_f} & SNR > \frac{\sqrt{c}}{\gamma_f} \\ 0 & otherwise, \end{cases}$$

where $\overset{a.s.}{\to}$ denotes almost sure convergence, and c is defined in (5.14). Here, γ_f is the gain/loss affecting the f-th eigenvector of $\boldsymbol{S}(\boldsymbol{\tau})$ due to $\boldsymbol{\tau}$ being different from $\boldsymbol{d}$, and is given by:

$$\gamma_f = \lambda_f\left(diag\left(\boldsymbol{s}\left(\boldsymbol{d} -_p \boldsymbol{\tau}\right) \otimes \bar{\boldsymbol{\sigma}}\right)^{\frac{1}{2}} \boldsymbol{R}_H\, diag\left(\boldsymbol{s}\left(\boldsymbol{d} -_p \boldsymbol{\tau}\right) \otimes \bar{\boldsymbol{\sigma}}\right)^{\frac{1}{2}}\right), \tag{5.15}$$

where $\boldsymbol{s}(\boldsymbol{t})$ and $\boldsymbol{R}_H$ are defined in (5.12) and (5.13), respectively.

See Appendix D.2 for a proof. This result is better understood graphically. Figures 5.1 and 5.2 show the average $\left|\left\langle \boldsymbol{v}_f\left(\boldsymbol{S}(\boldsymbol{0})\right), \boldsymbol{w}_f\right\rangle\right|^2$ computed over 50 random realizations

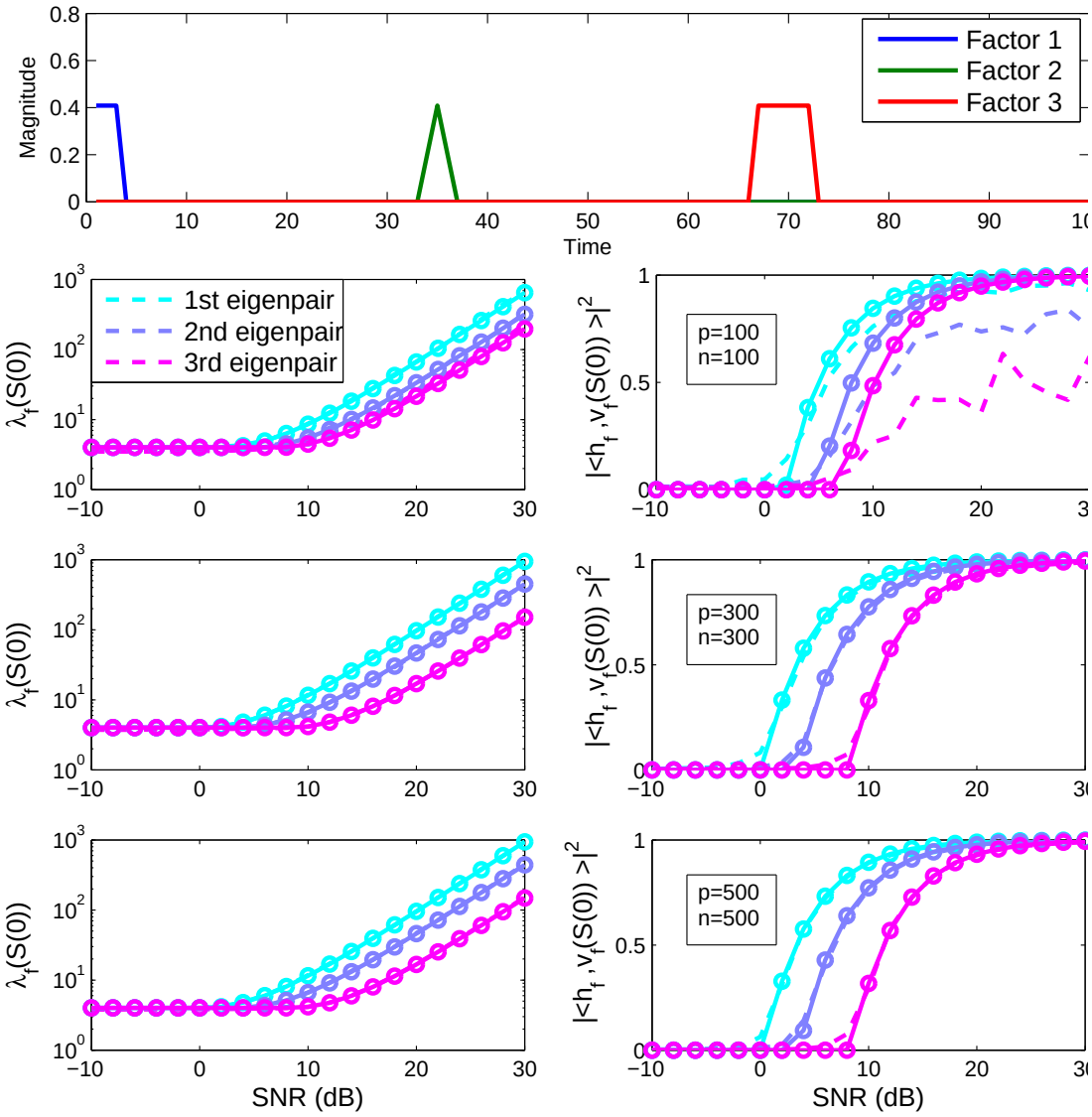


Figure 5.1: Predicted and average values of $\lambda_f(\boldsymbol{S}(\mathbf{0}))$ and $|\langle \boldsymbol{v}_f(\boldsymbol{S}(\mathbf{0})), \boldsymbol{v}_f(\boldsymbol{\Sigma}(\mathbf{0}))\rangle|^2$, $f = 1, \cdots, 3$, for $\boldsymbol{H} \in \mathbb{R}^{p\times 3}$ equal to three orthogonal pulses with narrow support (their support is much smaller than the dimension of the signal), shown in the top panel. The predictions of Theorem V.1 are shown in solid lines, the empirical average obtained over 50 random realizations are shown dashed. As p and n increase, the empirical results get closer to the predicted values. Notice that in this experiment the first three eigenvalues of the population covariance are close to each other, rendering the estimation of the corresponding eigenvectors harder. Figure 5.2 shows the results of the same experiment with pulses of larger width.

generated with model (5.1) for $\boldsymbol{H} \in \mathbb{R}^{p\times 3}$, $n = p$ samples and two choices of $\boldsymbol{H}$. Notice that the empirical results accurately match the asymptotic theory.

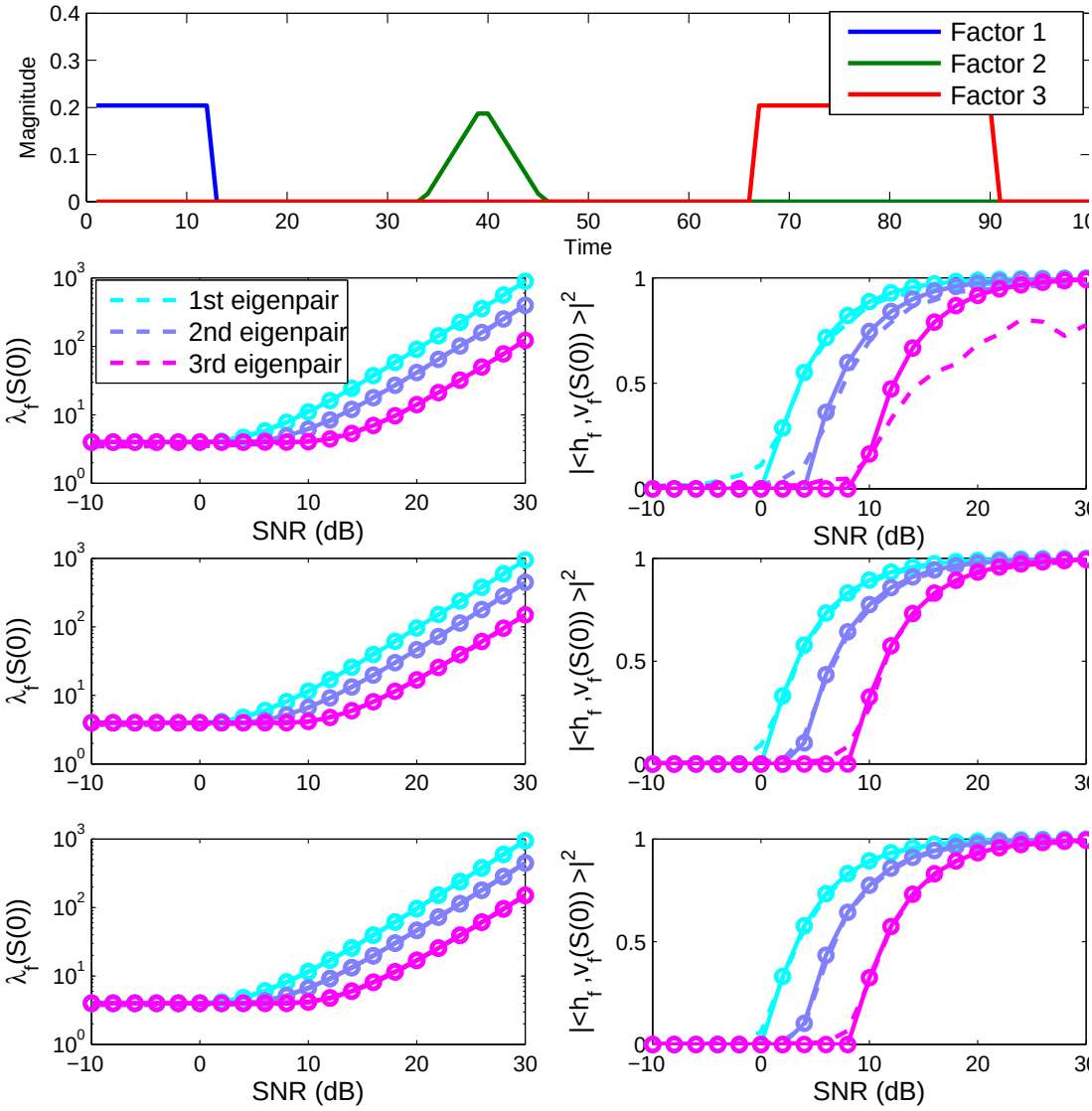


Figure 5.2: Same as in Figure 5.1 for $\boldsymbol{H} \in \mathbb{R}^{p\times 3}$ equal to three orthogonal pulses with large support (their support is in the order of the dimension of the signal). Notice that in this case the eigenvalues of the population covariance are more spaced than in the results of Figure 5.1, as reflected by the distance between the phase transition points of each eigenpair, and the convergence of the empirical results to the predictions is faster.

Theorem V.1 determines a "no-hope" regime for PCA and MisPCA. Consider for instance the PCA estimate, where $\boldsymbol{\tau} = \mathbf{0}$. Then Theorem V.1 implies that if the SNR

is smaller than

$$\frac{1}{\lambda_f\left(\operatorname{diag}\left(\boldsymbol{s}\left(\boldsymbol{d}\right)\otimes\bar{\boldsymbol{\sigma}}\right)^{\frac{1}{2}}\boldsymbol{R}_H\operatorname{diag}\left(\boldsymbol{s}\left(\boldsymbol{d}\right)\otimes\bar{\boldsymbol{\sigma}}\right)^{\frac{1}{2}}\right)}\sqrt{c}, \tag{5.16}$$

then each of the PCA estimates, defined in (5.9), is orthogonal to the corresponding eigenvector of $\boldsymbol{\Sigma}\left(\mathbf{0}\right)$, which contains partial information about the underlying signal $\boldsymbol{H}$. The scalar accompanying $\sqrt{c}$ in (5.16) can be interpreted as a tradeoff between the magnitude of the misalignments and the smoothness of the signal $\boldsymbol{H}$, as we will explore in the following sections.

More generally, if $SNR \leq \frac{\sqrt{c}}{\gamma_F}$ for any $\boldsymbol{\tau} \in \{0,\cdots,d_{\max}\}^n$, then the first part of Theorem V.1 asserts that the MisPCA objective in (5.7) is almost surely uninformative:

$$\sum_{f=1}^{F}\lambda_f\left(\boldsymbol{S}\left(\boldsymbol{\tau}\right)\right)\overset{\text{a.s.}}{\to}\left(1+\sqrt{c}\right)^2F \quad \text{as } n\to\infty,$$

and hence there is little hope for recovering $\boldsymbol{d}$ and $\boldsymbol{H}$.

In order to apply Theorem V.1 in a practical scenario, one would need to know both $\boldsymbol{H}$ and $\boldsymbol{d}$ beforehand, to subsequently compute the gain/losses due to misaligments, γ_f, and determine minimum operating SNR at which estimation is possible. In the following sections we give results that will allow us to characterize γ_f for each f, from only partial information about $\boldsymbol{H}$ and $\boldsymbol{d}$.

5.4.1 PCA under equispaced, deterministic misalignments

We consider first the simplest situation where we take the PCA estimate, $\boldsymbol{\tau}=\mathbf{0}$, and the misalignments $\boldsymbol{d}$ are deterministic and equispaced over $\{0,\cdots,d_{\max}\}$, in the following sense: For each $k\in\{0,\cdots,d_{\max}\}$, there are exactly $\frac{n}{d_{\max}+1}$ observations such that $d_i=k$. This implies that:

$$s_i\left(\boldsymbol{d}\right)=\begin{cases}\frac{1}{d_{\max}+1} & \text{if } 0\leq i\leq d_{\max}\\ 0 & \text{otherwise.}\end{cases} \tag{5.17}$$

Hence in such case Theorem V.1 implies that the "no-hope" regime includes all SNR smaller than

$$\frac{d_{\max}+1}{\lambda_1\left(\left[\boldsymbol{R}_H\right]_{\Omega,\Omega}\right)}\sqrt{c},\ \text{with } \Omega=\{1,\cdots,d_{\max}+1\}. \tag{5.18}$$

Of special interest is the case where $F = 1$, i.e. we have only a rank-1 signal. In such case $[\boldsymbol{R}_H]_{\Omega,\Omega}$ corresponds to a principal submatrix of the autocorrelation matrix of $\boldsymbol{H}$, which is a Toeplitz symmetric matrix parameterized by the autocorrelation sequence of $\boldsymbol{H}$, denoted by $r_h[k]$ and defined as:

$$r_h[k] = \boldsymbol{H}^T \boldsymbol{C}_k \boldsymbol{H}, \quad 0 \leq k \leq p-1. \tag{5.19}$$

The following result, which is an application of a result in (Fer92) to bound the eigenvalues of Toeplitz symmetric matrices, characterizes $\lambda_1\left([\boldsymbol{R}_h]_{\Omega,\Omega}\right)$ as a function of the Discrete Fourier spectrum of $r_h[k]$.

Theorem V.2. *Let $\boldsymbol{R}_h$ be the $(d_{\max}+1) \times (d_{\max}+1)$ autocorrelation matrix of a vector $\boldsymbol{h}$,*

$$[\boldsymbol{R}_h]_{i,j} = \boldsymbol{h}^T \boldsymbol{C}_{i-j} \boldsymbol{h} = r_h[|i-j|] \quad 0 \leq i,j \leq d_{\max}.$$

where $r_h[k]$ is its autocorrelation sequence, defined in (5.19). Then,

$$\omega_i + \underline{\delta} - 1 \leq \lambda_i(\boldsymbol{R}_h) \leq \omega_i + \bar{\delta} - 1 \tag{5.20}$$

where

$$\begin{aligned} \omega_i &= i\text{-}th\ [\mathfrak{Re}\ \hat{r}_h[0], \cdots, \mathfrak{Re}\ \hat{r}_h[d_{\max}+1]] \\ \underline{\delta} &= \min_k \mathfrak{Re}\ \hat{r}_h\left[k+\frac{1}{2}\right] \\ \bar{\delta} &= \max_k \mathfrak{Re}\ \hat{r}_h\left[k+\frac{1}{2}\right] \end{aligned}$$

where i-th $\boldsymbol{x}$ denotes the operator that returns the i-th largest element of a real vector $\boldsymbol{x}$, $\mathfrak{Re}\ x$ denotes the real part of a complex number x and $\hat{r}_h[k]$ is the Discrete Fourier Transform of the autocorrelation sequence $r_h[k]$:

$$\hat{r}_h[k] = \sum_{i=0}^{d_{\max}} r_h[i]\, e^{j\frac{2\pi i}{d_{\max}+1}k}$$

Proof. It is clear that $\boldsymbol{R}_h$ is a Toeplitz symmetric matrix, since $[\boldsymbol{R}_h]_{i,j} = [\boldsymbol{R}_h]_{j,i} = [\boldsymbol{R}_h]_{i+1,j+1} = r_h[|i-j|]$. We can hence use the results in (Fer92), which bound the eigenvalues of a Toeplitz symmetric matrix with functions of the eigenvalues of a circular matrix in which the original Toeplitz matrix can be embedded. Let

$\boldsymbol{c} \in \mathbb{R}^{2(d_{\max}+1)}$ be the vector parameterizing a $2(d_{\max}+1) \times 2(d_{\max}+1)$ circular matrix $\boldsymbol{C}$, defined as:

$$[\boldsymbol{C}]_{i,j} = c_{i-_{2(d_{\max}+1)}j}$$

where $-_p$ denotes a modulo $2(d_{\max}+1)$ subtraction. It is well-known that the eigenvalues of $\boldsymbol{C}$, denoted here by μ_i are given by the discrete Fourier transform of the sequence $\{c_i\}_{i=1}^{2(d_{\max}+1)}$:

$$\mu_k = \sum_{i=1}^{2(d_{\max}+1)} c_i e^{\frac{j2\pi(i-1)}{2(d_{\max}+1)}k}. \tag{5.21}$$

Notice that here the μ_i's are not necessarily sorted in descending order. Choose:

$$\boldsymbol{c} = [r_h[0], \cdots, r_h[d_{\max}], 0, r_h[d_{\max}], \cdots, r_h[1]] \tag{5.22}$$

and observe that the matrix $\boldsymbol{R}_h$ can be embedded in the circulant matrix $\boldsymbol{C}$ as follows:

$$[\boldsymbol{C}]_{\Omega,\Omega} = \boldsymbol{R}_h, \quad \text{with } \Omega = \{1, \cdots, d_{\max}+1\}.$$

Exploiting the properties of this embedding, the author in (Fer92) obtains the following bounds:

$$\frac{1}{2}\left[i\text{-th } \left[\mu_0, \mu_2, \cdots, \mu_{2(d_{\max}+1)}\right] + \min_k \mu_{2k+1}\right] \leq \lambda_i(\boldsymbol{R}_h) \tag{5.23}$$

$$\lambda_i(\boldsymbol{R}_h) \leq \frac{1}{2}\left[i\text{-th } \left[\mu_0, \mu_2, \cdots, \mu_{2(d_{\max}+1)}\right] + \max_k \mu_{2k+1}\right]. \tag{5.24}$$

The rest of our effort will be devoted to developing an expression of μ_{2k} and μ_{2k+1} in terms of the DFT of the sequence $r_h[k]$. From (5.21) and (5.22):

$$\begin{aligned}
\mu_k &= \sum_{i=1}^{d_{\max}+1} r_h[i-1] e^{\frac{j2\pi(i-1)}{2(d_{\max}+1)}k} + \sum_{i=d_{\max}+3}^{2(d_{\max}+1)} r_h[2(d_{\max}+1)-i+1] e^{\frac{j2\pi(i-1)}{2(d_{\max}+1)}k} \\
&= \hat{r}_h\left[\frac{k}{2}\right] + e^{j2\pi k} \sum_{t=2}^{d_{\max}+1} r_h[t-1] e^{\frac{-j2\pi(t-1)}{2(d_{\max}+1)}k} \\
&= \hat{r}_h\left[\frac{k}{2}\right] + \hat{r}_h\left[-\frac{k}{2}\right] - 1
\end{aligned}$$

where we have used the fact that $r_h[0] = 1$, and $\hat{r}_h[k]$ denotes the $d_{\max}+1$-points

DFT of r_h evaluated at k:

$$\hat{r}_h\,[k] = \sum_{i=1}^{d_{\max}+1} r_h[i-1] e^{\frac{j2\pi(i-1)}{d_{\max}+1}k}.$$

By properties of the DFT, we can conclude that:

$$\begin{aligned} \mu_{2k} &= 2\mathfrak{Re}\ \hat{r}_h\,[k] - 1 \\ \mu_{2k+1} &= 2\mathfrak{Re}\ \hat{r}_h \left[k + \frac{1}{2}\right] - 1 \end{aligned}$$

Combining these expressions with the bounds in (5.23) yields (5.20). □

Figure 5.3 shows the application of this result to bound the eigenvalues of two signals of dimension 20, a rectangular and a triangular signals of increasing width, denoted by W, $1 \leq W \leq d_{\max}+1$, with $d_{\max} = 10$.

Theorem V.2 seems to require knowledge of the spectrum of the autocorrelation sequence of $h[k]$ restricted to lags smaller or equal than $d_{\max}$. This demands less a priori information than knowing the signal itself, however, ideally, one would like to derive bounds on the eigenvalues of $\boldsymbol{R}_h$ that depend solely on fewer parameters of the signal. The answer to this question is affirmative, at least for a subset of signals, as the following example illustrates.

Consider a rectangular signal of width $1 < W < d_{\max}+1$:

$$\Pi_i = \begin{cases} \frac{1}{\sqrt{W}} & \text{if } 1 \leq i \leq W \\ 0 & \text{if } W \leq i \leq p \end{cases} \tag{5.25}$$

The corresponding autocorrelation function is given by:

$$r_{\Pi}[i] = \begin{cases} 1 - \frac{i}{W} & \text{if } 0 \leq i \leq W \\ 0 & \text{otherwise.} \end{cases}$$

and its $(d_{\max}+1)$-points DFT is:

$$\hat{r}_{\Pi}[k] = \sum_{i=0}^{W} \left(1 - \frac{i}{W}\right) e^{\frac{j2\pi ik}{d_{\max}+1}}$$

Using the formulae $\sum_{i=0}^{Q} \rho^i = \frac{1-\rho^{Q+1}}{1-\rho}$ and $\sum_{i=0}^{Q} i\rho^i = \frac{\rho-(Q+1)\rho^{Q+1}+Q\rho^{Q+2}}{(1-\rho)^2}$‘, we obtain

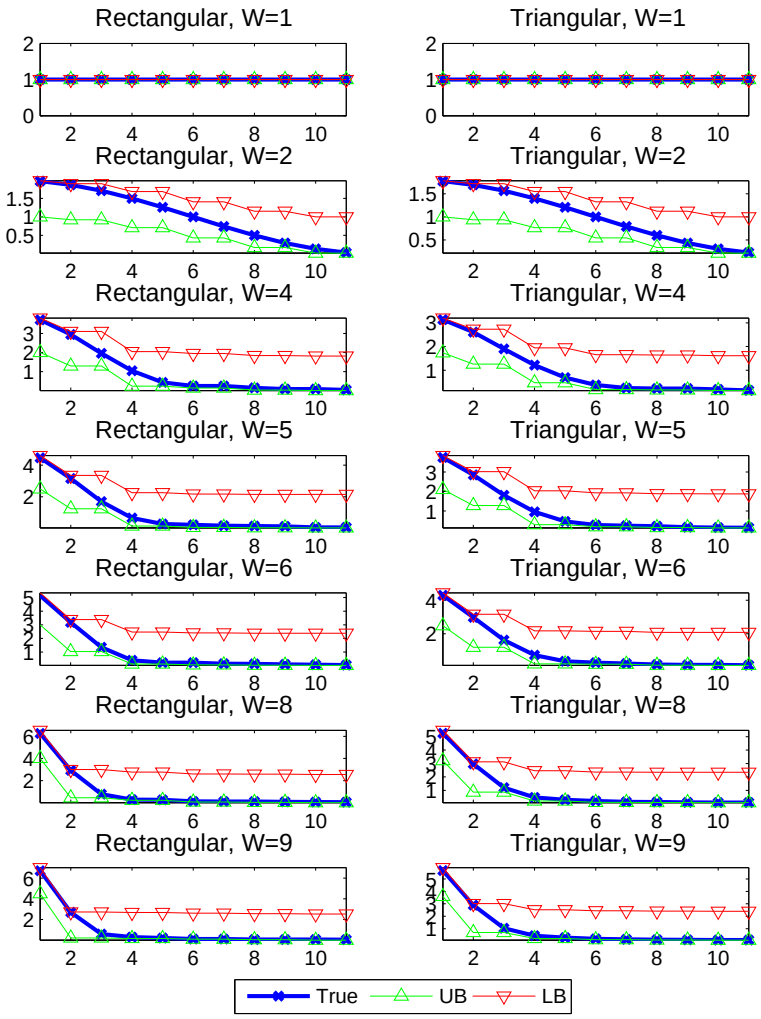


Figure 5.3: Eigenvalues for the autocorrelation matrix $\boldsymbol{R}_h$ for two 20-dimensional signals: a rectangular and a triangular signal of increasing width, denoted by W, and $d_{\max} = 10$. The upper and lower bounds for each eigenvalue are computed using Theorem V.2.

that:

$$\sum_{i=0}^{Q} \rho^i - \frac{1}{Q} \sum_{i=0}^{Q} i\rho^i \quad = \quad \frac{Q - (Q+1)\rho + \rho^{Q+1}}{Q\,(1-\rho)^2}$$

Furthermore, for a complex ρ with unit magnitude, we have:

$$\begin{aligned}\sum_{i=0}^{Q}\rho^i - \frac{1}{Q}\sum_{i=0}^{Q} i\rho^i &= \frac{Q-(Q+1)\rho+\rho^{Q+1}}{2Q\rho\left(\mathfrak{Re}\rho-1\right)} \\ &= \frac{Q\rho^{-1}-(Q+1)+\rho^{Q}}{2Q\left(\mathfrak{Re}\rho-1\right)}\end{aligned}$$

Applying the above expression with $Q=W$ and $\rho = e^{\frac{j2\pi k}{d_{\max}+1}}$ yields:

$$\hat{r}_{\Pi}[k] = \begin{cases} \frac{-(W+1)+We^{\frac{-j2\pi k}{d_{\max}+1}}+e^{\frac{j2\pi k}{d_{\max}+1}W}}{2W\left(\cos\left(\frac{2\pi k}{d_{\max}+1}\right)-1\right)} & \text{if } 1\le k\le d_{\max}+1 \\ \frac{W+1}{2} & \text{if } k=0 \end{cases}.$$

Taking the real part:

$$\mathfrak{Re}\ \hat{r}_{\Pi}[k] = \begin{cases} \frac{-(W+1)+W\cos\left(\frac{2\pi k}{d_{\max}+1}\right)+\cos\left(\frac{2\pi k}{d_{\max}+1}W\right)}{2W\left(\cos\left(\frac{2\pi k}{d_{\max}+1}\right)-1\right)} & \text{if } 1\le k\le d_{\max}+1 \\ \frac{W+1}{2} & \text{if } k=0 \end{cases}.$$

It is easy to check that $\mathfrak{Re}\ \hat{r}_{\Pi}[k]\ge 0$ and that:

$$\omega_1 = \max_k \mathfrak{Re}\ \hat{r}_{\Pi}[k] = \frac{W+1}{2},$$

which allow us to apply Theorem V.2 to obtain:

$$\frac{W-1}{2} \le \lambda_i\left(\boldsymbol{R}_{\Pi}\right).$$

Finally we combine this result with (5.18) to upper bound the "no-hope" SNR:

$$\frac{d_{\max}+1}{\lambda_1\left(\left[\boldsymbol{R}_H\right]_{\Omega,\Omega}\right)}\sqrt{c} \le \frac{2\left(d_{\max}+1\right)}{W-1}\sqrt{c}. \tag{5.26}$$

The right hand side of the above equation is an overestimator of the "no-hope" SNR below which rectangular signals of width W become undetectable. This has a direct application in certain multi-path communications environment, where the normalized channel impulse response, denoted by $c[k]$, can be well approximated as:

$$c[k] = \sum_{i=0}^{d_{\max}} a_i\delta[k-i]$$

with a_i independent, normal random amplitudes. In this setting, equation (5.26) gives a sufficient minimum SNR at which the signal needs to be transmitted in order to be able to recover it from the covariance matrix. Finally, observe that the "no-hope" SNR exhibits a tradeoff between the amplitude of the misalignments and the width of the signal, as the intuition might suggest.

5.4.2 PCA under random misalignments of small magnitude

We consider here the case where each element of the vector $\boldsymbol{d}$ is drawn independently from a uniform distribution over $\{0, \cdots, d_{\max}\}$. The "small magnitude" assumption here refers to the situation where $d_{\max} \ll p_n$ and grows very slowly with p_n.

First, we will characterize the asymptotic behavior of $\boldsymbol{s}(\boldsymbol{d})$. The i-th element of $\boldsymbol{s}(\boldsymbol{d})$ is given by

$$s_i(\boldsymbol{d}) = \begin{cases} \frac{1}{n}\sum_{j=1}^{n} I(d_j, i) & \text{if } 0 \leq i \leq d_{\max} \\ 0 & \text{if } d_{\max}+1 \leq i \leq p \end{cases},$$

where $I(x, y)$ is an indicator function that returns 1 if $x = y$ and 0 otherwise. Since the d_j's are drawn independently from one another and $\mathrm{Var}(I(d_j, i)) = \frac{d_{\max}}{(d_{\max}+1)^2}$, a simple application of Chebyshev's inequality shows that, as $n \to \infty$,

$$s_i(\boldsymbol{d}) \to_p \frac{1}{d_{\max}+1}, \quad 0 \leq i \leq d_{\max},$$

where $\to_p$ denotes convergence in probability. We can turn now to the problem of estimating $\lambda_1\left(\mathrm{diag}\,\boldsymbol{s}(\boldsymbol{d})^{\frac{1}{2}}\boldsymbol{R}_H\mathrm{diag}\,\boldsymbol{s}(\boldsymbol{d})^{\frac{1}{2}}\right)$ as $n \to \infty$. The function $\boldsymbol{x} \in \mathbb{R}_+^2 \to [\boldsymbol{R}_H]_{i,j}\sqrt{x_1 x_2}$ is continuous in $\boldsymbol{x}$ for any i, j, hence we have, by Proposition 8.5 of (Kee10):

$$\mathrm{diag}\,\boldsymbol{s}(\boldsymbol{d})^{\frac{1}{2}}\boldsymbol{R}_H\mathrm{diag}\,\boldsymbol{s}(\boldsymbol{d})^{\frac{1}{2}} \to_p \frac{1}{d_{\max}+1}[\boldsymbol{R}_H]_{\Omega,\Omega}, \quad \text{as } n \to \infty, \tag{5.27}$$

with $\Omega = \{1, \cdots, d_{\max}+1\}$, under a Frobenius norm metric. Finally, it is a well known result of matrix analysis (see for instance, Appendix D in (HJ90)) that the eigenvalue function of a real symmetric matrix is a continuous function of its argument. Combining this fact with Proposition 8.5 of (Kee10) allows us to conclude that:

$$\lambda_1\left(\mathrm{diag}\,\boldsymbol{s}(\boldsymbol{d})^{\frac{1}{2}}\boldsymbol{R}_H\mathrm{diag}\,\boldsymbol{s}(\boldsymbol{d})^{\frac{1}{2}}\right) \to_p \frac{1}{d_{\max}+1}\lambda_1\left([\boldsymbol{R}_H]_{\Omega,\Omega}\right),$$

where $\Omega = \{1, \cdots, d_{\max} + 1\}$. This means that, as one would expect, the small magnitude, random misalignment case essentially reduces to the case studied in the previous section, Section 5.4.1.

5.4.3 Asymptotic bias of PCA under deterministic equispaced misalignments

We have shown in the last section that misalignments can have a negative effect at low SNR's, where they may render the signal undetectable at the same level of SNR where it would have been detectable if no misalignment was present. There is however another perverse effect of misaligments that does not disappear as the SNR or the sample size increases: the introduction of bias to the traditional PCA estimate.

In this section we will characterize the asymptotic bias in terms of the distance between the estimated and the original subspace, which is spanned by the matrix $\boldsymbol{H}$. A reasonable measure of distance between the subspaces spanned by two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is given by:

$$d(\boldsymbol{A}, \boldsymbol{B}) = \left\| \boldsymbol{A} \left(\boldsymbol{A}^T \boldsymbol{A} \right)^\dagger \boldsymbol{A}^T - \boldsymbol{B} \left(\boldsymbol{B}^T \boldsymbol{B} \right)^\dagger \boldsymbol{B}^T \right\|_F^2, \tag{5.28}$$

that is, the Frobenius distance between the projection matrices associated to each one of these matrices. Precisely, we will compute the asymptotic distance between the signal subspace spanned by $\boldsymbol{H}$ and the subspace spanned by the PCA estimate

$$\mathcal{V}_F(\boldsymbol{S}(\boldsymbol{0})) := [\boldsymbol{v}_1(\boldsymbol{S}(\boldsymbol{0})), \cdots, \boldsymbol{v}_F(\boldsymbol{S}(\boldsymbol{0}))], \tag{5.29}$$

as $n \to \infty$ and SNR $\to \infty$, for the deterministic equispaced misalignment model discussed in Section 5.4.1. A word of caution is required here since $\mathcal{V}_F(\boldsymbol{S}(\boldsymbol{0}))$ is only uniquely defined whenever the first F eigenvalues of $\boldsymbol{S}(\boldsymbol{0})$ counting multiplicities are strictly larger than the $p-F$ subsequent eigenvalues. If this is not the case, then there is more than one possible $\mathcal{V}_F(\boldsymbol{S}(\boldsymbol{0}))$. As an example, if $\lambda_1(\boldsymbol{S}(\boldsymbol{0}))$ is of multiplicity $M > F$, then there are $\binom{F}{M}$ possibilities to construct $\mathcal{V}_F(\boldsymbol{S}(\boldsymbol{0}))$ and hence one would need to define a criteria to decide which one of these possibilities is the right one to compare $\boldsymbol{H}$ to.

In order to simplify our development, we will assume in the sequel that $\bar{\boldsymbol{\sigma}} = \boldsymbol{1}$, that is, that all the signal components have the same energy. Similar results could be obtain with a generic $\bar{\boldsymbol{\sigma}}$, at the price of introducing additional notational burden.

First, we observe that, by Theorem V.1, as $n \to \infty$ and SNR $\to \infty$ we have:

$$\mathcal{V}_F\left(\boldsymbol{S}\left(\mathbf{0}\right)\right) \to \mathcal{V}_F\left(\boldsymbol{\Sigma}\left(\mathbf{0}\right)\right). \tag{5.30}$$

We will thus focus on the quantity $d\left(\boldsymbol{H}, \mathcal{V}_F\left(\boldsymbol{\Sigma}\left(\mathbf{0}\right)\right)\right)$, instead of its sample analog $d\left(\boldsymbol{H}, \mathcal{V}_F\left(\boldsymbol{S}\left(\mathbf{0}\right)\right)\right)$. Since both $\boldsymbol{H}$ and $\mathcal{V}_F\left(\boldsymbol{\Sigma}\left(\mathbf{0}\right)\right)$ are unitary matrices, this quantity simplifies to:

$$\begin{aligned} d\left(\boldsymbol{H}, \mathcal{V}_F\left(\boldsymbol{\Sigma}\right)\right) &= \left\| \boldsymbol{H}\boldsymbol{H}^T - \mathcal{V}_F\left(\boldsymbol{\Sigma}\right)\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)^T \right\|_F^2 \\ &= 2\left(F - \mathrm{tr}\left(\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)^T \boldsymbol{H}\boldsymbol{H}^T\right)\right), \end{aligned} \tag{5.31}$$

where we have denoted $\boldsymbol{\Sigma}\left(\mathbf{0}\right)$ by $\boldsymbol{\Sigma}$ in order to alleviate the notation. It is clear from the equation above that if

$$\boldsymbol{H}^T\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)^T\boldsymbol{H} = \boldsymbol{I}_F \tag{5.32}$$

then $d\left(\boldsymbol{H}, \mathcal{V}_F\left(\boldsymbol{\Sigma}\right)\right) = 0$. The following result shows that this is also a necessary condition, and gives an expression for $d\left(\boldsymbol{H}, \mathcal{V}_F\left(\boldsymbol{\Sigma}\right)\right)$ that depends exclusively on the spectrum of the multidimensional autocorrelation matrix of $\boldsymbol{H}$, $\boldsymbol{R}_H$, defined in (5.13).

Theorem V.3. *Assume that $\bar{\boldsymbol{\sigma}} = \mathbf{1}_{F\times 1}$ and that the misalignments are equispaced and deterministic, so that $\boldsymbol{s}\left(\boldsymbol{d}\right)$ is given by (5.17). Assume further that the first F leading eigenvalues of $\boldsymbol{\Sigma}\left(\mathbf{0}\right)$, counting multiplicities, are strictly larger than the subsequent $p - F$, so that $\mathcal{V}_F\left(\boldsymbol{\Sigma}\left(\mathbf{0}\right)\right)$ is uniquely defined. Then, the asymptotic bias between the subspace spanned by the PCA estimate and the original signal, as $n \to \infty$ and SNR $\to \infty$, is strictly positive and given by:*

$$d\left(\boldsymbol{H}, \mathcal{V}_F\left(\boldsymbol{\Sigma}\left(\mathbf{0}\right)\right)\right) = 2\left(F - tr\left(\mathcal{V}_H\Lambda_F\left(\tilde{\boldsymbol{R}}_H\right)\mathcal{V}_H^T\right)\right) > 0$$

unless

$$\mathcal{V}_H\Lambda_F\left(\tilde{\boldsymbol{R}}_H\right)\mathcal{V}_H^T = \boldsymbol{I}_F, \tag{5.33}$$

where $\mathcal{V}_H = \left[\mathcal{V}_F\left(\tilde{\boldsymbol{R}}_H\right)\right]_{\{1,\cdots,F\},\cdot}$, and $\tilde{\boldsymbol{R}}_H$ denotes the $(d_{\max}+1)\times(d_{\max}+1)$ upper left principal submatrix of the autocorrelation matrix of $\boldsymbol{H}$, defined in (5.13).

Proof. First we observe that, since $\bar{\boldsymbol{\sigma}} = \mathbf{1}_{F\times 1}$ and $\boldsymbol{s}(\boldsymbol{d})$ is given by (5.17),

$$\begin{aligned}\boldsymbol{\Sigma} &= \text{SNR}\, \boldsymbol{\mathcal{H}}\text{diag}\left(\left[\mathbf{1}_{(d_{\max}+1)\times 1}, \mathbf{0}_{(p-d_{\max}-1)\times 1}\right] \otimes \mathbf{1}_{F\times 1}\right) \boldsymbol{\mathcal{H}}^T + \boldsymbol{I}_p, \\ &= \frac{\text{SNR}}{d_{\max}+1}\, \tilde{\boldsymbol{\mathcal{H}}}\tilde{\boldsymbol{\mathcal{H}}}^T + \boldsymbol{I}_p,\end{aligned}$$

where

$$\tilde{\boldsymbol{\mathcal{H}}} = \left[\boldsymbol{H}, \boldsymbol{C}_1\boldsymbol{H}, \ \cdots \ , \boldsymbol{C}_{d_{\max}}\boldsymbol{H}\right].$$

Thus,

$$\mathcal{V}_F\left(\boldsymbol{\Sigma}\right) = \mathcal{V}_F\left(\tilde{\boldsymbol{\mathcal{H}}}\tilde{\boldsymbol{\mathcal{H}}}^T\right). \tag{5.34}$$

In addition, by definition, the eigenvectors of $\tilde{\boldsymbol{\mathcal{H}}}^T\tilde{\boldsymbol{\mathcal{H}}}$ are given by,

$$\tilde{\boldsymbol{\mathcal{H}}}^T\tilde{\boldsymbol{\mathcal{H}}}\boldsymbol{v}_i\left(\tilde{\boldsymbol{\mathcal{H}}}^T\tilde{\boldsymbol{\mathcal{H}}}\right) = \lambda_i\left(\tilde{\boldsymbol{\mathcal{H}}}^T\tilde{\boldsymbol{\mathcal{H}}}\right)\boldsymbol{v}_i\left(\tilde{\boldsymbol{\mathcal{H}}}^T\tilde{\boldsymbol{\mathcal{H}}}\right)$$

and consequently,

$$\tilde{\boldsymbol{\mathcal{H}}}\tilde{\boldsymbol{\mathcal{H}}}^T\left(\tilde{\boldsymbol{\mathcal{H}}}\boldsymbol{v}_i\left(\tilde{\boldsymbol{\mathcal{H}}}^T\tilde{\boldsymbol{\mathcal{H}}}\right)\right) = \lambda_i\left(\tilde{\boldsymbol{\mathcal{H}}}^T\tilde{\boldsymbol{\mathcal{H}}}\right)\left(\tilde{\boldsymbol{\mathcal{H}}}\boldsymbol{v}_i\left(\tilde{\boldsymbol{\mathcal{H}}}^T\tilde{\boldsymbol{\mathcal{H}}}\right)\right),$$

which shows that the right hand side is a multiple of an eigenvector of $\tilde{\boldsymbol{\mathcal{H}}}\tilde{\boldsymbol{\mathcal{H}}}^T$. Hence, by (5.34):

$$\mathcal{V}_F\left(\boldsymbol{\Sigma}\right) = \tilde{\boldsymbol{\mathcal{H}}}\mathcal{V}_F\left(\tilde{\boldsymbol{R}}_H\right)\Lambda_F^{-\frac{1}{2}}\left(\tilde{\boldsymbol{R}}_H\right) \tag{5.35}$$

where we define the autocorrelation matrix restricted to misalignments of magnitude smaller or equal than $d_{\max}$:

$$\tilde{\boldsymbol{R}}_H = \tilde{\boldsymbol{\mathcal{H}}}^T\tilde{\boldsymbol{\mathcal{H}}}. \tag{5.36}$$

Fan's inequality (BL06) asserts that, for any two $p \times p$ symmetric matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\text{tr}\left(\boldsymbol{X}\boldsymbol{Y}\right) \leq \sum_{i=1}^{p} \lambda_i\left(\boldsymbol{X}\right)\lambda_i\left(\boldsymbol{Y}\right) \tag{5.37}$$

with equality only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ have a simultaneous ordered spectral decomposition.

Since $\lambda_i\left(\boldsymbol{H}\boldsymbol{H}^T\right) = \lambda_i\left(\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)^T\right) = 1$ for $i = 1, \cdots, F$, and 0 for $F < i \leq p$, we can use (5.37), to establish:

$$\mathrm{tr}\left(\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)^T\boldsymbol{H}\boldsymbol{H}^T\right) \leq F \tag{5.38}$$

with equality if and only if there exists a symmetric, orthogonal matrix $\boldsymbol{Q}$ such that

$$\begin{aligned}\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)^T &= \boldsymbol{Q}\mathrm{diag}\left(\left[\mathbf{1}_{F\times 1}, \mathbf{0}_{p-F\times 1}\right]\right)\boldsymbol{Q}^T \\ \boldsymbol{H}\boldsymbol{H}^T &= \boldsymbol{Q}\mathrm{diag}\left(\left[\mathbf{1}_{F\times 1}, \mathbf{0}_{p-F\times 1}\right]\right)\boldsymbol{Q}^T\end{aligned}$$

It follows that, for any such $\boldsymbol{Q}$,

$$\boldsymbol{Q}^T\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)^T\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{H}\boldsymbol{H}^T\boldsymbol{Q}^T$$

which implies that $\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)^T = \boldsymbol{H}\boldsymbol{H}^T$ and that the equality in (5.38) is attained whenever:

$$\boldsymbol{H}^T\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)^T\boldsymbol{H} = \boldsymbol{I}.$$

Replacing the expression for $\mathcal{V}_F\left(\boldsymbol{\Sigma}\right)$ obtained in (5.35) leads to the condition:

$$\boldsymbol{H}^T\tilde{\boldsymbol{\mathcal{H}}}\mathcal{V}_F\left(\tilde{\boldsymbol{R}}_H\right)\Lambda_F^{-1}\left(\tilde{\boldsymbol{R}}_H\right)\mathcal{V}_F\left(\tilde{\boldsymbol{R}}_H\right)^T\tilde{\boldsymbol{\mathcal{H}}}^T\boldsymbol{H} = \boldsymbol{I} \tag{5.39}$$

Observe now that the matrix $\tilde{\boldsymbol{R}}_H$ defined in (5.36) is a block Toeplitz matrix parameterized by the $F \times F$ matrix function:

$$\boldsymbol{r}_H\left(d\right) := \boldsymbol{H}^T\boldsymbol{C}_d\boldsymbol{H}, \tag{5.40}$$

that is,

$$\left[\tilde{\boldsymbol{R}}_H\right]_{B_i,B_j} = \boldsymbol{r}_H\left(j-i\right), \text{ where } B_i = \left\{(i-1)F+1, \cdots, iF\right\}.$$

Note also that, by definition of $\boldsymbol{r}_H\left(d\right)$,

$$\begin{aligned}\tilde{\boldsymbol{\mathcal{H}}}^T\boldsymbol{H} &= \left[\boldsymbol{r}_H\left(0\right), \boldsymbol{r}_H\left(-1\right), \cdots, \boldsymbol{r}_H\left(-d_{\max}\right)\right]^T \\ &= \left[\boldsymbol{r}_H\left(0\right), \boldsymbol{r}_H\left(1\right)^T, \cdots, \boldsymbol{r}_H\left(d_{\max}\right)^T\right]^T.\end{aligned} \tag{5.41}$$

which happens to correspond to the first column of the multidimensional autocorre-

lation matrix $\tilde{\boldsymbol{R}}_H$. This has an interesting implication: the eigenvectors of a block-Toeplitz matrix $\tilde{\boldsymbol{R}}_H$ parameterized by a vector $[\boldsymbol{r}_H(0), \cdots, \boldsymbol{r}_H(d_{\max})]$ satisfy:

$$[\boldsymbol{r}_H(0), \cdots, \boldsymbol{r}_H(d_{\max})]\, \boldsymbol{v}_i\left(\tilde{\boldsymbol{R}}_H\right) = \lambda_i\left(\tilde{\boldsymbol{R}}_H\right)\left[\boldsymbol{v}_i\left(\tilde{\boldsymbol{R}}_H\right)\right]_{\mathcal{I}} \tag{5.42}$$

where $\mathcal{I} = \{1, \cdots, F\}$.

Thus condition (5.39) reduces to:

$$\left[\mathcal{V}_F\left(\tilde{\boldsymbol{R}}_H\right)\right]_{\mathcal{I},\cdot} \Lambda_F\left(\tilde{\boldsymbol{R}}_H\right)\left[\mathcal{V}_F\left(\tilde{\boldsymbol{R}}_H\right)\right]_{\mathcal{I},\cdot}^T = \boldsymbol{I}, \tag{5.43}$$

which proves (5.33). □

This result is perhaps surprising in that the asymptotic bias solely depends on the first F elements of the first F eigenvectors of the multi-dimensional correlation matrix of $\boldsymbol{H}$, denoted by $\boldsymbol{R}_H$. To illustrate its potential application, consider now the following special cases:

$$\boldsymbol{R}_H^1 = \mathbf{1}_{d_{\max}+1 \times d_{\max}+1} \otimes \boldsymbol{I}_F \quad \text{and} \quad \boldsymbol{R}_H^2 = \boldsymbol{I}_{d_{\max}+1} \otimes \boldsymbol{I}_F.$$

Using properties of the Kronecker product, we can assert that:

$$\begin{aligned} \lambda_i\left(\boldsymbol{R}_H^1\right) &= \begin{cases} d_{\max}+1, & i = 1, \cdots, F \\ 1 & F < i \leq F(d_{\max}+1) \end{cases} \\ \lambda_i\left(\boldsymbol{R}_H^2\right) &= 1, \end{aligned} \tag{5.44}$$

and:

$$\begin{aligned} \boldsymbol{v}_i\left(\boldsymbol{R}_H^1\right) &= \frac{1}{\sqrt{d_{\max}+1}} \mathbf{1}_{d_{\max}+1} \otimes \boldsymbol{e}_i^F, i = 1, \cdots, F \\ \boldsymbol{v}_i\left(\boldsymbol{R}_H^2\right) &= \boldsymbol{e}_i^{F(d_{\max}+1)}, i = 1, \cdots, F \end{aligned} \tag{5.45}$$

where $\boldsymbol{e}_i^p$ is the i-th canonical vector in $\mathbb{R}^p$. First, we notice that $\boldsymbol{R}_H^2$ does not verify the necessary condition to define $\mathcal{V}_F(\boldsymbol{\Sigma}^2)$ uniquely and hence Theorem V.3 does not apply. In fact, for $F = 1$ and a signal $\boldsymbol{H}^2$ with autocorrelation matrix equal to $\boldsymbol{R}_H^2$, one can show that $\mathcal{V}_1(\boldsymbol{\Sigma}^2)$ is given by any element of the set of circularly shifted $\boldsymbol{H}^2$. If we do not happen to choose the right shift, the bias will be non-null whereas wrongly applying Theorem V.3 would lead us to assert that it will be zero.

On the other hand, consider $\boldsymbol{H}^1 = \frac{1}{\sqrt{p}}\mathbf{1}_p$ which verifies $\boldsymbol{R}_H^1 = \mathbf{1}_{d_{\max}+1 \times d_{\max}+1}$. The constant signal $\boldsymbol{H}^1$ is invariant to circular shifts, and this intuition is confirmed by

the application of Theorem V.3 which asserts that the PCA estimate from misaligned copies of such signal will be asymptotically unbiased.

We will numerically investigate these asymptotic results for other choices of signals in Section 5.5.2.

5.5 Experiments

In this section, we present numerical results that illustrate the application of the theory developed in the preceding sections and study the performance of the MisPCA algorithms described in Section 5.3.

5.5.1 Phase transitions in misaligned signals

In this section we investigate the non-asymptotic accuracy of the phase transition predictions for the PCA estimate from Section 5.4.1 and 5.4.2, for the rank-1 signal case, $F = 1$.

We generate data according to model (5.1) with $p = 300$, $F = 1$, uniform and independently distributed misalignments $d_i \in \{0, \cdots, d_{\max}\}$ and varying n and $d_{\max}$. We experiment with three choices for $\boldsymbol{H}$, depicted in the top panel of Figure 5.4: (i) a rectangular signal with width $W = 30$ as defined in (5.25), (ii) the same signal convoluted with a triangular pulse of width 10 and (iii) a sum of two sinusoids with periods $T_1 = \frac{1}{2}$ and $T_1 = \frac{1}{3}$. In the first experiment, we fix $d_{\max} = 30$ and let the SNR vary between SNR $= -10$ and SNR $= 30$dB and and n between $n = 100$ and $n = 1000$. At each point (SNR, n), we compute the affinity between the PCA estimate $\boldsymbol{H}^{\text{PCA}}$ and the original signal $\boldsymbol{H}$, which we define as:

$$a\left(\boldsymbol{H}^{\text{PCA}}, \boldsymbol{H}\right) = \max_{j \in \{1, \cdots, p\}} \langle \boldsymbol{C}_j \boldsymbol{H}^{\text{PCA}}, \boldsymbol{H} \rangle^2. \tag{5.46}$$

Notice that this measure is invariant to translations of $\boldsymbol{H}$ and that it is a reciprocal of the Frobenius distance between the estimated and true signal:

$$\left|\left|\boldsymbol{H}^{\text{PCA}} - \boldsymbol{H}\right|\right|_F^2 \geq 2\left(1 - \sqrt{a\left(\boldsymbol{H}^{\text{PCA}}, \boldsymbol{H}\right)}\right).$$

The second experiment is identical to the first one except that now we fix $n = 100$, and vary SNR between SNR $= -10$ and SNR $= 30$dB and $d_{\max}$ between $d_{\max} = 2$ and $d_{\max} = 99$. The results of both experiments are shown in Figure 5.4. Superimposed to the heatmap of affinity, we plot the phase transition bounds obtained by the

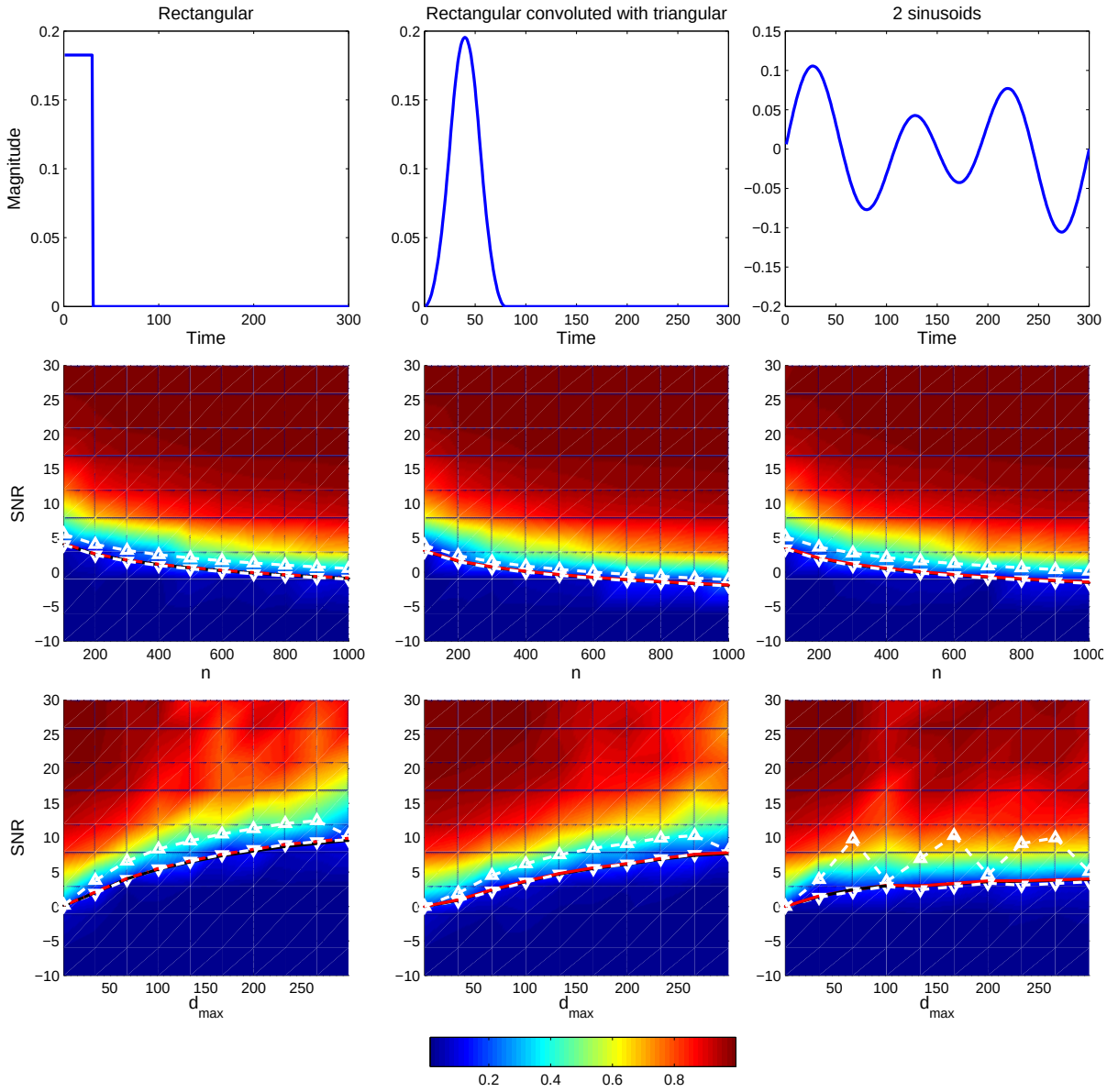


Figure 5.4: Heatmaps depicting the average value over 30 random realizations of the affinity $a\left(\boldsymbol{H}^{\mathrm{PCA}}, \cdot\right)$ between the PCA estimate and the true signal, as a function of $(\mathrm{SNR}, d_{\max})$ (middle panel) or (SNR, n) (bottom panel), for each of three rank-1 signals shown on the top panel. The red line corresponds to the computed phase transition SNR, given by Theorem V.1. The white dashed lines depict the upper and lower bounds obtained by combined application of Theorems V.1 and V.2.

direct application of Theorem V.1 (red), which uses knowledge of the signal $\boldsymbol{H}$ and the misalignments $\boldsymbol{d}$. The white dashed lines depict the upper and lower bounds

obtained by combined application of Theorems V.1 and V.2.

The results in Figure 5.4 shows how the asymptotic theory developed in Section 5.4 is of practical use at a non-asymptotic regime. As predicted by Theorem V.1 and Section 5.4.2, $a\left(\boldsymbol{H}^{\mathrm{PCA}}, \boldsymbol{H}\right)$ shows a clear phase transition frontier at

$$\frac{d_{\max}+1}{\lambda_1\left(\left[\boldsymbol{R}_H\right]_{\Omega,\Omega}\right)}\sqrt{c},$$

shown in solid red. Figure 5.4 also highlights the advantage of pooling misaligned observations: despite the misalignment, the phase transition point decreases significantly as n increases.

5.5.2 Asymptotic Bias predictions

In this section we study the asymptotic bias predictions developed in Section 5.4.3. Intuitively, one expects signals with larger temporal support to be less affected by misalignments than signals with support concentrated on a very small region of the temporal axis. In order to validate this intuition, we consider now three different signals of increasing temporal support, and generate data according to model (5.1) with $p = 300$, $F = 3$, $n = 500$ and uniform and independently distributed misalignments of magnitude smaller than $d_{\max}$. For each realization, we plot in Figure 5.5 the subspace distance between the PCA estimate and the true signal $\boldsymbol{H}$, defined in (5.31), as a function of the SNR level and the magnitude of the misalignments, $d_{\max}$. We also plot superimposed the sample average of these distances (solid blue) and the predicted bias (red).

It is clear that as the SNR increases, the empirical results accurately match the predictions, even for the relatively small $n = 500$. In addition, Figure 5.5 shows that the sensitivity to misalignments is much less pronounced for signals of larger support (rightmost plots) compared to the small support signals on the left.

5.5.3 Numerical comparison of MisPCA Algorithms

In this section we compare the various MisPCA algorithms described in Section II. As a reference benchmark, we compute the Oracle-PCA, which assumes knowledge of $\boldsymbol{d}$ and consists of performing PCA on $\boldsymbol{S}(\boldsymbol{d})$.

In order to compare the performance of each method, we estimate the minimum SNR needed for each algorithm to attain a certain level of fidelity ρ with respect to

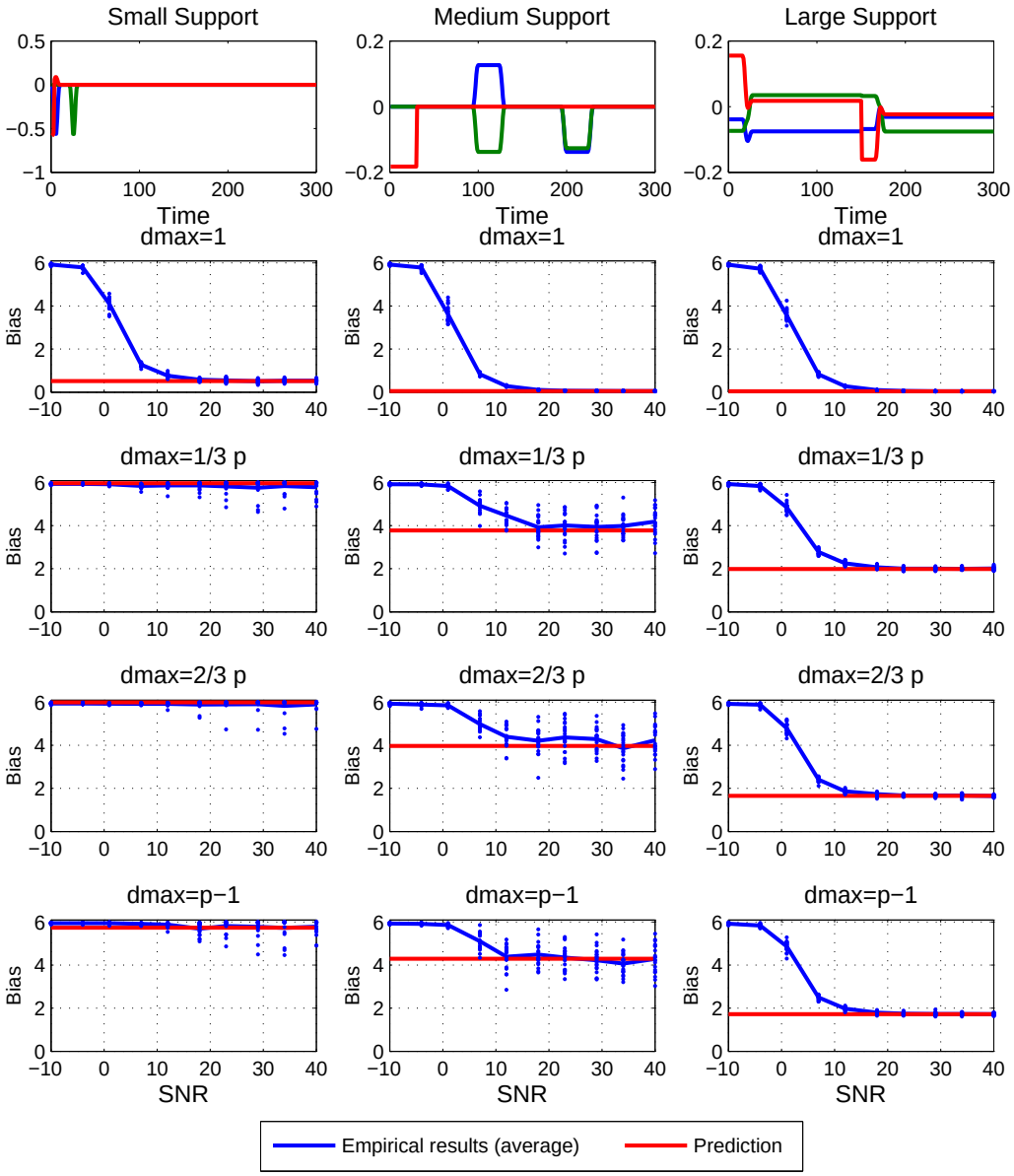


Figure 5.5: Asymptotic bias for the PCA estimate of a misaligned rank 3 signal of dimension $p = 300$ with uniformly distributed misalignments of increasing magnitude. We consider three signals, depicted in the upper panels, with increasingly larger support. Our asymptotic predictions demonstrate the intuitive fact that signals with wider support are more robust to misalignments: the bias for the signal on the rightmost plots is about one third of the bias for the narrow signal on the leftmost plots.

the generative $\boldsymbol{H}$:

$$\min \left\{ \text{SNR} : d\left(\hat{\boldsymbol{H}}, \boldsymbol{H}\right) \leq \rho \right\},$$

where the distance $d\left(\hat{\boldsymbol{H}}, \boldsymbol{H}\right)$ between the estimated and the true $\boldsymbol{H}$ is defined in (5.31).

We perform our experiments for varying n and $d_{\max}$, and we consider three choices of unitary $\boldsymbol{H} \in \mathbb{R}^{100\times F}$ with ranks $F = 1$, $F = 3$ and $F = 10$. The top plots of Figure 5.6 show the results for the case $F = 1$ as a function of the number of samples n with $d_{\max} = 40$. The bottom plots of the same figure show the results for the case $F = 1$ as a function of $d_{\max}$ for fixed $n = 200$. The same experiment is performed for the cases $F = 3$ and $F = 10$, and the corresponding results are shown in Figures 5.7 and 5.8, respectively.

These results demonstrate the advantage of A-MisPCA over S-MisPCA and PCA in almost every regime. Only when $d_{\max} = 1$ does PCA compare to A-MisPCA. In that regime, the misalignments are small compared to the width of the rectangular signal and hence affect little the PCA estimate. Interestingly, the results also show that the A-MisPCA algorithm is remarkably robust to the magnitude of the misalignments.

5.5.4 Numerical comparison to OPFA

In this section we compare the performance of MisPCA and the OPFA of Chapter 4 under the two respective generative models.

We consider two measures of performance, the distance between the subspaces of the true and the estimated factors, and the MSE of the data estimator constructed by fitting a least-squares model to the estimated factors. The MisPCA and the OPFA model are inherently different: the latter allows for different (but order-preserving) shifts to apply to each factor and the former restricts the shifts to be the same for each factor. In addition, the OPFA incorporates structural constraints about the non-negativity and the smoothness of the data that MisPCA does not enforce. It is thus expected that each of this algorithms works better than the other when the generative model is the right one.

In order to test this hypothesis, we generate 100-dimensional non-negative data from a 2-factor model with random order preserving (OPFA model) and circular (MisPCA model) misalignments. For each SNR level, and each realization, we compute the distance between the true and the estimated subspaces and the reconstruction MSE. Figure 5.9 (left) shows the results under the MisPCA model, and Figure 5.9 (right) shows the results of the same experiment under an OPFA model.

As expected, each algorithm outperforms the other in both performance measures under its respective correct model. In addition, the saturation of MSE curves at

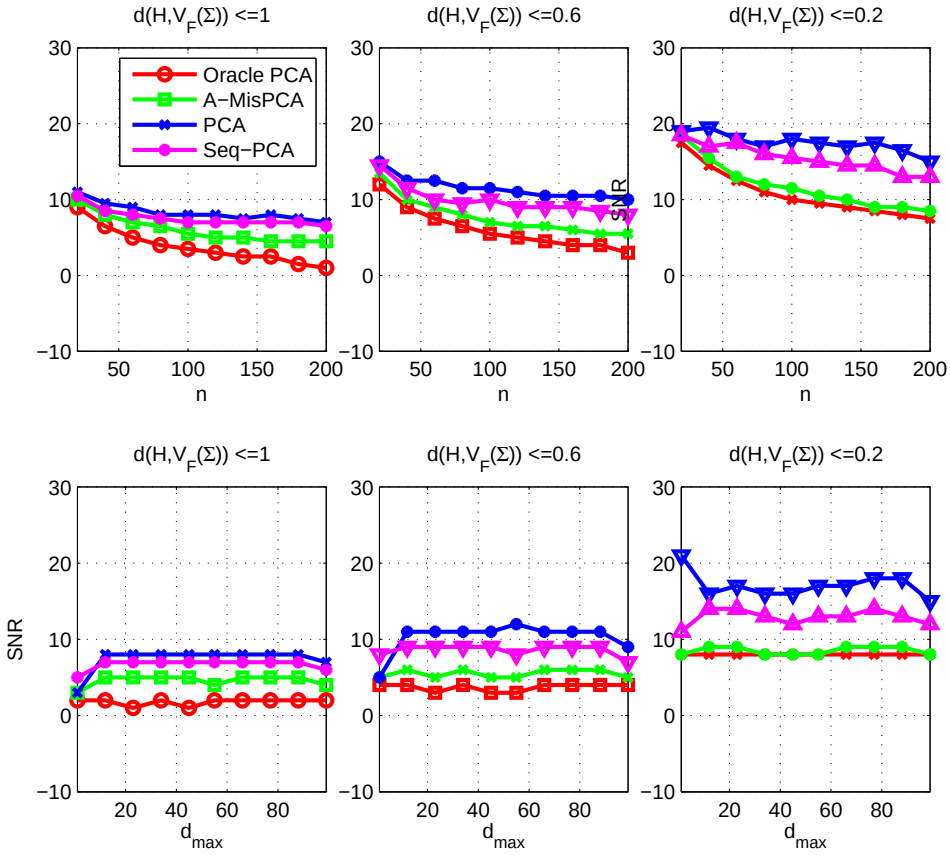


Figure 5.6: Estimated SNR levels needed for each of the algorithms to attain a level of fidelity ρ, defined as $\min\left\{\text{SNR} : d\left(\hat{\boldsymbol{H}}, \boldsymbol{H}\right) \leq \rho\right\}$, for $\rho \in \left\{F, \frac{F}{3}, \frac{2F}{3}\right\}$, as a function of the number of samples n, and as a function of $d_{\max}$, the maximum misalignment, for a rank-1 signal ($F = 1$).

high SNR reflect the bias of OPFA and MisPCA under the misspecified model. It is also interesting to note that both algorithms suffer from a similar phase transition phenomenon; only when the SNR is large enough do their factor estimates correlate with the true signal subspace.

5.5.5 A-MisPCA: Initialization and comparison to Brute Force MisPCA.

The Alternating MisPCA is a sub-optimal iterative procedure that depends on the initialization choice. In this section we show that the A-MisPCA algorithm is in fact quite robust with respect to the initialization choice, and hence we provide a justification for the random initialization criteria we use in the applications. In addition, we will also show that, despite its sub-optimality, the A-MisPCA estimator shows performance comparable to that of the brute-force estimator obtained when solving the MisPCA problem (5.7) exactly through an exhaustive search.

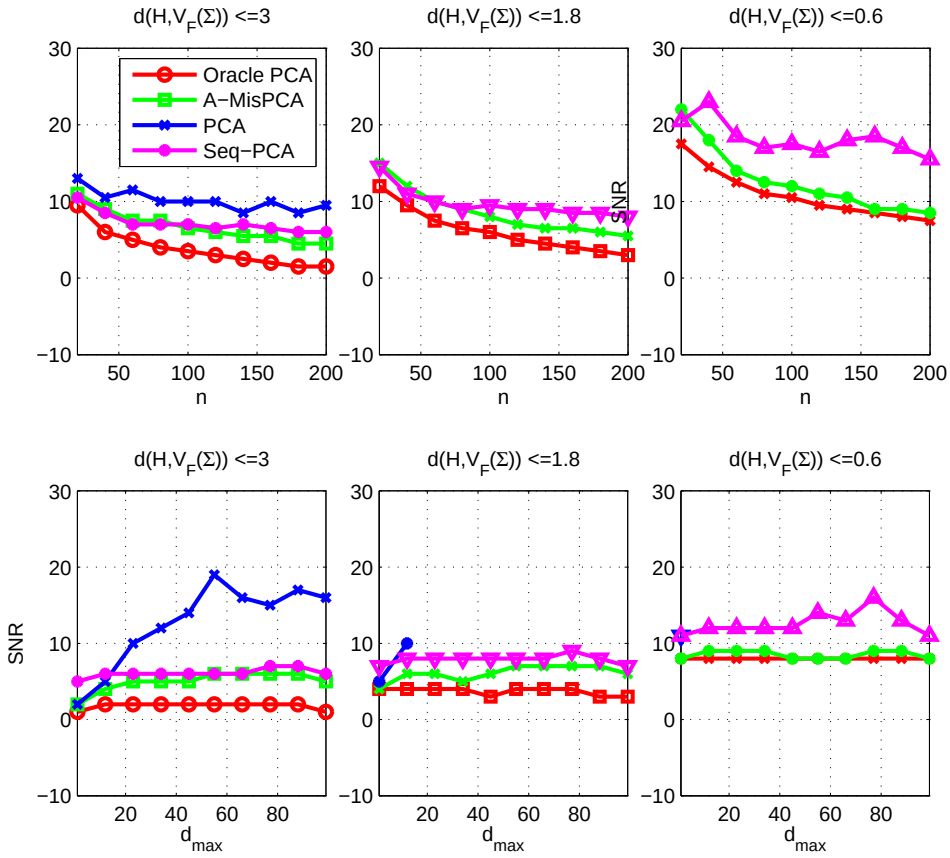


Figure 5.7: Same as in Figure 5.6 for the case $F = 3$. Notices that since PCA is biased, here it fails to attain the target fidelity level in several regimes.

Due to the computational burden of the latter, we have to limit our study here to $S = 5$, $p = 100$, $n = 50$. We chose the generative signal to be a rank-2 signal with support width equal to 10. In order to study the sensitivity of A-MisPCA to the initialization choice, we initialize the algorithm as follows. For a given $\theta \in [0, 1]$, we set the initial value $\boldsymbol{H}^0$ to:

$$\boldsymbol{H}^0 = \theta \boldsymbol{\Pi} + (1 - \theta)\boldsymbol{n},$$

where $\boldsymbol{n}$ is a random gaussian vector with unitary variance, normalized so that $||\boldsymbol{n}||_2 = 1$. Thus $\theta = 1$ corresponds to initializing with the true signal $\boldsymbol{\Pi}$ and $\theta = 0$ to a totally random initialization. We compute the affinity $d\left(\hat{\boldsymbol{H}}, \boldsymbol{H}\right)$ with respect to $\boldsymbol{\Pi}$ for the A-MisPCA estimator over a small grid of values for $(\text{SNR}, d_{\max})$ and $\theta \in \{0, .5, 1\}$. We also compute the affinity for the brute-force MisPCA (BF-MisPCA) estimator obtained by solving (5.7). The results, shown in Table 5.1, evidence the robustness of A-MisPCA with respect to initialization: the largest difference among the affinity

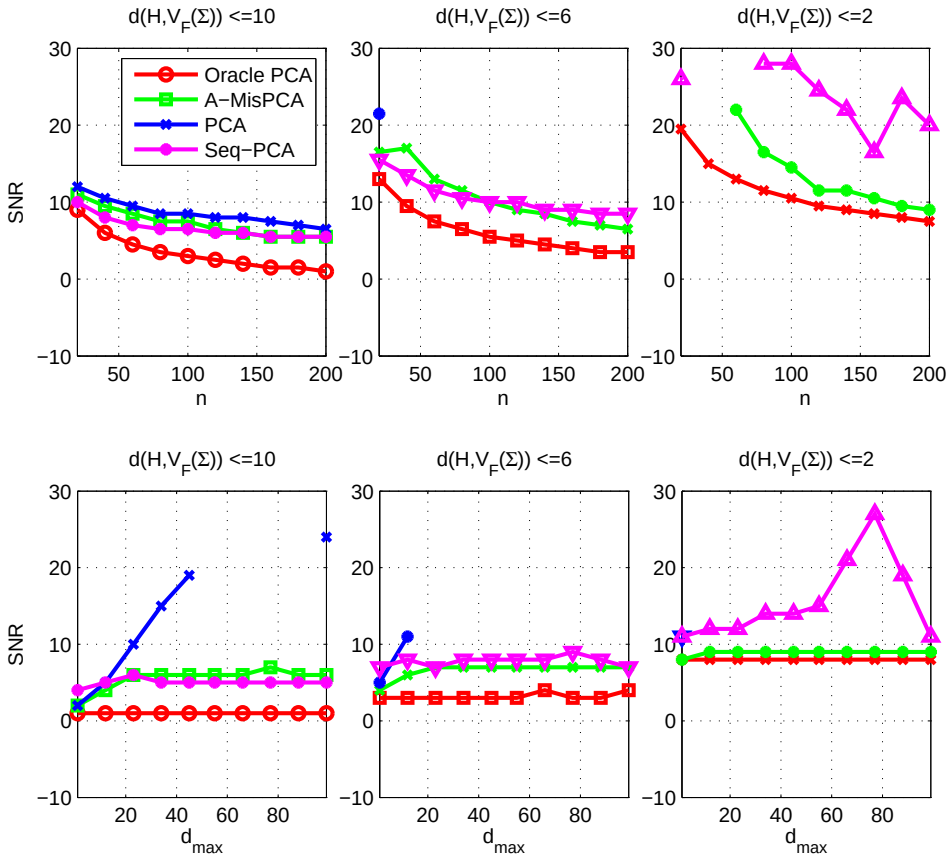


Figure 5.8: Same as in Figure 5.6 for the case $F = 10$.

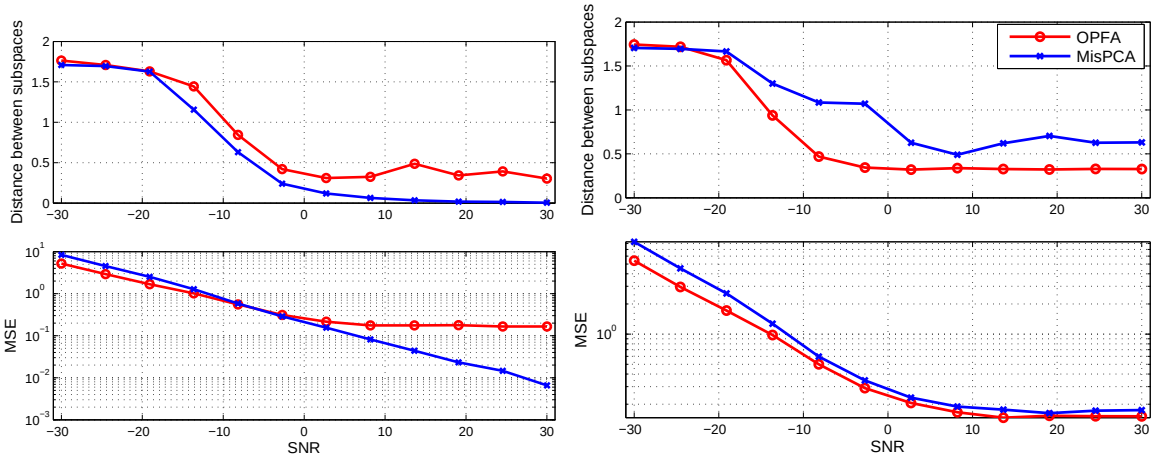


Figure 5.9: *Right*: MisPCA vs OPFA under a non-negative MisPCA generative model. *Left*: MisPCA vs OPFA under an OPFA generative model.

values obtained for different θ is smaller than 0.5. In addition, comparing to Table 5.2 shows that the differences between the affinity values obtained by BF-MisPCA

	$d\left(\boldsymbol{H}^{\text{A-MisPCA}}, \boldsymbol{\Pi}^r\right)$ [mean(std.dev.)]		
SNR	$d_{\max} = 1$	$d_{\max} = 50$	$d_{\max} = 99$
	$\theta = 0$		
-10	3.7(0.076)	3.7(0.065)	3.7(0.049)
3	1.1(0.14)	2.5(0.63)	2.6(0.63)
17	0.031(0.0038)	0.59(0.44)	0.65(0.36)
30	0.0015(0.00016)	0.66(0.53)	0.67(0.43)
	$\theta = .5$		
-10	3.8(0.059)	3.8(0.058)	3.7(0.056)
3	1.2(0.16)	2.9(0.52)	3.1(0.43)
17	0.045(0.04)	0.74(0.6)	0.71(0.38)
30	0.047(0.069)	0.7(0.6)	1(0.71)
	$\theta = 1$		
-10	3.8(0.061)	3.8(0.063)	3.7(0.068)
3	1.2(0.19)	2.7(0.6)	3.1(0.38)
17	0.1(0.087)	0.84(0.68)	0.78(0.52)
30	0.056(0.066)	0.83(0.72)	1(0.74)

Table 5.1: Sensitivity of the A-MisPCA estimates to the initialization choice.

	$d\left(\boldsymbol{H}^{\text{BF-MisPCA}}, \boldsymbol{\Pi}^r\right)$ [mean(std.dev.)]		
SNR	$d_{\max} = 1$	$d_{\max} = 50$	$d_{\max} = 99$
-10	3.7(0.072)	3.7(0.079)	3.7(0.056)
3	1.2(0.16)	1.7(0.28)	2.1(0.45)
17	0.031(0.0038)	0.44(0.19)	0.6(0.24)
30	0.0015(0.00016)	0.36(0.18)	0.55(0.23)

Table 5.2: Performance of the Brute Force MisPCA estimator.

and the A-MisPCA algorithms are smaller or equal to .1. This highlights the fact that A-MisPCA achieves performance comparable to BF-MisPCA at a much smaller computational cost.

5.5.6 Application to longitudinal gene expression data clustering

In this section we apply our methodology to the study of an influenza challenge study which is part of the (DARPA) Predicting Health and Disease program (HZR+11). This dataset consists of a collection of 272 microarray samples of dimension $G = 12023$ genes obtained from 17 individuals. All of these subjects were inoculated with influenza A H3N2Wisconsin and $T = 16$ blood samples were extracted

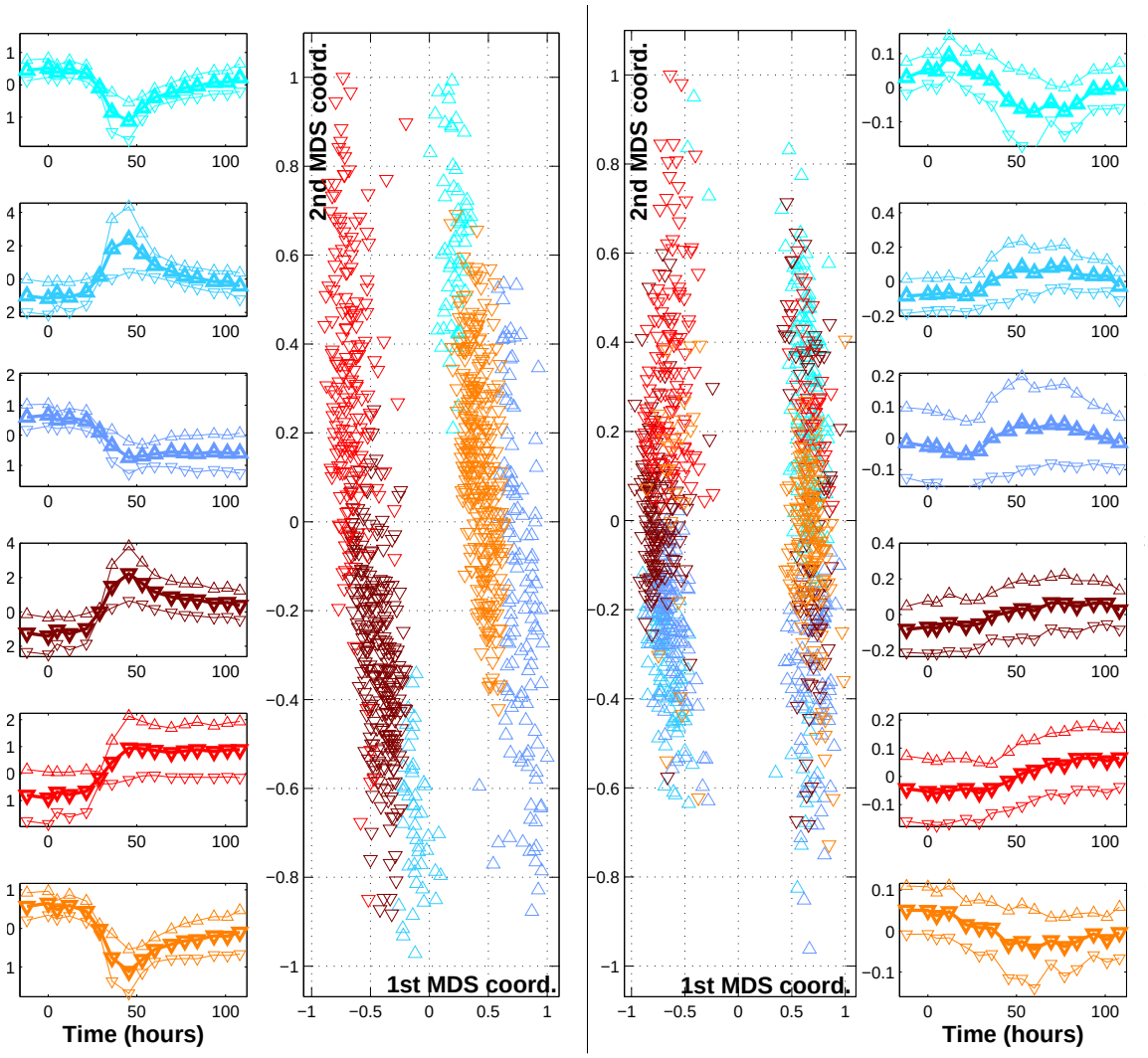


Figure 5.10: Hierarchical Clustering results obtained after MisPCA and PCA-based dimensionality reduction. The leftmost and the right most panels show the centroids (+/− standard deviations) after MisPCA and PCA, respectively. The middle panels correspond to a 2-dimensional embedding of the data projected on the MisPC's (left) and the PC's (right).

before and after inoculation at prespecified time points. Finally, the clinicians on the team established which of these subjects developed symptoms, based on a standardized symptom scoring method. In previous work, we showed that the trajectories of the gene expression values for different subjects are misaligned with respect to one another (TPWZ+11).

An important problem in the analysis of temporal gene expression data is that of performing *temporal clustering*, which consists in identifying groups of genes with similar temporal pattern. These genes are likely to be part of a biological pathway

and their temporal responses relate to the mechanistics of the process under study. In this section, we use A-MisPCA as a dimensionality reduction tool prior to clustering, and we show its advantage with respect to dimensionality reduction using standard PCA. Our approach is based on two steps. The first step consists in finding an aligned decomposition of the $T \times G$ gene expression matrix $\boldsymbol{X}_s$ for each subject, where G is the number of genes and $T = 16$ is the number of points. In this analysis, we assume that each column of each subject's gene expression matrix follows the Misaligned Linear Factor model of Section 5.2:

$$[\boldsymbol{X}_s]_{*,i} = \boldsymbol{C}_{d_s} \boldsymbol{H} [\boldsymbol{A}_s]_{*,i} + \boldsymbol{n}, \quad i = 1, \cdots, G, \quad s = 1, \cdots, S$$

Here the matrix $\boldsymbol{H} \in \boldsymbol{R}^{T \times F}$ is a matrix of temporal factors, and $\boldsymbol{C}_d$ is a circulant shift matrix parameterized by d. This is a coarser model than the OPFA one of Chapter 4: the temporal factors here are forced to be aligned equally for all gene profiles of a given subject. In addition, the model is likely to be biased due to the assumption of uncorrelation of $[\boldsymbol{A}_s]_{*,i}$, $[\boldsymbol{A}_s]_{*,j}$ essential to the MisPCA model. We view this model mismatch as part of a tradeoff between model accuracy and computational complexity.

We estimate d_s^{MisPCA} and $\boldsymbol{H}^{\text{MisPCA}}$, using MisPCA and compute $\boldsymbol{A}_s^{\text{MisPCA}}$ as the projection of the data on the aligned principal components:

$$\boldsymbol{A}_s^{\text{MisPCA}} = \left(\boldsymbol{H}^{\text{MisPCA}^T} \boldsymbol{H}^{\text{MisPCA}} \right)^{-1} \boldsymbol{H}^{\text{MisPCA}^T} \boldsymbol{C}_{d_s^{\text{MisPCA}}}^T \boldsymbol{X}_s$$

We use the same procedure with $d_s^{\text{PCA}} = 0$ to obtain the PCA estimates $\boldsymbol{H}^{\text{PCA}}$ and $\boldsymbol{A}_s^{\text{PCA}}$. The number F of Principal Components ($F = 4$) is chosen as to minimize the cross validation error, using the cross-validation procedure described in (TPWZ$^+$11). As is common in gene-expression data analysis, we apply an Analysis-of-Variance pre-processing step to select $G = 1000$ genes exhibiting high temporal variability.

The second step consist on applying a hierarchical clustering algorithm[1] to the columns of the matrices

$$\tilde{\boldsymbol{A}}^{\text{MisPCA}} = \left[\boldsymbol{A}_1^{\text{MisPCA}^T}, \cdots, \boldsymbol{A}_S^{\text{MisPCA}^T} \right]^T$$
$$\tilde{\boldsymbol{A}}^{\text{PCA}} = \left[\boldsymbol{A}_1^{\text{PCA}^T}, \cdots, \boldsymbol{A}_S^{\text{PCA}^T} \right]^T,$$

[1]The hierarchical clustering algorithm is used with standardized Euclidean distance and complete linkage. Different choices of the number of clusters were explored and 6 was shown to give the most interpretable results.

constructed from the projection of the data on the aligned subspace, $\boldsymbol{A}_s^{\text{MisPCA}}$, and the misaligned subspace, $\boldsymbol{A}_s^{\text{PCA}}$, obtained in the first step. The results are shown in Figure 5.10. The MisPCA-based centroids, shown on the leftmost panel, have on average 30% less variance that those obtained using PCA. The second and the third panel show a 2-dimensional embedding, computed using Multidimensional Scaling (MDS), of the projection of the data on the MisPC's and the Principal Components (PC's). It is clear that the clusters corresponding to up-regulated genes (low-to-high variation) are better separated from the down-regulated ones (high-to-low variation) in the MisPCA-based projections.

5.6 Conclusion

In this work we have addressed the problem of estimating a common underlying subspace from second-order statistics of misaligned time series. We have shown that misalignments introduce bias in the usual PCA estimator and increase the phase transition SNR, which is the minimum SNR required to be able to detect the signal from second order statistics. These results motivate us to propose an approximate Misaligned PCA algorithm that estimates the principal component while compensating for the misalignments. We demonstrate the accuracy of our theoretical predictions on several prototypical signals. We also show the effectiveness of our relaxed MisPCA algorithm, which outperforms PCA at little additional computational cost.

CHAPTER VI

Conclusions and future work

6.1 Conclusions

The leitmotif underlying this research work has been to demonstrate the advantage of using complex statistical models to extract information from high-dimensional noisy data. A recurring theme throughout all our projects has been the quest for a balance between computational complexity and statistical gain with respect to less refined models that do not account for data intricacies. As we have seen, the most interesting parsimonious models are often combinatorially hard to fit: one has to explore an exponentially growing number of configurations to find the optimal parameter estimate. Our approach has been based on the following principle: sometimes there is no need to solve the exact combinatorial problem; instead, a relaxation or a greedy approximation will yield estimates that have good statistical properties while being computationally solvable by today's computers.

For example, in the second and third chapter, we have shown that there exist low-dimensional reformulations of a class of non-differentiable optimization problems arising in the relaxation of structured-sparsity constraints. These reformulations have the potential of leading to efficient algorithms for complicated penalized estimation problems, enabling the introduction of complex, finely-crafted structure priors in the analysis of high-dimensional data such as gene expression time course studies. These algorithms have empirical complexities much lower than the worst-case third-order polynomial complexity traditionally associated to generalist convex optimization solvers, making its application possible to datasets with large dimension and/or large number of structural constraints.

In the fourth chapter, we have proposed a generative factor model that accounts for order-preserving misalignments between the temporal factors of observations from a random, high-dimensional multivariate time series. In this model, both the fitting of

the sparse factor scores and the estimation of the order-preserving alignments for each observation's factors are inherently combinatorial. To overcome this difficulty, we have designed convex and global optimization algorithms that exploit the structure of the problem to fit this model approximately, but efficiently. Our numerical simulations suggest that the loss due to our approximations is outweighed by the gain due to the tailoring of our model to the specific learning problem.

Finally, in Chapter 5 we have addressed the problem of subspace estimation under a noisy environment with sampling synchronization problems. This joint estimation problem is again exponentially complex: in the discrete misalignment setting, the number of possible misalignments grows exponentially with the number of observations. We have asymptotically quantified the degradation of the estimates due to the misalignments between observations, and proposed a simple algorithm that partially compensates for their negative effects. Our analysis and algorithms suggest again that in many practical cases it is not worthwhile to solve the combinatorial misaligned PCA estimation problem; instead, a simple approximate algorithm that partially compensates for the nuisance misalignments will be good enough.

6.2 Future work

For better or worse, doctoral studies are a limited-time enterprise, and many interesting questions have to be left out for future exploration. This work is no exception, and there are a number of open questions that could be the subject of interesting future research. The following list is by no means exhaustive, but illustrates the author's most interesting unexplored topics:

- *Efficient algorithms to evaluate the proximity operator of group-ℓ_2 penalties*: Our results from Chapter 3 show that the p-dimensional proximity operator of general group-ℓ_2 penalties can be evaluated through the solution of an m-dimensional, convex, smooth optimization problem, where m is potentially much smaller than p. Such an efficient evaluation would open the door to the application of fine-tuned structured sparsity penalties to a variety of high-dimensional statistical learning problems other than the penalized linear regression problems we explored.

- *Applications of group-ℓ_2 penalties for manifold-based sparsity*: Theorem III.2 shows that the GSTO is bound to satisfy

$$\mathcal{T}_{\lambda,\boldsymbol{H}}(\boldsymbol{g}) \in \mathrm{Ker}\,(\boldsymbol{A}_{D,\cdot}) \tag{6.1}$$

where $D = \cup_{i \in \mathcal{A}} G_i$ for some active set $\mathcal{A} \subseteq \{1, \cdots, m\}$. Immediate applications of the GSTO and the group-ℓ_2 penalties arise anywhere where one wants to enforce sparsity on a basis other than the canonical basis of $\mathbb{R}^p$. For example, one can choose $\boldsymbol{A}$ to be composed from chosen rows of the Discrete Fourier Transform matrix, or $\boldsymbol{A}$ to be a discretization of the differential operator, in order to enforce different types of smoothness in the operator's output. Another interesting and more general approach would be to choose $\boldsymbol{A}$ to model a linearization of a smooth manifold on $\mathbb{R}^p$. In this case, the operator output would have a sparse representation on the linearization of the manifold, allowing for a whole new class of structural constraints.

- *Development of conditions under A-MisPCA/Seq-MisPCA solve the MisPCA problems*: We have given in Section 5.3 and Alternating-MisPCA algorithm to approximate the solution to the combinatorial MisPCA problem. It is not hard to show that, for $F = 1$, under high SNR conditions and signals with non-flat autocorrelation, the A-MisPCA and the Seq-MisPCA algorithms solve the MisPCA problem. An interesting project would be to characterize the conditions under which the A-MisPCA algorithm leads to the MisPCA solution, as a function of the SNR and the characteristics of the underlying signal.

- *PCA under uniform misalignment*: There is another aspect of the MisPCA problem that we have left unexplored. In fact, from our developments in Chapter 5, it follows that under the deterministic, equispaced misaligned model, the covariance matrix of the misaligned observations reduces to:

$$\boldsymbol{\Sigma}(\boldsymbol{\tau}) = \frac{\text{SNR}}{d_{\max} + 1} \boldsymbol{\mathcal{H}} \text{diag}(\mathbf{1} \otimes \bar{\boldsymbol{\sigma}}) \boldsymbol{\mathcal{H}}^T + \boldsymbol{I}_p.$$

  In this setting, an effective algorithm to estimate $\boldsymbol{H}$ could be based on MUSIC/SPIRIT-like techniques, where one estimates a basis for the noise subspace, denoted by $\boldsymbol{N}$, and consequently estimates $\boldsymbol{H}$ by solving:

$$\min_{\boldsymbol{H}^T \boldsymbol{H} = \boldsymbol{I}} \left|\left| \boldsymbol{N}^T \boldsymbol{\mathcal{H}} \right|\right|_2,$$

  that is, one tries to estimate the signal $\boldsymbol{H}$, whose equispaced translated version, given by $\boldsymbol{\mathcal{H}}$, is most orthogonal to the noise subspace.

APPENDICES

APPENDIX A

Appendix to Chapter 2

A.1 Derivation of the gradient and Hessian for the Projected Newton method.

First we derive the formulae (2.24) used in the Projected Newton Algorithm of Section 2.2.1 , for the case $\boldsymbol{W} = \boldsymbol{I}$. For other $\boldsymbol{W} \succ \boldsymbol{0}$, the formulas follow through a change of variable. Letting

$$q(\eta, \boldsymbol{\mu}) \;\; := -\tfrac{1}{2}(\boldsymbol{g} - \lambda\boldsymbol{\mu})^T \boldsymbol{H}^{\dagger}(\boldsymbol{g} - \lambda\boldsymbol{\mu}) - \eta(\boldsymbol{\mu}'\boldsymbol{\mu} - 1)$$

and using the fact that $w(\eta) = \max_{\boldsymbol{\mu}\in\mathcal{R}(\boldsymbol{H})} q(\eta, \boldsymbol{\mu})$, we have:

$$w'(\eta) \;\; := \tfrac{dw(\eta)}{d\eta} = \left.\tfrac{dq(\eta,\boldsymbol{\mu})}{d\eta}\right|_{\boldsymbol{\mu}=\boldsymbol{\mu}_{opt}} = 1 - \boldsymbol{\mu}_{opt}^T\boldsymbol{\mu}_{opt} = 1 - \tfrac{\lambda^2}{4}\boldsymbol{g}'\boldsymbol{B}^{-2}(\eta)\,\boldsymbol{g}$$

where $\boldsymbol{\mu}_{opt} := \arg\max_{\boldsymbol{\mu}\in\mathcal{R}(\boldsymbol{H})} q(\eta, \boldsymbol{\mu}) = \frac{\lambda}{2}\boldsymbol{B}^{-1}(\eta)\,\boldsymbol{g}$. Also,

$$\begin{aligned} w''(\eta) \;\; &:= \tfrac{d^2w(\eta)}{(d\eta)^2} = \tfrac{\lambda^2}{2}\boldsymbol{g}^T\tfrac{d\boldsymbol{B}^{-1}(\eta)}{d\eta}\boldsymbol{g} \\ &= \tfrac{\lambda^2}{2}\boldsymbol{g}'\boldsymbol{C}(\eta)\,\boldsymbol{g}. \end{aligned}$$

where $\boldsymbol{C}(\eta) := \boldsymbol{B}^{-3}(\eta)\,\boldsymbol{H}$. Since $\eta \geq 0$, $\boldsymbol{g} \in \mathcal{R}(\boldsymbol{H})$, $\mathcal{R}(\boldsymbol{H}) = \mathcal{R}(\boldsymbol{C}(\eta))$ and $\lambda > 0$ then $\frac{\lambda^2}{2}\boldsymbol{g}^T\boldsymbol{C}(\eta)\,\boldsymbol{g} > 0$. Furthermore, if $\lambda_i(\boldsymbol{H}) < \infty$, it holds that:

$$\lambda_i(\boldsymbol{C}(\eta)) \leq \max_i \frac{\lambda_i(\boldsymbol{H})}{\left(\eta\lambda_i(\boldsymbol{H}) + \frac{\lambda^2}{2}\right)^3} \leq \max_i 8\frac{\lambda_i(\boldsymbol{H})}{\lambda^6}$$

$$\infty > \frac{4\|\boldsymbol{g}\|_2^2}{\lambda^4} \max_i \lambda_i (\boldsymbol{H}) \geq w''(\eta) > 0$$

for $\forall \eta \geq 0$. It follows that $w(\eta)$ is strictly convex and $w''(\eta)$ is uniformly bounded over $\eta \succeq 0$. Since $w(\eta) \geq 0$ is lower bounded and $\eta \succeq 0$ is a compact set, we can invoke Theorem 4.1 in (Dun80) to conclude that the Goldstein Projected Newton iterate (2.23) converges method to the minimizer of (2.5).

APPENDIX B

Appendix to Chapter 3

B.1 Proof of Corollary III.3

We will assume without loss of generality that the groups G_i are sorted in ascending order according to the indices they contain. First, we identify the objective in (3.25) with that in (3.11) by setting $\boldsymbol{A}_{G_i,G_i} = \boldsymbol{I}_{n_i}$ and $\boldsymbol{A}_{G_i,\bar{G}_i} = \boldsymbol{0}$. Since $G_i \cap G_j = \emptyset$ for $i \neq j$, this implies that $\boldsymbol{A} = \boldsymbol{I}_p$ and $\sum_{i=1}^m n_i = p$. It is clear that for any $S \subset \{1, \cdots, p\}$ and its complementary $\bar{S} = \{1, \cdots, p\} \setminus S$, $\boldsymbol{I}_p$ verifies:

$$\mathrm{Ker}\left(\boldsymbol{I}_{S,.}\right) \cap \mathrm{Ker}\left(\boldsymbol{I}_{\bar{S},.}\right) = \{\boldsymbol{0}\}.$$

and hence we can safely invoke Theorem III.2 with $\boldsymbol{A} = \boldsymbol{I}_p$ and $I_i = G_i$ verifying the conditions in the statement of this Corollary.

Second, for any given $\boldsymbol{\eta}$, using the definition of $\boldsymbol{B}_i$ in (3.44) allows us to conclude that the sets D and $\bar{D}$ are simply given by:

$$D = \cup_{i:\eta_i=0} G_i \quad \text{and} \quad \bar{D} = \cup_{i:\eta_i>0} G_i.$$

It is easy to check that, for any given $\boldsymbol{\eta}$,

$$\mathrm{Ker}\left(\boldsymbol{A}_{D,.}\right) = \{\boldsymbol{v} \in \mathbb{R}^p : \boldsymbol{v}_D = \boldsymbol{0}\}$$

and hence we can choose the basis $\boldsymbol{B}_D = \left[\boldsymbol{I}_{p-|D|}, \boldsymbol{0}^T_{|D|\times p-|D|}\right]^T$ for $\mathrm{Ker}\left(\boldsymbol{A}_{D,.}\right)$. Identifying the sets $Z = D$ and $\bar{Z} = \bar{D}$ and using $\boldsymbol{A} = \boldsymbol{I}_p$, the matrix $\boldsymbol{\Gamma}\left(\boldsymbol{\eta}\right)$ in (3.13)

specializes to:

$$\boldsymbol{\Gamma}(\boldsymbol{\eta}) = \boldsymbol{S}(\boldsymbol{\eta}) + \boldsymbol{X}_{\bar{Z},\cdot}^T \boldsymbol{X}_{\bar{Z},\cdot},$$

with $\boldsymbol{S}(\boldsymbol{\eta})$ defined in (3.28). Finally, we apply the fact that for any vector $\boldsymbol{v} \in \mathbb{R}^{p-|D|}$, and the basis $\boldsymbol{B}_D$ defined above, $\boldsymbol{B}_D \boldsymbol{v} = \left[\boldsymbol{v}_{\bar{Z}}^T, \boldsymbol{0}_{|Z|}^T\right]^T$ to obtain (3.26) from (3.12).

B.2 Proof of Corollary III.4

First, we let $\boldsymbol{A} \in \mathbb{R}^{\sum_{i=1}^m n_i \times p}$ be defined as follows:

$$\boldsymbol{A}_{I_i,G_i} = \operatorname{diag}(\boldsymbol{w}_i) \quad \boldsymbol{A}_{I_i,\bar{G}_i} = \boldsymbol{0}_{n_i \times p - n_i} \quad i = 1, \cdots, m \tag{B.1}$$

where $I_i = \left[\sum_j^{i-1} n_j, \sum_j^{i-1} n_j + 1, \cdots, \sum_j^i n_j - 1\right]$. For any subset $S \subseteq \{1, \cdots, \sum_{i=1}^m n_i\}$, we claim that:

$$\operatorname{Ker}(\boldsymbol{A}_{S,\cdot}) = \{\boldsymbol{v} \in \mathbb{R}^p : \boldsymbol{v}_Z = \boldsymbol{0},\ Z = \cup_{i:S\cap I_i \neq \emptyset} \operatorname{supp}(\boldsymbol{A}_{S\cap I_i,\cdot})\} \tag{B.2}$$

where $\operatorname{supp}(\boldsymbol{X})$ is the set of indices corresponding to columns of $\boldsymbol{X}$ with at least one non-zero element. We prove this claim as follows. First, it is clear that if $\boldsymbol{v}$ belongs to the set on the left hand side, then

$$\boldsymbol{A}_{S\cap G_i,\cdot} \boldsymbol{v} = \boldsymbol{0}_{|S\cap G_i|}$$

for any i such that $S \cap G_i \neq \emptyset$. On the other hand, suppose that $\boldsymbol{k} \in \operatorname{Ker}(\boldsymbol{A}_{S,\cdot})$. Then, for every i such that $S \cap G_i \neq \emptyset$,

$$\boldsymbol{A}_{S\cap G_i,\cdot} \boldsymbol{k} = \operatorname{diag}\left([\boldsymbol{w}_i]_{\operatorname{supp}(A_{S\cap G_i,\cdot})}\right) [\boldsymbol{k}]_{\operatorname{supp}(A_{S\cap G_i,\cdot})} = \boldsymbol{0}.$$

Since $\operatorname{diag}\left([\boldsymbol{w}_i]_{\operatorname{supp}(A_{S\cap G_i,\cdot})}\right) \succ 0$, this implies that $[\boldsymbol{k}]_{\operatorname{supp}(A_{S\cap G_i,\cdot})}$ has to be zero in order to comply with the leftmost equality, which completes the proof of the claim made in (B.2).

By (B.2), we have that if $\exists \boldsymbol{v} \in \operatorname{Ker}(\boldsymbol{A}_{S,\cdot}) \cap \operatorname{Ker}\left(\boldsymbol{A}_{\bar{S},\cdot}\right)$ with $\boldsymbol{v} \neq \boldsymbol{0}$, then this vector must satisfy:

$$\boldsymbol{v}_{Z_1 \cup Z_2} = \boldsymbol{0}, \quad \boldsymbol{v}_{\{1,\cdots,p\} \setminus Z_1 \cup Z_2} \neq \boldsymbol{0}.$$

where

$$Z_1 = \cup_{i:S\cap I_i\neq\emptyset}\text{supp}\left(\boldsymbol{A}_{S\cap I_i,\cdot}\right)$$

and

$$Z_2 = \cup_{i:\bar{S}\cap I_i\neq\emptyset}\text{supp}\left(\boldsymbol{A}_{\bar{S}\cap I_i,\cdot}\right).$$

Since we have assumed that $\cup_{i=1}^m G_i = \{1,\cdots,p\}$, it follows that $\{1,\cdots,p\}\setminus Z_1\cup Z_2 = \emptyset$ and hence such $\boldsymbol{v}$ has to be $\boldsymbol{v}=\boldsymbol{0}$. It follows that $\boldsymbol{A}$ verifies condition (3.10) and we can invoke Theorem III.2. Second, for any given $\boldsymbol{\eta}$, and using the equivalence in (B.2), we can characterize $\text{Ker}\left(\boldsymbol{A}_{D,\cdot}\right)$ as follows:

$$\text{Ker}\left(\boldsymbol{A}_{D,\cdot}\right) = \{\boldsymbol{v}\in\mathbb{R}^p : \boldsymbol{v}_Z = \boldsymbol{0},\ \text{where } Z = \cup_{i:\eta_i=0}G_i\}$$

where D is defined in (3.14). It follows that $\boldsymbol{B}_D = \left[\boldsymbol{I}_{p-|Z|}, \boldsymbol{0}^T_{|Z|\times p-|Z|}\right]^T$ is a basis for $\text{Ker}\left(\boldsymbol{A}_{D,\cdot}\right)$, with $Z = \cup_{i:\eta_i=0}G_i$. Substituting this choice of $\boldsymbol{B}_D$ and $\boldsymbol{A}$ defined by (B.1) in (3.13) leads to:

$$\boldsymbol{\Gamma}\left(\boldsymbol{\eta}\right) = \boldsymbol{S}\left(\boldsymbol{\eta}\right) + \boldsymbol{X}^T_{\bar{Z},\cdot}\boldsymbol{X}_{\bar{Z},\cdot},$$

with $\boldsymbol{S}\left(\boldsymbol{\eta}\right)$ defined in (3.33). Finally, we apply the fact that for any vector $\boldsymbol{v}\in\mathbb{R}^{p-|Z|}$, and the basis $\boldsymbol{B}_D$ defined above, $\boldsymbol{B}_D\boldsymbol{v} = \left[\boldsymbol{v}^T_{\bar{Z}}, \boldsymbol{0}^T_{|Z|}\right]^T$ to obtain (3.31) from (3.12).

B.3 Proof of Theorem III.6

First we will need to prove two auxiliary results. First, for any $\boldsymbol{\eta}\succ 0$ and a small enough α such that $\boldsymbol{\eta}+\alpha\boldsymbol{d}\succ 0$, we claim that:

$$\boldsymbol{C}^{-1}\left(\boldsymbol{\eta}+\alpha\boldsymbol{d}\right) = \boldsymbol{C}^{-1}\left(\boldsymbol{\eta}\right) - \alpha\boldsymbol{C}^{-1}\left(\boldsymbol{\eta}\right)\boldsymbol{C}\left(d\right)\boldsymbol{C}^{-1}\left(\boldsymbol{\eta}\right) + O\left(\alpha^2\right) \tag{B.3}$$

To see this, observe that by definition of the matrix $\boldsymbol{C}\left(\boldsymbol{\eta}\right)$:

$$\boldsymbol{C}\left(\boldsymbol{\eta}+\alpha\boldsymbol{d}\right) = \boldsymbol{C}\left(\boldsymbol{\eta}\right) + \alpha\boldsymbol{C}\left(\boldsymbol{d}\right).$$

Now for any two symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{E}$, repeatedly apply the formula:

$$\left(\boldsymbol{A}+\boldsymbol{E}\right)^{-1} = \boldsymbol{A}^{-1} - \left(\boldsymbol{A}+\boldsymbol{E}\right)^{-1}\boldsymbol{E}\boldsymbol{A}^{-1}$$

to obtain:

$$(\boldsymbol{A}+\boldsymbol{E})^{-1}=\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{E}\boldsymbol{A}^{-1}+(\boldsymbol{A}+\boldsymbol{E})^{-1}\left(\boldsymbol{E}\boldsymbol{A}^{-1}\right)^{2} \tag{B.4}$$

Substituting $\boldsymbol{A}=\boldsymbol{C}(\boldsymbol{\eta})$ and $\boldsymbol{E}=\alpha\boldsymbol{C}(\boldsymbol{d})$ and observing that both $\boldsymbol{C}^{-1}(\boldsymbol{\eta})$ and $(\boldsymbol{C}(\boldsymbol{\eta})+\alpha\boldsymbol{C}(\boldsymbol{d}))^{-1}$ exist when $\boldsymbol{\eta}\succ\mathbf{0}$ for a small enough α, leads to (B.3). Our second claim is that:

$$\begin{aligned}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta}+\alpha\boldsymbol{d})= &\ \boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta})+\frac{\alpha}{2}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta})\boldsymbol{A}^{T}\boldsymbol{C}^{-1}(\boldsymbol{\eta})\boldsymbol{C}(d)\boldsymbol{C}^{-1}(\boldsymbol{\eta})\boldsymbol{A}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta})\\ &+O\left(\alpha^{2}\right)\end{aligned} \tag{B.5}$$

To prove this, first we will use (B.3) to write:

$$\begin{aligned}\boldsymbol{\Gamma}(\boldsymbol{\eta}+\alpha\boldsymbol{d}) &= \frac{1}{2}\boldsymbol{A}^{T}\boldsymbol{C}^{-1}(\boldsymbol{\eta}+\alpha\boldsymbol{d})\boldsymbol{A}+\boldsymbol{X}^{T}\boldsymbol{X}\\ &= \boldsymbol{\Gamma}(\boldsymbol{\eta})-\frac{\alpha}{2}\boldsymbol{A}^{T}\boldsymbol{C}^{-1}(\boldsymbol{\eta})\boldsymbol{C}(d)\boldsymbol{C}^{-1}(\boldsymbol{\eta})\boldsymbol{A}+O\left(\alpha^{2}\right)\end{aligned}$$

As proved during the proof of Theorem III.2, $\boldsymbol{\Gamma}(\boldsymbol{\eta})\succ 0$ and so is $\boldsymbol{\Gamma}(\boldsymbol{\eta}+\alpha\boldsymbol{d})$ if $\boldsymbol{\eta}+\alpha\boldsymbol{d}\succ\mathbf{0}$. It follows that we can apply (B.4) again with $\boldsymbol{A}=\boldsymbol{\Gamma}(\boldsymbol{\eta})$ and $\boldsymbol{E}=-\frac{\alpha}{2}\boldsymbol{A}^{T}\boldsymbol{C}^{-1}(\boldsymbol{\eta})\boldsymbol{C}(d)\boldsymbol{C}^{-1}(\boldsymbol{\eta})\boldsymbol{A}+O(\alpha^{2})$ to obtain:

$$\begin{aligned}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta}+\alpha\boldsymbol{d})= &\ \boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta})+\frac{\alpha}{2}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta})\boldsymbol{A}^{T}\boldsymbol{C}^{-1}(\boldsymbol{\eta})\boldsymbol{C}(d)\boldsymbol{C}^{-1}(\boldsymbol{\eta})\boldsymbol{A}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta})\\ &-O\left(\alpha^{2}\right)\end{aligned}$$

as $\alpha\to 0$.

We are now ready to compute the directional derivative of $w(\boldsymbol{\eta})$ at $\boldsymbol{\eta}\succ 0$:

$$\begin{aligned}w'(\boldsymbol{\eta},\boldsymbol{d}) &= \lim_{\alpha\to 0}\frac{w(\boldsymbol{\eta}+\alpha\boldsymbol{d})-w(\boldsymbol{\eta})}{\alpha}\\ &= \lim_{\alpha\to 0}\frac{-\frac{1}{2}\boldsymbol{y}^{T}\boldsymbol{X}\boldsymbol{B}_{D}^{T}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta}+\alpha\boldsymbol{d})\boldsymbol{B}_{D}\boldsymbol{X}^{T}\boldsymbol{y}+\frac{1}{2}\boldsymbol{y}^{T}\boldsymbol{X}\boldsymbol{B}_{D}^{T}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta})\boldsymbol{B}_{D}\boldsymbol{X}^{T}\boldsymbol{y}}{\alpha}+\lambda^{2}\boldsymbol{d}^{T}\boldsymbol{c}\\ &= -\frac{1}{4}\boldsymbol{y}^{T}\boldsymbol{X}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta})\boldsymbol{A}^{T}\boldsymbol{C}^{-1}(\boldsymbol{\eta})\boldsymbol{C}(d)\boldsymbol{C}^{-1}(\boldsymbol{\eta})\boldsymbol{A}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta})\boldsymbol{X}^{T}\boldsymbol{y}+\lambda^{2}\boldsymbol{d}^{T}\boldsymbol{c}\\ &= -\boldsymbol{\nu}^{T}\boldsymbol{C}(d)\boldsymbol{\nu}+\lambda^{2}\boldsymbol{d}^{T}\boldsymbol{c},\end{aligned}$$

which shows that the directional derivative exists and is linear in $\boldsymbol{d}$, an hence $w(\boldsymbol{\eta})$, is a differentiable function. The gradient is then given by (3.41). To show that $w(\boldsymbol{\eta})$ is also twice differentiable, we will proceed as follows. First, we denote each coordinate

of the gradient of $w(\boldsymbol{\eta})$ by $g_i(\boldsymbol{\eta})$, i.e.:

$$g_i(\boldsymbol{\eta}) = \lambda^2 c_i - ||\boldsymbol{\nu}_{G_i}||_2^2 = \lambda^2 c_i - ||\boldsymbol{B}_i \boldsymbol{\nu}||_2^2 .$$

where

$$\boldsymbol{\nu} = \frac{1}{2} \boldsymbol{C}^{-1}(\boldsymbol{\eta}) \boldsymbol{A} \boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta}) \boldsymbol{X}^T \boldsymbol{y}$$

The directional derivative of the gradient's coordinates is then given by:

$$g_i'(\boldsymbol{\eta}; d) = \lim_{\alpha \to 0} \frac{-\frac{1}{4} \left|\left| \boldsymbol{B}_i \boldsymbol{C}^{-1}(\boldsymbol{\eta} + \alpha \boldsymbol{d}) \boldsymbol{A} \boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta} + \alpha \boldsymbol{d}) \boldsymbol{X}^T \boldsymbol{y} \right|\right|_2^2 + \frac{1}{4} ||\boldsymbol{B}_i \boldsymbol{\nu}||_2^2}{\alpha} \quad \text{(B.6)}$$

We will first focus on the first element in the denominator above. Using (B.3) we can conclude that

$$\boldsymbol{C}^{-1}(\boldsymbol{\eta} + \alpha \boldsymbol{d}) \boldsymbol{A} \boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta} + \alpha \boldsymbol{d}) \boldsymbol{X}^T \boldsymbol{y}$$

is equal to

$$\boldsymbol{C}^{-1}(\boldsymbol{\eta}) \boldsymbol{A} \boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta} + \alpha \boldsymbol{d}) \boldsymbol{X}^T \boldsymbol{y} - \alpha \boldsymbol{C}^{-1}(\boldsymbol{\eta}) \boldsymbol{C}(d) \boldsymbol{C}^{-1}(\boldsymbol{\eta}) \boldsymbol{A} \boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta} + \alpha \boldsymbol{d}) \boldsymbol{X}^T \boldsymbol{y} + O\left(\alpha^2\right).$$

Substituting $\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta} + \alpha \boldsymbol{d})$ above using (B.5), we obtain:

$$\boldsymbol{\nu} + \frac{\alpha}{2} \boldsymbol{C}^{-1}(\boldsymbol{\eta}) \boldsymbol{A} \boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta}) \boldsymbol{A}^T \boldsymbol{C}^{-1}(\boldsymbol{\eta}) \boldsymbol{C}(d) \boldsymbol{\nu} - \alpha \boldsymbol{C}^{-1}(\boldsymbol{\eta}) \boldsymbol{C}(d) \boldsymbol{\nu} + O\left(\alpha^2\right),$$

which yields:

$$\begin{aligned} \frac{1}{4} \left|\left| \boldsymbol{B}_i \boldsymbol{C}^{-1}(\boldsymbol{\eta} + \alpha \boldsymbol{d}) \boldsymbol{A} \boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta} + \alpha \boldsymbol{d}) \boldsymbol{X}^T \boldsymbol{y} \right|\right|_2^2 &= ||\boldsymbol{B}_i \boldsymbol{\nu}||_2^2 + \\ & \alpha \boldsymbol{\nu}^T \boldsymbol{B}_i \boldsymbol{C}^{-1}(\boldsymbol{\eta}) \boldsymbol{A} \boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta}) \boldsymbol{A}^T \boldsymbol{C}^{-1}(\boldsymbol{\eta}) \boldsymbol{C}(d) \boldsymbol{\nu} \\ & -2\alpha \boldsymbol{B}_i \boldsymbol{\nu}^T \boldsymbol{C}^{-1}(\boldsymbol{\eta}) \boldsymbol{C}(d) \boldsymbol{\nu} + O\left(\alpha^2\right). \end{aligned}$$

Plugging this expression back to (B.6) leads to:

$$g_i'(\boldsymbol{\eta}; d) = 2\boldsymbol{\nu}^T \boldsymbol{B}_i \boldsymbol{C}^{-1}(\boldsymbol{\eta}) \boldsymbol{C}(d) \boldsymbol{\nu} - \boldsymbol{\nu}^T \boldsymbol{B}_i \boldsymbol{C}^{-1}(\boldsymbol{\eta}) \boldsymbol{A} \boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta}) \boldsymbol{A}^T \boldsymbol{C}^{-1}(\boldsymbol{\eta}) \boldsymbol{C}(d) \boldsymbol{\nu},$$

which shows that $g_i(\boldsymbol{\eta})$ is differentiable and hence $w(\boldsymbol{\eta})$ is twice differentiable with Hessian given in (3.42).

B.4 Proof of Theorem III.7

First observe that if $\boldsymbol{\eta}^*$ is optimal, it also verifies:

$$\begin{aligned} \boldsymbol{\eta}^*_{\bar{\Omega}} = \quad & \arg\min \quad w_{\bar{\Omega}}\left(\boldsymbol{\eta}_{\bar{\Omega}}\right) \\ & \boldsymbol{\eta}^*_{\bar{\Omega}} \succ \mathbf{0}. \end{aligned} \tag{B.7}$$

Since $\boldsymbol{\eta}_{\bar{\Omega}} \succ 0$, it follows by Theorem III.6 that $w_{\bar{\Omega}}\left(\boldsymbol{\eta}_{\bar{\Omega}}\right)$ is twice differentiable and that $\boldsymbol{\eta}^*_{\bar{\Omega}} \succ \mathbf{0}$ has to verify the optimality conditions:

$$\nabla w_{\bar{\Omega}}\left(\boldsymbol{\eta}_{\bar{\Omega}}\right) = \mathbf{0}. \tag{B.8}$$

Or, using the expression for the gradient given in (3.41), $\boldsymbol{\eta}^*_{\bar{\Omega}}$ solves:

$$\lambda^2 c_i - \boldsymbol{\nu}_{G_i}^T \boldsymbol{\nu}_{G_i} = 0 \quad \text{for each } i \in \bar{\Omega}, \tag{B.9}$$

where we have let:

$$\boldsymbol{\nu} = \frac{1}{2}\boldsymbol{C}^{-1}\left(\boldsymbol{\eta}_{\bar{\Omega}}\right)\boldsymbol{A}\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\eta}_{\bar{\Omega}}\right)\boldsymbol{B}_D^T\boldsymbol{X}^T\boldsymbol{y} \tag{B.10}$$

By Theorem III.6, the function

$$g_i\left(\boldsymbol{\eta}_{\bar{\Omega}}, \lambda\right) = \lambda^2 c_i - \boldsymbol{\nu}^T \boldsymbol{B}_i \boldsymbol{\nu}$$

is continuous and differentiable with respect to λ and $\boldsymbol{\eta}_{\bar{\Omega}}$. Its transposed gradient is the Hessian of $w_{\bar{\Omega}}\left(\boldsymbol{\eta}_{\bar{\Omega}}\right)$, and is given by (3.42). If $\nabla_{\boldsymbol{\eta}_{\bar{\Omega}}}\boldsymbol{g}\left(\boldsymbol{\eta}_{\bar{\Omega}}, \lambda\right) = \nabla^2 w_{\bar{\Omega}}\left(\boldsymbol{\eta}_{\bar{\Omega}}\right)$ is non-singular at $\boldsymbol{\eta}^*_{\bar{\Omega}}$, then by the implicit function theorem (see for instance Proposition 1.1.14 in (BNO$^+$03)):

$$\frac{d\boldsymbol{\eta}^*_{\bar{\Omega}}}{d\lambda} = -\left(\nabla^2 w_{\bar{\Omega}}\left(\boldsymbol{\eta}^*_{\bar{\Omega}}\right)\right)^{-1} \nabla_\lambda \boldsymbol{g}\left(\boldsymbol{\eta}_{\bar{\Omega}}, \lambda\right). \tag{B.11}$$

Substituting

$$\nabla_\lambda \boldsymbol{g}\left(\boldsymbol{\eta}_{\bar{\Omega}}, \lambda\right) = 2\lambda \boldsymbol{c} \tag{B.12}$$

above yields (3.55).

B.5 Proof of Theorem III.5

Applying Lemma III.1 with $\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{g} = -\boldsymbol{y}$, problem (3.35) is equivalent to:

$$\begin{array}{ll} \max & -\frac{1}{2}\left(-\boldsymbol{y} - \boldsymbol{A}^T\boldsymbol{\nu}\right)^T\left(-\boldsymbol{y} - \boldsymbol{A}^T\boldsymbol{\nu}\right) \\ & \boldsymbol{\nu}^T\boldsymbol{B}_i\boldsymbol{\nu} \leq \lambda_i^2 \qquad\qquad i = 1, \cdots, m \end{array}$$

The Lagrange dual function of this problem is given by:

$$l(\boldsymbol{\eta}) \;\; = \sup_{\boldsymbol{\nu}} -\tfrac{1}{2}\left(-\boldsymbol{y} - \boldsymbol{A}^T\boldsymbol{\nu}\right)^T\left(-\boldsymbol{y} - \boldsymbol{A}^T\boldsymbol{\nu}\right) - \tfrac{1}{2}\boldsymbol{\nu}^T 2\boldsymbol{C}(\boldsymbol{\eta})\,\boldsymbol{\nu} + \boldsymbol{\eta}^T\boldsymbol{\lambda}^2$$

Since $\boldsymbol{A}\boldsymbol{A}^T \succeq 0$, the above problem is concave and its optimality conditions are given by:

$$\left(\boldsymbol{A}\boldsymbol{A}^T + 2\boldsymbol{C}(\boldsymbol{\eta})\right)\boldsymbol{\nu} = -\boldsymbol{A}\boldsymbol{y}$$

It is easy to check that $\boldsymbol{A}\boldsymbol{y} \in \mathcal{R}\left(\boldsymbol{A}\boldsymbol{A}^T + 2\boldsymbol{C}(\boldsymbol{\eta})\right)$ and since $\boldsymbol{A}\boldsymbol{A}^T \succ \boldsymbol{0}$ the supremum is unique and attained at:

$$\boldsymbol{\nu}^* = -\left(\boldsymbol{A}\boldsymbol{A}^T + 2\boldsymbol{C}(\boldsymbol{\eta})\right)^{-1}\boldsymbol{A}\boldsymbol{y},$$

which yields (3.36) through equation (3.7):

$$\mathcal{P}_\lambda(\boldsymbol{y}) = \boldsymbol{y} + \boldsymbol{A}^T\boldsymbol{\nu}^* = \left(\boldsymbol{I} - \boldsymbol{A}^T\left(\boldsymbol{A}\boldsymbol{A}^T + 2\boldsymbol{C}(\boldsymbol{\eta})\right)^{-1}\boldsymbol{A}\right)\boldsymbol{y}.$$

Finally, evaluating $l(\boldsymbol{\eta})$ at $\boldsymbol{\nu}^*$ yields:

$$l(\boldsymbol{\eta}) \;\; = \tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{A}^T\left(\boldsymbol{A}\boldsymbol{A}^T + 2\boldsymbol{C}(\boldsymbol{\eta})\right)^{-1}\boldsymbol{A}\boldsymbol{y} + \boldsymbol{\eta}^T\boldsymbol{\lambda}^2 - \tfrac{1}{2}\boldsymbol{y}^T\boldsymbol{y}.$$

Strong duality holds and hence the equivalence between (3.35) and (3.37) follows. To prove (3.38) it suffices to observe that, letting $g_i(\boldsymbol{\nu}) = \lambda_i^2 - \boldsymbol{\nu}^T\boldsymbol{B}_i\boldsymbol{\nu}$, we have:

$$\begin{array}{rcl} l(\boldsymbol{\eta}) & = & \max\limits_{\boldsymbol{\nu}} -\dfrac{1}{2}\left(-\boldsymbol{y} - \boldsymbol{A}^T\boldsymbol{\nu}\right)^T\left(-\boldsymbol{y} - \boldsymbol{A}^T\boldsymbol{\nu}\right) - \boldsymbol{\eta}^T\boldsymbol{g}(\boldsymbol{\nu}) - \dfrac{1}{2}\boldsymbol{y}^T\boldsymbol{y}. \\ & \geq & -\dfrac{1}{2}\left(-\boldsymbol{y} - \boldsymbol{A}^T\hat{\boldsymbol{\nu}}\right)^T\left(-\boldsymbol{y} - \boldsymbol{A}^T\hat{\boldsymbol{\nu}}\right) - \boldsymbol{\eta}^T\boldsymbol{g}(\hat{\boldsymbol{\nu}}) - \dfrac{1}{2}\boldsymbol{y}^T\boldsymbol{y}. \\ & = & l(\hat{\boldsymbol{\eta}}) - (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})^T\boldsymbol{g}(\hat{\boldsymbol{\nu}}) \end{array}$$

hence,

$$l(\boldsymbol{\eta}) - l(\hat{\boldsymbol{\eta}}) \geq (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})^T (-\boldsymbol{g}(\hat{\boldsymbol{\nu}}))$$

and $-\boldsymbol{g}(\hat{\boldsymbol{\nu}})$ is a subgradient of $l(\hat{\boldsymbol{\eta}})$.

APPENDIX C

Appendix to Chapter 4

C.1 Circulant time shift model

Using circular shifts in (4.3) introduces periodicity into our model (4.2). Some types of gene expression may display periodicity, e.g. circadian transcripts, while others, e.g. transient host response, may not. For transient gene expression profiles such as the ones we are interested in here, we use a truncated version of this periodic model, where we assume that each subject's response arises from the observation of a longer periodic vector within a time window (see Figure C.1):

$$\boldsymbol{X}_s = \left[\boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right) \boldsymbol{A}_s + \boldsymbol{\epsilon}_s\right]_\Omega . \tag{C.1}$$

Here, the factors are of dimension $n_F \geq n$ and the window size is of dimension n (in the special case that $n = n_F$, we have the original periodic model). In this model, the observed features are non-periodic as long as the delays $\boldsymbol{d}^s$ are sufficiently small as compared to n_F. More concretely, if the maximum delay is $d_{\max}$, then in order to avoid wrap-around effects the dimension should be chosen as at least $n_F = n + d_{\max}$. Finally, we define the index set Ω corresponding to the observation window as:

$$\Omega = \left\{\omega^1, \omega^1 + 1, \omega^1 + 2..., \omega^2\right\}^p \tag{C.2}$$

where ω^1 and ω^2 are the lower and upper observation endpoints, verifying $n = \omega^2 - \omega^1$.

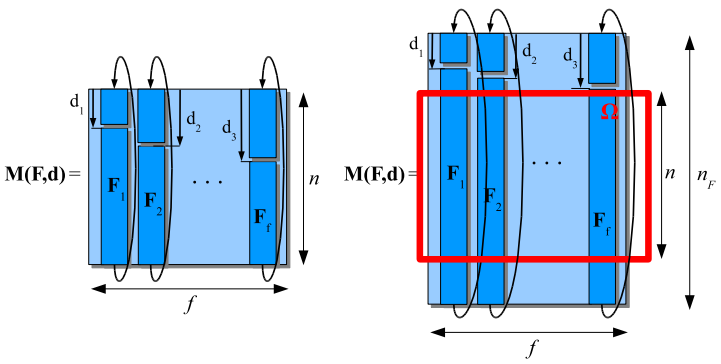


Figure C.1: *Right*: Each subject's factor matrix $\boldsymbol{M}_i$ is obtained by applying a circular shift to a common set of factors $\boldsymbol{F}$ parameterized by a vector of misalignment. $\boldsymbol{d}$. *Left*: In order to avoid wrap-around effects when modeling transient responses, we consider instead a higher dimensional truncated model of larger dimension from which we only observe the elements within a window characterized by Ω.

C.2 Delay estimation and time-course alignment

The solution to problem (4.9) yields an estimate $\hat{\boldsymbol{d}}^s$ for each subject's intrinsic factor delays. These delays are relative to the patterns found in the estimated factors and therefore require conversion to a common reference time.

For a given up-regulation or down-regulation motif, I, which we call the feature of interest, found in factor g, we choose a time point of interest t_I. See Figure C.2 (a) for an example of choice of t_I for an up-regulation feature.

Then, given t_I and for each subject s and each factor k, we define the *absolute feature occurrence time* as follows:

$$t_{s,k} = \left(\hat{\boldsymbol{d}}_k^s + t_I\right) \bmod n_F - \omega^1. \tag{C.3}$$

where $\hat{\boldsymbol{d}}_g^s$ is the estimated delay corresponding to factor k and subject s and ω^1 is the lower endpoint of the observation window (see (C.2)). Figure C.2 illustrates the computation of $t_{s,k}$ in a 2-factor example.

The quantities $t_{s,k}$ can be used for interpretation purposes or to *realign* temporal profiles in order to find common, low-variance gene expression signatures, as shown in Section 4.4.2.

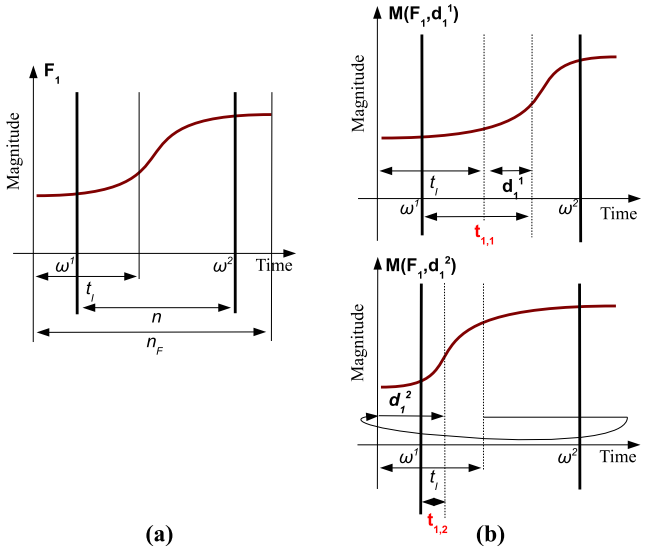


Figure C.2: (a) Time point of interest (t_I) for the up-regulation feature of factor 1. (b) The absolute time points $t_{1,1}$, $t_{1,2}$ are shown in red font for two different subjects and have been computed according to their corresponding relative delays and the formula in (C.3).

C.3 Implementation of EstimateFactors and EstimateScores

We consider here the implementation of EstimateFactors and EstimateScores under the presence of missing data. Let $\Omega_s = \left[\omega_1^s, \cdots, \omega_p^s\right] \in \{0,1\}^{n \times p}$ be the set of observed entries in observation $\boldsymbol{X}_s$. The objective in (4.9) is then:

$$\sum_{s=1}^{S} \left|\left|\left[\boldsymbol{X}_s - \boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right) \boldsymbol{A}_s\right]_{\Omega_s}\right|\right|_F^2 \tag{C.4}$$

We will show how to reformulate problems EstimateFactors (4.14) and EstimateScores (4.15) in a standard quadratic objective with linear and/or quadratic constraints.

First, we rewrite the objective (C.4) as:

$$\sum_{s=1}^{S} \sum_{j=1}^{p} \left|\left|\operatorname{diag}\left(\omega_j^s\right) \left[\boldsymbol{X}_s\right]_{\cdot,j} - \operatorname{diag}\left(\omega_j^s\right) \boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right) \left[\boldsymbol{A}_s\right]_{\cdot,j}\right|\right|_F^2 .$$

Expanding the square we obtain:

$$\begin{aligned}\sum_{s=1}^{S}\left|\left|\left[\boldsymbol{X}_s-\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^s\right)\boldsymbol{A}_s\right]_{\Omega_s}\right|\right|_F^2 &= \textstyle\sum_{s=1}^{S}\sum_{j=1}^{p}\left[\boldsymbol{A}_s\right]_{\cdot,j}^T\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^s\right)^T\operatorname{diag}\left(\omega_j^s\right)\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^s\right)\left[\boldsymbol{A}_s\right]_{\cdot,j}\\ &\quad -2\left[\boldsymbol{X}_s\right]_{\cdot,j}^T\operatorname{diag}\left(\omega_j^s\right)\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^s\right)\left[\boldsymbol{A}_s\right]_{\cdot,j}\\ &\quad +\textstyle\sum_{s=1}^{S}\left|\left|\left[\boldsymbol{X}_s\right]_{\Omega_s}\right|\right|_F^2. \end{aligned}\tag{C.5}$$

To obtain the EstimateFactors objective, first we will rewrite the OPFA model (C.1) using matrix notation. Let $\boldsymbol{U}_i$ be a circular shift matrix $\boldsymbol{U}_i$ parameterized by the i-th delay $\boldsymbol{d}$ component. Then

$$\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}\right)=\left[\boldsymbol{U}_1\boldsymbol{F}_1,\cdots,\boldsymbol{U}_f\boldsymbol{F}_f\right]=\boldsymbol{H}\tilde{\boldsymbol{F}}$$

where $\boldsymbol{F}_j$ denotes the j-th column of $\boldsymbol{F}$, $\boldsymbol{H}$ is the concatenation of the $\boldsymbol{U}_i$ matrices and $\tilde{\boldsymbol{F}}$ is a matrix containing the columns of $\boldsymbol{F}$ with the appropriate padding of zeros. With this notation and (C.5) we obtain:

$$\begin{aligned}\sum_{s=1}^{S}\left|\left|\left[\boldsymbol{X}_s-\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^s\right)\boldsymbol{A}_s\right]_{\Omega_s}\right|\right|_F^2 &\propto_{\boldsymbol{F}} \textstyle\sum_{s=1}^{S}\sum_{j=1}^{p}\operatorname{tr}\left(\left[\boldsymbol{A}_s\right]_{\cdot,j}\left[\boldsymbol{A}_s\right]_{\cdot,j}^T\tilde{\boldsymbol{F}}^T\boldsymbol{H}_s^T\operatorname{diag}\left(\omega_j^s\right)\boldsymbol{H}_s\tilde{\boldsymbol{F}}\right)\\ &\quad -2\operatorname{tr}\left(\left[\boldsymbol{A}_s\right]_{\cdot,j}\left[\boldsymbol{X}_s\right]_{\cdot,j}^T\operatorname{diag}\left(\omega_j^s\right)\boldsymbol{H}_s\tilde{\boldsymbol{F}}\right).\end{aligned}$$

We now use the identity ((NM99), Thm. 3 Sec. 4):

$$\operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{W}\right)=\operatorname{vec}\left(\boldsymbol{W}\right)^T\boldsymbol{Z}\otimes\boldsymbol{Y}^T\operatorname{vec}\left(\boldsymbol{X}\right),$$

to write:

$$\begin{aligned}\sum_{s=1}^{S}\left|\left|\left[\boldsymbol{X}_s-\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^s\right)\boldsymbol{A}_s\right]_{\Omega_s}\right|\right|_F^2 &\propto_{\boldsymbol{F}} \operatorname{vec}\left(\tilde{\boldsymbol{F}}\right)^T\left(\textstyle\sum_{s=1}^{S}\sum_{j=1}^{p}\left[\boldsymbol{A}_s\right]_{\cdot,j}\left[\boldsymbol{A}_s\right]_{\cdot,j}^T\otimes\boldsymbol{H}_s^T\operatorname{diag}\left(\omega_j^s\right)\boldsymbol{H}_s\right)\operatorname{vec}\left(\right.\\ &\quad -2\operatorname{vec}\left(\textstyle\sum_{s=1}^{S}\sum_{j=1}^{p}\boldsymbol{H}_s^T\operatorname{diag}\left(\omega_j^s\right)\left[\boldsymbol{X}_s\right]_{\cdot,j}\left[\boldsymbol{A}_s\right]_{\cdot,j}^T\right)\operatorname{vec}\left(\tilde{\boldsymbol{F}}\right).\end{aligned}$$

Finally, making use of the fact that $\tilde{\boldsymbol{F}}$ is a block-column matrix with the columns of $\boldsymbol{F}$ padded by zeros, we conclude:

$$\sum_{s=1}^{S}\left|\left|\left[\boldsymbol{X}_s-\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^s\right)\boldsymbol{A}_s\right]_{\Omega_s}\right|\right|_F^2 \propto_{\boldsymbol{F}} \operatorname{vec}\left(\boldsymbol{F}\right)^T\boldsymbol{Q}_F\operatorname{vec}\left(\boldsymbol{F}\right)-2\boldsymbol{q}_F^T\operatorname{vec}\left(\boldsymbol{F}\right)$$

where we have defined

$$\boldsymbol{Q}_F = \left[\sum_{s=1}^{S}\sum_{j=1}^{p}\left[\boldsymbol{A}_s\right]_{\cdot,j}\left[\boldsymbol{A}_s\right]_{\cdot,j}^T \otimes \boldsymbol{H}_s^T \mathrm{diag}\left(\omega_j^s\right)\boldsymbol{H}_s\right]_{\mathcal{J},\mathcal{J}} \tag{C.6}$$

$$\boldsymbol{q}_F = \left[\mathrm{vec}\left(\sum_{s=1}^{S}\sum_{j=1}^{p}\boldsymbol{H}_s^T \mathrm{diag}\left(\omega_j^s\right)\left[\boldsymbol{X}_s\right]_{\cdot,j}\left[\boldsymbol{A}_s\right]_{\cdot,j}^T\right)\right]_{\mathcal{J}} \tag{C.7}$$

and $\mathcal{J}$ are the indices corresponding to the non-zero elements in $\mathrm{vec}\left(\tilde{\boldsymbol{F}}\right)$. Hence, EstimateFactors can be written as:

$$\min_{\boldsymbol{F}} \quad \mathrm{vec}\left(\boldsymbol{F}\right)^T\left(\boldsymbol{Q}_F + \beta\mathrm{diag}\left(\boldsymbol{W}^T\boldsymbol{W},\cdots,\boldsymbol{W}^T\boldsymbol{W}\right)\right)\mathrm{vec}\left(\boldsymbol{F}\right) - 2\boldsymbol{q}_F^T\mathrm{vec}\left(\boldsymbol{F}\right) \tag{C.8}$$

$$\text{s.t.} \qquad \begin{cases} ||\boldsymbol{F}||_{\mathcal{F}}^2 \leq \delta & \\ \boldsymbol{F}_{i,j} \geq 0 & i = 1,\cdots,n, \\ & j = 1,\cdots,f \end{cases}$$

The dimension of the variables in this problem is nf. In the applications considered here, both n and f are relatively small and hence this program can be solved with a standard convex solver such as SeDuMi (Stu99) (upon conversion to a standard conic problem), or the Projected Newton method (Ber82) on the dual problem that we describe next. To alleviate the notation, we will consider the following problem, which has essentially the same structure as (C.8):

$$\begin{aligned} \min_{\boldsymbol{f}} \quad & \tfrac{1}{2}\boldsymbol{f}^T\boldsymbol{Q}\boldsymbol{f} + \boldsymbol{q}^T\boldsymbol{f} \\ \text{s.t.} \quad & \boldsymbol{f}^T\boldsymbol{f} \leq \delta \\ & \boldsymbol{f} \succeq 0 \end{aligned}$$

where $\succeq$ denotes an element-wise inequality (also the generalized inequality corresponding to the positive orthant cone). The lagrange dual function of this differentiable, convex optimization problem is given by:

$$w\left(\boldsymbol{\eta},\epsilon\right) = \inf_{\boldsymbol{f}} \frac{1}{2}\boldsymbol{f}^T\left(\boldsymbol{Q} + 2\epsilon\boldsymbol{I}\right)\boldsymbol{f} + \left(\boldsymbol{q}-\boldsymbol{\eta}\right)^T\boldsymbol{f} - \epsilon\delta$$

If $\boldsymbol{q}-\boldsymbol{\eta} \in \mathcal{R}\left(bQ + 2\epsilon\boldsymbol{I}\right)$, the infimum is finite, and we obtain:

$$w\left(\boldsymbol{\eta},\epsilon\right) = \begin{cases} -\frac{1}{2}\left(\boldsymbol{q}-\boldsymbol{\eta}\right)^T\left(\boldsymbol{Q}+2\epsilon\boldsymbol{I}\right)^{-1}\left(\boldsymbol{q}-\boldsymbol{\eta}\right) - \epsilon\delta & \boldsymbol{q}-\boldsymbol{\eta} \in \mathcal{R}\left(\boldsymbol{Q}+2\epsilon\boldsymbol{I}\right) \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem is hence:

$$\begin{aligned} \min \quad & -\tfrac{1}{2}\left(\boldsymbol{q}-\boldsymbol{\eta}\right)^{T}\left(\boldsymbol{Q}+2\epsilon\boldsymbol{I}\right)^{-1}\left(\boldsymbol{q}-\boldsymbol{\eta}\right)-\epsilon\delta \\ \text{s.t.} \quad & \epsilon \geq 0 \\ & \boldsymbol{\eta} \succeq 0 \\ & \boldsymbol{q}-\boldsymbol{\eta} \in \mathcal{R}\left(\boldsymbol{Q}+2\epsilon\boldsymbol{I}\right) \end{aligned}$$

In general, $\boldsymbol{Q}$ is positive definite in our problems, due to the quadratic part of the smoothness penalty ($\beta\text{diag}\left(\boldsymbol{W}^{T}\boldsymbol{W},\cdots,\boldsymbol{W}^{T}\boldsymbol{W}\right)$). Hence, we can drop the range constraint and solve:

$$\begin{aligned} \min \quad & -\tfrac{1}{2}\left(\boldsymbol{q}-\boldsymbol{\eta}\right)^{T}\left(\boldsymbol{Q}+2\epsilon\boldsymbol{I}\right)^{-1}\left(\boldsymbol{q}-\boldsymbol{\eta}\right)-\epsilon\delta \\ \text{s.t.} \quad & \epsilon \geq 0 \\ & \boldsymbol{\eta} \succeq 0 \end{aligned} \tag{C.9}$$

Through a simple projected Newton method (Ber82) which only requires evaluating the Hessian and gradient of the objective in (C.9).

On the other hand, we can follow a similar procedure to reformulate the objective in EstimateScores (4.15) into a penalized quadratic form. First we use (C.5) and (C.6) to write:

$$\sum_{s=1}^{S}\left|\left|\left[\boldsymbol{X}_{s}-\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^{s}\right)\boldsymbol{A}_{s}\right]_{\Omega_{s}}\right|\right|_{F}^{2} \quad \propto_{\{\boldsymbol{A}_{s}\}_{i=1}^{S}} \sum_{s=1}^{S}\text{vec}\left(\boldsymbol{A}_{s}\right)^{T}\boldsymbol{Q}_{A}^{s}\text{vec}\left(\boldsymbol{A}_{s}\right)-2\boldsymbol{q}_{A}^{s\,T}\text{vec}\left(\boldsymbol{A}_{s}\right)$$

where

$$\boldsymbol{Q}_{A}^{s}=\begin{bmatrix} \boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^{s}\right)^{T}\text{diag}\left(\omega_{1}^{s}\right)\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^{s}\right) & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^{s}\right)^{T}\text{diag}\left(\omega_{p}^{s}\right)\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^{s}\right) \end{bmatrix} \tag{C.10}$$

$$\boldsymbol{q}_{A}^{s}=\begin{bmatrix} \left[\boldsymbol{X}_{s}\right]_{\cdot,1}^{T}\text{diag}\left(\omega_{1}^{s}\right)\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^{s}\right) & \cdots & \left[\boldsymbol{X}_{s}\right]_{\cdot,p}^{T}\text{diag}\left(\omega_{p}^{s}\right)\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^{s}\right) \end{bmatrix} \tag{C.11}$$

Thus EstimateScores can be written as:

$$\begin{aligned} \min_{\boldsymbol{F}} \quad & \sum_{s=1}^{S}\text{vec}\left(\boldsymbol{A}_{s}\right)^{T}\boldsymbol{Q}_{A}^{s}\text{vec}\left(\boldsymbol{A}_{s}\right)-2\boldsymbol{q}_{A}^{s\,T}\text{vec}\left(\boldsymbol{A}_{s}\right)+\lambda\sum_{i=1}^{p}\sum_{j=1}^{f}\left\|\left[\boldsymbol{A}_{1}\right]_{j,i}\cdots\left[\boldsymbol{A}_{S}\right]_{j,i}\right\|_{2} \\ \text{s.t.} \quad & \begin{cases} \left|\left|\boldsymbol{F}\right|\right|_{\mathcal{F}}^{2}\leq\delta \\ \boldsymbol{F}_{i,j}\geq 0 \quad i=1,\cdots,n, \\ \qquad\qquad j=1,\cdots,f \end{cases} \end{aligned} \tag{C.12}$$

This is a convex, non-differentiable and potentially high-dimensional problem. For this type of optimization problems, there exists a class of simple and scalable algorithms which has recently received much attention (ZE10), (DDDM04b), (CW06), (PCP08). These algorithms rely only on first-order updates of the type:

$$\boldsymbol{x}^t \leftarrow \mathcal{T}_{\Gamma,\lambda}\left(\boldsymbol{v} - 2\boldsymbol{x}^{t-1}\left(\alpha\boldsymbol{I} - \boldsymbol{Q}\right)\right), \tag{C.13}$$

which only involves matrix-vector multiplications and evaluation of the operator $\mathcal{T}$, which is called the proximal operator (CW06) associated to Γ and $\mathcal{C}$ and is defined as:

$$\begin{aligned}\mathcal{T}_{\Gamma,\lambda}\left(\boldsymbol{v}\right) \quad &:= \min \quad \frac{1}{2}\boldsymbol{x}'\boldsymbol{x} + \boldsymbol{v}'\boldsymbol{x} + \lambda\Gamma\left(\boldsymbol{x}\right) \\ &\quad\ \text{s.t.} \quad\ \boldsymbol{x} \in \mathcal{C}.\end{aligned}$$

This operator takes the vector $\boldsymbol{v}$ as an input and outputs a shrunk/thresholded version of it depending on the nature of the penalty Γ and the constraint set $\mathcal{C}$. For some classes of penalties Γ (e.g. l_1, l_2, mixed $l_1 - l_2$) and the positivity constraints considered here, this operator has a closed form solution (TPWOH09), (CW06). Weak convergence of the sequence (C.13) to the optimum of (C.8) is assured for a suitable choice of the constant α (PCP08), (BT09).

C.4 Delay Estimation lower bound in the presence of Missing Data

As we mentioned earlier, the lower bound (4.23) does not hold anymore under the presence of missing data. We derive here another bound that can be used in such case. From expression (C.5), we first obtain the objective in EstimateDelays (4.17) in a quadratic form:

$$\begin{aligned}\sum_{s=1}^{S}\left|\left|\left[\boldsymbol{X}_s - \boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^s\right)\boldsymbol{A}_s\right]_{\Omega_s}\right|\right|_F^2 \quad &\propto_{\boldsymbol{d}^s} \sum_{s=1}^{S}\sum_{j=1}^{p}\operatorname{tr}\left(\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^s\right)^T\operatorname{diag}\left(\omega_j^s\right)\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^s\right)\boldsymbol{P}_{s,j}\right) \\ &\qquad -2\operatorname{tr}\left(\boldsymbol{Q}_{s,j}\boldsymbol{M}\left(\boldsymbol{F},\boldsymbol{d}^s\right)\right).\end{aligned}$$

Where we have let $\boldsymbol{P}_{s,j} := \left[\boldsymbol{A}_s\right]_{\cdot,j}\left[\boldsymbol{A}_s\right]_{\cdot,j}^T$ and $\boldsymbol{Q}_{s,j} = \left[\boldsymbol{A}_s\right]_{\cdot,j}\left[\boldsymbol{X}_s\right]_{\cdot,j}^T\operatorname{diag}\left(\omega_j^s\right)$. Notice that each of the terms indexed by s is independent of the others. Using (C.6), we

obtain

$$\begin{aligned}\left|\left|[\boldsymbol{X}_s - \boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\boldsymbol{A}_s]_{\Omega_s}\right|\right|_F^2 \;&\propto \operatorname{vec}\left(\boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\right)^T \sum_{j=1}^p \left(\boldsymbol{P}_{s,j} \otimes \operatorname{diag}\left(\omega_j^s\right)\right) \operatorname{vec}\left(\boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\right) \\ &\quad -2\operatorname{vec}\left(\sum_{j=1}^p \boldsymbol{Q}_{s,j}^T\right)^T \operatorname{vec}\left(\boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\right).\end{aligned}$$

We now can minorize the function above by:

$$\begin{aligned}\left|\left|[\boldsymbol{X}_s - \boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\boldsymbol{A}_s]_{\Omega_s}\right|\right|_F^2 \;&\geq \underline{\lambda}\left(\sum_{j=1}^p \boldsymbol{P}_{s,j} \otimes \operatorname{diag}\left(\omega_j^s\right)\right) \operatorname{vec}\left(\boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\right)^T \operatorname{vec}\left(\boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\right) \\ &\quad -2\operatorname{vec}\left(\sum_{j=1}^p \boldsymbol{Q}_{s,j}^T\right)^T \operatorname{vec}\left(\boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\right) + \sum_{s=1}^S \left|\left|[\boldsymbol{X}_s]_{\Omega_s}\right|\right|_F^2\end{aligned}$$

which leads to:

$$\begin{aligned}\left|\left|[\boldsymbol{X}_s - \boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\boldsymbol{A}_s]_{\Omega_s}\right|\right|_F^2 \;&\geq \underline{\lambda}\left(\sum_{j=1}^p \boldsymbol{P}_{s,j} \otimes \operatorname{diag}\left(\omega_j^s\right)\right) ||\boldsymbol{F}||_F^2 \\ &\quad -2\operatorname{vec}\left(\sum_{j=1}^p \boldsymbol{Q}_{s,j}^T\right)^T \operatorname{vec}\left(\boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\right) + \left|\left|[\boldsymbol{X}_s]_{\Omega_s}\right|\right|_F^2.\end{aligned}$$

Using the relaxation in (4.21), we can now compute a lower bound $\Phi_{lb}\left(\mathcal{I}_t\right)$ as

$$\begin{aligned}\Phi_{lb}\left(\mathcal{I}_t\right) \;&= \underline{\lambda}\left(\sum_{j=1}^p \boldsymbol{P}_{s,j} \otimes \operatorname{diag}\left(\omega_j^s\right)\right) ||\boldsymbol{F}||_F^2 + \left|\left|[\boldsymbol{X}_s]_{\Omega_s}\right|\right|_F^2 \\ &\quad -2\max_{\boldsymbol{d}\in\mathcal{S}_t} \operatorname{vec}\left(\sum_{j=1}^p \boldsymbol{Q}_{s,j}^T\right)^T \operatorname{vec}\left(\boldsymbol{M}\left(\boldsymbol{F}, \boldsymbol{d}^s\right)\right).\end{aligned}$$

APPENDIX D

Appendix to Chapter 5

D.1 Derivation of the MLE estimate for the signal-to-noise ratio of each component

We will only consider the case $F = 1$. Since $l\left(\boldsymbol{h}^{\text{MisPCA}}, \boldsymbol{d}^{\text{MisPCA}}, \boldsymbol{\sigma}\right)$ is separable with respect to σ_f, the general case $F > 1$ follows immediately from the case $F = 1$.

It is easy to verify that if $\lambda^{\text{MisPCA}} < 1$, then $l\left(\boldsymbol{h}^{\text{MisPCA}}, \boldsymbol{d}^{\text{MisPCA}}, \sigma_1\right)$ is monotonically decreasing over $\sigma_1 \geq 0$. Otherwise, it has a positive stationary point at:

$$\sigma_1^{\circ} = \lambda^{\text{MisPCA}} - 1.$$

The second derivative of $l\left(\boldsymbol{h}^{\text{MisPCA}}, \boldsymbol{d}^{\text{MisPCA}}, \sigma_1\right)$ with respect to σ_1 is negative at σ_1°, hence σ_1° is at least a local maxima. It is easy to check that $l\left(\boldsymbol{h}^{\text{MisPCA}}, \boldsymbol{d}^{\text{MisPCA}}, \sigma_1\right)$ is strictly increasing over $0 \leq \sigma_1 < \sigma_1^{\circ}$ and strictly decreasing over $\sigma_1^{\circ} > \sigma_1$, thus the local maxima is also a global maxima. This finalizes the proof of (5.8).

D.2 Proof of Theorem V.1

The eigenvalue decomposition of $\boldsymbol{\Sigma}(\boldsymbol{\tau})$ is denoted by $\boldsymbol{Q}_{\Sigma}\boldsymbol{\Delta}_{\Sigma}\boldsymbol{Q}_{\Sigma}^{T}$, where $\boldsymbol{Q}_{\Sigma}$ is a unitary matrix containing its eigenvectors and $\boldsymbol{\Delta}_{\Sigma}$ is a diagonal matrix containing its eigenvalues:

$$[\boldsymbol{\Delta}_{\Sigma}]_{i,i} = \begin{cases} \text{SNR } \lambda_i\left(\boldsymbol{\mathcal{H}}\text{diag}\left(\boldsymbol{s}\left(\boldsymbol{d}_{-p}\boldsymbol{\tau}\right) \otimes \bar{\boldsymbol{\sigma}}\right)\boldsymbol{\mathcal{H}}^{T}\right) + 1 & 1 \leq i \leq r \\ 1 & r < i \leq p \end{cases}$$

where $r = \text{rank}\left(\boldsymbol{\mathcal{H}}\text{diag}\left(\boldsymbol{s}\left(\boldsymbol{d}-_p\boldsymbol{\tau}\right)\otimes\bar{\boldsymbol{\sigma}}\right)\boldsymbol{\mathcal{H}}^T\right) \leq 2d_{\max}$. A well-known property of the eigenvalues of Grammian matrices allows us to conclude that $\lambda_i\left(\boldsymbol{\mathcal{H}}\text{diag}\left(\boldsymbol{s}\left(\boldsymbol{d}-_p\boldsymbol{\tau}\right)\otimes\bar{\boldsymbol{\sigma}}\right)\boldsymbol{\mathcal{H}}^T\right)$ is equal to:

$$\lambda_i\left(\text{diag}\left(\boldsymbol{s}\left(\boldsymbol{d}-_p\boldsymbol{\tau}\right)\otimes\bar{\boldsymbol{\sigma}}\right)^{\frac{1}{2}}\boldsymbol{R}_H\text{diag}\left(\boldsymbol{s}\left(\boldsymbol{d}-_p\boldsymbol{\tau}\right)\otimes\bar{\boldsymbol{\sigma}}\right)^{\frac{1}{2}}\right)$$

where we have used the definition of $\boldsymbol{R}_H$ from (5.13).

Using properties of the Gaussian distribution, we can write:

$$\boldsymbol{S}\left(\boldsymbol{\tau}\right) = \boldsymbol{Q}_\Sigma\tilde{\boldsymbol{S}}\boldsymbol{Q}_\Sigma^T.$$

where $\tilde{\boldsymbol{S}} = \frac{\boldsymbol{Z}\boldsymbol{Z}^T}{n}$ and each column of the $p\times n$ matrix $\boldsymbol{Z}$ follows a zero-mean multivariate Gaussian distribution with covariance $\boldsymbol{\Delta}_\Sigma$. The result for $\lambda_1\left(\boldsymbol{S}\left(\boldsymbol{\tau}\right)\right)$ follows from observing that $\lambda_f\left(\boldsymbol{S}\left(\boldsymbol{\tau}\right)\right) = \lambda_f\left(\tilde{\boldsymbol{S}}\right)$ and applying Theorems 1 and 2 from (Pau07). The result concerning $\boldsymbol{v}_f\left(\boldsymbol{S}\left(\boldsymbol{\tau}\right)\right)$ follows from observing that:

$$\begin{aligned}\left\langle\boldsymbol{v}_f\left(\boldsymbol{S}\left(\boldsymbol{\tau}\right)\right),\boldsymbol{v}_f\left(\boldsymbol{\Sigma}\left(\boldsymbol{\tau}\right)\right)\right\rangle &= \left\langle\boldsymbol{Q}_\Sigma\boldsymbol{v}_f\left(\tilde{\boldsymbol{S}}\right),\boldsymbol{v}_f\left(\boldsymbol{\Sigma}\left(\boldsymbol{\tau}\right)\right)\right\rangle \\ &= \left\langle\boldsymbol{v}_f\left(\tilde{\boldsymbol{S}}\right),\boldsymbol{e}_f\right\rangle,\end{aligned}$$

where $\boldsymbol{e}_f$ denotes the vector of all zeros except for a 1 in the f-th coordinate, and applying Theorem 4 from (Pau07). See (BGN11) for an alternative derivation and insight into the origin of the phase transition.

BIBLIOGRAPHY

BIBLIOGRAPHY

[ABDF11] M.V. Afonso, J.M. Bioucas-Dias, and M.A.T. Figueiredo, *An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems*, Image Processing, IEEE Transactions on (2011), no. 99, 1–1.

[AEB06] M. Aharon, M. Elad, and A. Bruckstein, *K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation*, IEEE Transactions on signal processing **54** (2006), no. 11, 4311–4322.

[AU00] A. Aderem and R.J. Ulevitch, *Toll-like receptors in the induction of the innate immune response*, Nature **406** (2000), no. 6797, 782–787.

[Bac08] F.R. Bach, *Bolasso: model consistent lasso estimation through the bootstrap*, Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 33–40.

[Bac10] F. Bach, *Structured sparsity-inducing norms through submodular functions*, Advances in Neural Information Processing Systems 23 (J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, eds.), 2010, pp. 118–126.

[BBAP05] J. Baik, G. Ben Arous, and S. Péché, *Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices*, Annals of Probability **33** (2005), no. 5, 1643–1697.

[BBC09] S. Becker, J. Bobin, and E. Candes, *Nesta: A fast and accurate first-order method for sparse recovery*, Arxiv preprint arXiv:0904.3367 (2009).

[BD05] T. Blumensath and M. Davies, *Sparse and shift-invariant representations of music*, Audio, Speech, and Language Processing, IEEE Transactions on **14** (2005), no. 1, 50–57.

[BD06] ———, *Sparse and shift-invariant representations of music*, IEEE Transactions on Speech and Audio Processing **14** (2006), no. 1, 50.

[Ber82] D.P. Bertsekas, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control and Optimization **20** (1982), no. 2.

[Ber99] ———, *Nonlinear programming*, Athena Scientific, 1999.

[BGN11] F. Benaych-Georges and R. R. Nadakuditi, *The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices*, Advances in Mathematics **227** (2011), no. 1, 494 – 521.

[BIAS03] B.M. Bolstad, R.A. Irizarry, M. Astrand, and T.P. Speed, *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias*, Bioinformatics **19** (2003), no. 2, 185.

[BJ04] Z. Bar-Joseph, *Analyzing time series gene expression data*, Bioinformatics **20** (2004), no. 16, 2493.

[BL06] J.M. Borwein and A.S. Lewis, *Convex analysis and nonlinear optimization: Theory and examples*, vol. 3, Springer Verlag, 2006.

[BL10] G. Bergqvist and E.G. Larsson, *The higher-order singular value decomposition: Theory and an application [lecture notes]*, Signal Processing Magazine, IEEE **27** (2010), no. 3, 151 –154.

[BMG10] J.A. Bazerque, G. Mateos, and G.B. Giannakis, *Group-Lasso on Splines for Spectrum Cartography*, Arxiv preprint arXiv:1010.0274 (2010).

[BNO+03] D.P. Bertsekas, A. Nedi, A.E. Ozdaglar, et al., *Convex analysis and optimization*.

[BT09] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci **2** (2009), 183–202.

[BV] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press.

[Can06] E.J. Candès, *Compressive sampling*, Proceedings of the International Congress of Mathematicians, vol. 3, Citeseer, 2006, pp. 1433–1452.

[CB83] R. Chamberlain and J. Bridle, *ZIP: A dynamic programming algorithm for time-aligning two indefinitely long utterances*, Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83., vol. 8, IEEE, 1983, pp. 816–819.

[CC70] J.D. Carroll and J.J. Chang, *Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition*, Psychometrika **35** (1970), no. 3, 283–319.

[Com02] P. Comon, *Tensor decompositions*, Mathematics in signal processing V, Oxford University Press, USA, 2002.

[CT07] E. Candes and T. Tao, *The dantzig selector: Statistical estimation when p is much larger than n*, The Annals of Statistics **35** (2007), no. 6, 2313–2351.

[CW06] P.L. Combettes and V.R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Modeling and Simulation **4** (2006), no. 4, 1168–1200.

[DDDM04a] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math **57** (2004), no. 11, 1413–1457.

[DDDM04b] ———, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Communications on Pure and Applied Mathematics **57** (2004), no. 11, 1413–1457.

[dEGJL07] A. d'Aspremont, L. El Ghaoui, M.I. Jordan, and G.R.G. Lanckriet, *A Direct Formulation for Sparse PCA Using Semidefinite Programming*, SIAM Review **49** (2007), no. 3, 434–448.

[DLDMV00] L. De Lathauwer, B. De Moor, and J. Vandewalle, *A multilinear singular value decomposition*, SIAM Journal on Matrix Analysis and Applications **21** (2000), no. 4, 1253–1278.

[Don00] D.L. Donoho, *High-dimensional data analysis: The curses and blessings of dimensionality*, AMS Math Challenges Lecture (2000), 1–32.

[DS04] D. Donoho and V. Stodden, *When does non-negative matrix factorization give a correct decomposition into parts*, Advances in neural information processing systems **16** (2004), 1141–1148.

[Dun80] J.C. Dunn, *Newton's Method and the Goldstein Step-Length Rule for Constrained Minimization Problems*, SIAM Journal on Control and Optimization **18** (1980), 659.

[EGL97] L. El Ghaoui and H. Lebret, *Robust solutions to least squares problems with uncertain data*, SIAM Journal Matrix Analysis and Applications (October 1997).

[EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, The Annals of statistics **32** (2004), no. 2, 407–499.

[Fer92] P.J.S.G. Ferreira, *Localization of the eigenvalues of Toeplitz matrices using additive decomposition, embedding in circulants, and the Fourier transform*, Proc. 10th IFAC Symposium on System Identification SYSID'94, Copenhagen, Citeseer, 1992.

[FFL08] J. Fan, Y. Fan, and J. Lv, *High dimensional covariance matrix estimation using a factor model*, Journal of Econometrics **147** (2008), no. 1, 186–197.

[FGR+06] B. Fischer, J. Grossmann, V. Roth, W. Gruissem, S. Baginsky, and J.M. Buhmann, *Semi-supervised LC/MS alignment for differential proteomics*, Bioinformatics **22** (2006), no. 14, e132.

[Fis36] R.A. Fisher, *The use of multiple measurements in taxonomic problems*, Annals of Eugenics **7** (1936), no. 2, 179–188.

[Fri01] J.H. Friedman, *The role of statistics in the data revolution?*, International Statistical Review **69** (2001), no. 1, 5–10.

[GST+99] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*, science **286** (1999), no. 5439, 531.

[HJ90] R.A. Horn and C.R. Johnson, *Matrix analysis*, Cambridge Univ Pr, 1990.

[Hot33] H. Hotelling, *Analysis of a complex of statistical variables into principal components*, Journal of educational psychology **24** (1933), no. 6, 417–441.

[HTF05] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, Springer, 2005.

[HZR+11] Y. Huang, A. K. Zaas, A. Rao, N. Dobigeon, P. J. Woolf, T. Veldman, N. C. Oien, M. T. McClain, J. B. Varkey, B. Nicholson, L. Carin, S Kingsmore, C. W. Woods, G. S. Ginsburg, and A. O. Hero, III, *Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection*, PLoS Genetics **7** (2011).

[IHC+03] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed, *Exploration, normalization, and summaries of high density oligonucleotide array probe level data*, Biostatistics **4** (2003), no. 2, 249.

[JAB09] R. Jenatton, J.Y. Audibert, and F. Bach, *Structured variable selection with sparsity-inducing norms*, Arxiv preprint arXiv:0904.3523 (2009).

[JL08] I.M. Johnstone and A.Y. Lu, *Sparse principal components analysis*, J. Amer. Statist. Assoc (2008).

[JL09] ———, *On consistency and sparsity for principal components analysis in high dimensions*, Journal of the American Statistical Association **104** (2009), no. 486, 682–693.

[JMOB10] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, *Proximal methods for sparse hierarchical dictionary learning*, Proceedings of the International Conference on Machine Learning (ICML), Citeseer, 2010.

[JOB10] R. Jenatton, G. Obozinski, and F. Bach, *Structured Sparse Principal Component Analysis*, Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, vol. 9, 2010.

[JOV09] L. Jacob, G. Obozinski, and J.P. Vert, *Group lasso with overlap and graph lasso*, Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 433–440.

[JSD10] Y. Jia, M. Salzmann, and T. Darrell, *Factorized latent spaces with structured sparsity*, Tech. Report UCB/EECS-2010-99, EECS Department, University of California, Berkeley, Jun 2010.

[KB09] T.G. Kolda and B.W. Bader, *Tensor decompositions and applications*, SIAM Review **51** (2009), no. 3, 455–500.

[KDMR+03] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T.W. Lee, and T.J. Sejnowski, *Dictionary learning algorithms for sparse representation*, Neural computation **15** (2003), no. 2, 349–396.

[Kee10] R.W. Keener, *Theoretical statistics: Topics for a core course*, Springer Verlag, 2010.

[LBC+06] X. Li, F. Bian, M. Crovella, C. Diot, R. Govindan, G. Iannaccone, and A. Lakhina, *Detection and identification of network anomalies using sketch subspaces*, Proceedings of the 6th ACM SIGCOMM conference on Internet measurement, ACM, 2006, pp. 147–152.

[LCD04] A. Lakhina, M. Crovella, and C. Diot, *Diagnosing network-wide traffic anomalies*, ACM SIGCOMM Computer Communication Review, vol. 34, ACM, 2004, pp. 219–230.

[LS99a] D.D. Lee and H.S. Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature **401** (1999), no. 6755, 788–791.

[LS99b] M.S. Lewicki and T.J. Sejnowski, *Coding time-varying signals using sparse, shift-invariant representations*, Advances in neural information processing systems (1999), 730–736.

[LVBL98] M.S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, *Applications of second-order cone programming*, Linear Algebra and its Applications **284** (1998), no. 1-3, 193–228.

[LW66] E.L. Lawler and D.E. Wood, *Branch-and-bound methods: A survey*, Operations research **14** (1966), no. 4, 699–719.

[LY10] J. Liu and J. Ye, *Fast overlapping group lasso*, Arxiv preprint arXiv:1009.0306 (2010).

[Mas07] P. Massart, *Concentration inequalities and model selection*, Lecture Notes in Mathematics **1896** (2007).

[MB06] N. Meinshausen and P. Buhlmann, *High-dimensional graphs and variable selection with the lasso*, The Annals of Statistics **34** (2006), no. 3, 1436–1462.

[MBP+10] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, *Non-local sparse models for image restoration*, Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2010, pp. 2272–2279.

[MCW05] D. Malioutov, M. Cetin, and AS Willsky, *A sparse signal reconstruction perspective for source localization with sensor arrays*, IEEE Transactions on Signal Processing **53** (2005), no. 8 Part 2, 3010–3022.

[MDCM02] E. Moulines, P. Duhamel, J.F. Cardoso, and S. Mayrargue, *Subspace methods for the blind identification of multichannel FIR filters*, Signal Processing, IEEE Transactions on **43** (2002), no. 2, 516–525.

[MLG+08] B. Mailhé, S. Lesage, R. Gribonval, F. Bimbot, and P. Vandergheynst, *Shift-invariant dictionary learning for sparse representations: extending K-SVD*, Proceedings of the 16th European Signal Processing Conference, vol. 4, 2008.

[Mor65] J.J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France **93** (1965), no. 2, 273–299.

[MVVR10] S. Mosci, S. Villa, A. Verri, and L. Rosasco, *A primal-dual algorithm for group sparse regularization with overlapping groups*, Neural Information Processing Systems, 2010.

[Nat95] B.K. Natarajan, *Sparse approximate solutions to linear systems*, SIAM journal on computing **24** (1995), 227.

[NF01] R.D. Nowak and M.A.T. Figueiredo, *Fast wavelet-based image deconvolution using the EM algorithm*, Conference Record of the 35th Asilomar Conference, vol. 1, 2001.

[NM96] Boon C. N. and Chong M.S.S., *Sensor-array calibration using a maximum-likelihood approach*, Antennas and Propagation, IEEE Transactions on **44** (1996), no. 6, 827 –835.

[NM99] H. Neudecker and J.R. Magnus, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 1999.

[NRWY10] S. Negahban, P. Ravikumar, M.J. Wainwright, and B. Yu, *A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers*, Arxiv preprint arXiv:1010.2731 (2010).

[NW08] S. Negahban and M.J. Wainwright, *Joint support recovery under high-dimensional scaling: Benefits and perils of l1,inf-regularization*, Advances in Neural Information Processing Systems **21** (2008), 1161–1168.

[NWK+96] J. Neter, W. Wasserman, M.H. Kutner, et al., *Applied linear statistical models*, Irwin Burr Ridge, Illinois, 1996.

[OF97] B.A. Olshausen and D.J. Field, *Sparse coding with an overcomplete basis set: A strategy employed by V1?*, Vision research **37** (1997), no. 23, 3311–3325.

[OP09] A.B. Owen and P.O. Perry, *Bi-cross-validation of the SVD and the non-negative matrix factorization*, The Annals of Applied Statistics **3** (2009), no. 2, 564–594.

[OPT00] M.R. Osborne, B. Presnell, and B.A. Turlach, *A new approach to variable selection in least squares problems*, IMA journal of numerical analysis **20** (2000), no. 3, 389.

[OWJ08] G. Obozinski, M.J. Wainwright, and M.I. Jordan, *High-dimensional union support recovery in multivariate regression*, Advances in Neural Information Processing Systems **22** (2008).

[Pau07] D. Paul, *Asymptotics of sample eigenstructure for a large dimensional spiked covariance model*, Statistica Sinica **17** (2007), no. 4, 1617.

[PCP08] N. Pustelnik, C. Chaux, and J.C. Pesquet, *A constrained forward-backward algorithm for image recovery problems*, Proceedings of the 16th European Signal Processing Conference, 2008, pp. 25–29.

[Pea01] K. Pearson, *Liii. on lines and planes of closest fit to systems of points in space*, Philosophical Magazine Series 6 **2** (1901), no. 11, 559–572.

[PHIP+03] N. Patwari, A.O. Hero III, M. Perkins, N.S. Correal, and R.J. O'dea, *Relative location estimation in wireless sensor networks*, Signal Processing, IEEE Transactions on **51** (2003), no. 8, 2137–2148.

[PK93] R.W. Picard and T. Kabir, *Finding similar patterns in large image databases*, vol. 5, apr. 1993, pp. 161 –164 vol.5.

[Pop02] K.R. Popper, *The logic of scientific discovery*, Psychology Press, 2002.

[Ram97] J.O. Ramsay, *Functional data analysis*, Encyclopedia of Statistical Sciences (1997).

[SGB00] N.D. Sidiropoulos, G.B. Giannakis, and R. Bro, *Blind PARAFAC receivers for DS-CDMA systems*, IEEE Transactions on Signal Processing **48** (2000), no. 3, 810–823.

[SK00] B.M. Sadler and R.J. Kozick, *Bounds on uncalibrated array signal processing*, Statistical Signal and Array Processing, 2000. Proceedings of the Tenth IEEE Workshop on, IEEE, 2000, pp. 73–77.

[SL06] Leif Srnmo and Pablo Laguna, *Electrocardiogram (ecg) signal processing*, John Wiley & Sons, Inc., 2006.

[SLMB07] L. Sacchi, C. Larizza, P. Magni, and R. Bellazzi, *Precedence Temporal Networks to represent temporal relationships in gene expression data*, Journal of Biomedical Informatics **40** (2007), no. 6, 761–774.

[SRJ05] N. Srebro, J.D.M. Rennie, and T.S. Jaakkola, *Maximum-margin matrix factorization*, Advances in neural information processing systems **17** (2005), 1329–1336.

[SRSE10a] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. Eldar, *C-hilasso: A collaborative hierarchical sparse modeling framework*, Arxiv preprint arXiv:1006.1346 (2010).

[SRSE10b] ———, *Collaborative hierarchical sparse modeling*, Information Sciences and Systems (CISS), 2010 44th Annual Conference on, IEEE, 2010, pp. 1–6.

[Stu99] J.F. Sturm, *Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones*, Optimization methods and software **11** (1999), no. 1, 625–653.

[TAJ77] A.N. Tikhonov, V.Y. Arsenin, and F. John, *Solutions of ill-posed problems*, VH Winston Washington, DC, 1977.

[TB99] M.E. Tipping and C.M. Bishop, *Probabilistic principal component analysis*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61** (1999), no. 3, 611–622.

[TBM79] H.L. Taylor, SC Banks, and JF McCoy, *Deconvolution with the l 1 norm*, Geophysics **44** (1979), no. 1, 39.

[THNC02] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, *Diagnosis of multiple cancer types by shrunken centroids of gene expression*, Proceedings of the National Academy of Sciences **99** (2002), no. 10, 6567.

[THNC03] ———, *Class prediction by nearest shrunken centroids, with applications to DNA microarrays*, Statistical Science (2003), 104–117.

[Tho18] W. M. Thorburn, *The myth of occam's razor*, Mind **XXVII** (1918), no. 3, 345–353.

[Tib96] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) (1996), 267–288.

[TP07] F. Tai and W. Pan, *Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data*, Bioinformatics **23** (2007), no. 23, 3170.

[TPWH10] A. Tibau Puig, A. Wiesel, and A.O. Hero, *Order-preserving factor analysis*, Tech. report, University of Michigan, June 2010.

[TPWOH09] A. Tibau Puig, A. Wiesel, and A. O. Hero, *A multidimensional shrinkage-thresholding operator*, Statistical Signal Processing, 2009. SSP '09. IEEE/SP 15th Workshop on, 2009, pp. 113 – 116.

[TPWZ+11] A. Tibau-Puig, A. Wiesel, A.K. Zaas, C.W. Woods, G.S. Ginsburg, G. Fleury, and A.O. Hero, *Order-preserving factor analysis: Application to longitudinal gene expression*, Signal Processing, IEEE Transactions on **59** (2011), no. 9, 4447 –4458.

[Tro06] J.A. Tropp, *Just relax: Convex programming methods for identifying sparse signals in noise*, Information Theory, IEEE Transactions on **52** (2006), no. 3, 1030–1051.

[Tse01] P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, Journal of Optimization Theory and Applications **109** (2001), no. 3, 475–494.

[VdVVP02] A.J. Van der Veen, M.C. Vanderveen, and A. Paulraj, *Joint angle and delay estimation using shift-invariance techniques*, Signal Processing, IEEE Transactions on **46** (2002), no. 2, 405–418.

[WDS+05] M.B. Wakin, M.F. Duarte, S. Sarvotham, D. Baron, and R.G. Baraniuk, *Recovery of jointly sparse signals from few random projections*, Proc. Neural Inform. Processing Systems–NIPS, 2005.

[Wig55] E.P. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, The Annals of Mathematics **62** (1955), no. 3, 548–564.

[WNF09] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, *Sparse reconstruction by separable approximation*, Signal Processing, IEEE Transactions on **57** (2009), no. 7, 2479–2493.

[Wol78] S. Wold, *Cross-validatory estimation of the number of components in factor and principal components models*, Technometrics **20** (1978), no. 4, 397–405.

[Wri] M.H. Wright, *The interior-point revolution in optimization: History, recent developments, and lasting consequences*, American Mathematical Society **42**, no. 1, 39–56.

[WTH09] D.M. Witten, R. Tibshirani, and T. Hastie, *A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis*, Biostatistics **10** (2009), no. 3, 515.

[YL06a] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society, Series B **68** (2006), 49–67.

[YL06b] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society Series B Statistical Methodology **68** (2006), no. 1, 49.

[ZCV+09] A.K. Zaas, M. Chen, J. Varkey, T. Veldman, A.O. Hero, J. Lucas, Y. Huang, R. Turner, A. Gilbert, R. Lambkin-Williams, et al., *Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans*, Cell Host & Microbe **6** (2009), no. 3, 207–217.

[ZE10] M. Zibulevsky and M. Elad, *L1-l2 optimization in signal and image processing*, Signal Processing Magazine, IEEE **27** (2010), no. 3, 76 –88.

[ZRY09] P. Zhao, G. Rocha, and B. Yu, *The composite absolute penalties family for grouped and hierarchical variable selection*, The Annals of Statistics **37** (2009), no. 6A, 3468–3497.

Learning from high-dimensional multivariate signals

by

Doctor of Philosophy

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Arnau Tibau-Puig
(Electrical Engineering: Systems)
in The University of Michigan
2012

Doctoral Committee:

Professor Alfred O. Hero III, Chair
Professor Anna Catherine Gilbert
Assistant Professor Rajesh Rao Nadakuditi
Assistant Professor Clayton D. Scott

ABSTRACT

Learning from high-dimensional multivariate signals
by
Arnau Tibau-Puig

Chair: Alfred O. Hero III

Modern measurement systems monitor a growing number of variables at low cost. In the problem of statistically characterizing the observed measurements, budget limitations usually constrain the number n of samples that one can acquire, leading to situations where the number p of variables is much larger than n. In this situation, classical statistical methods, founded on the assumption that n is large and p is fixed, fail both in theory and in practice. A successful approach to overcome this problem is to assume a parsimonious generative model characterized by a number k of free parameters, where k is much smaller than p.

In this dissertation we develop algorithms to fit low-dimensional generative models and extract relevant information from high-dimensional, multivariate time series. First, we define extensions of the well-known Scalar Shrinkage-Thresholding Operator, that we name Multidimensional and Generalized Shrinkage-Thresholding Operators, and show that these extensions arise in numerous algorithms for structured-sparse linear and non-linear regression. Using convex optimization techniques, we show that these operators, defined as the solutions to a class of convex, non-differentiable, optimization problems have an equivalent convex, low-dimensional reformulation. Our equivalence results shed light on the behavior of a general class of penalties that includes classical sparsity-inducing penalties such as the LASSO and the Group LASSO. In addition, our reformulation leads in some cases to new efficient algorithms for a variety of high-dimensional penalized estimation problems.

Second, we introduce two new classes of low-dimensional factor models that account for temporal shifts commonly occurring in multivariate time series. Our first contribution, called Order Preserving Factor Analysis, can be seen as an extension of the non-negative, sparse matrix factorization model to allow for order-preserving

temporal translations in the data. We develop an efficient descent algorithm to fit this model using techniques from convex and non-convex optimization. Our second contribution extends Principal Component Analysis to the analysis of observations suffering from arbitrary circular shifts, and we call it Misaligned Principal Component Analysis. We quantify the effect of the misalignments in the spectrum of the sample covariance matrix in the high-dimensional regime and develop simple algorithms to jointly estimate the principal components and the misalignment parameters.

All our algorithms are validated with both synthetic and real data. The real data is a high-dimensional longitudinal gene expression dataset obtained from blood samples of individuals inoculated by different types of viruses. Our results demonstrate the benefit of applying tailored, low-dimensional models to learn from high-dimensional multivariate time series.

ABSTRACT

Learning from high-dimensional multivariate signals
by
Arnau Tibau-Puig

Chair: Alfred O. Hero III

Modern measurement systems monitor a growing number of variables at low cost. In the problem of statistically characterizing the observed measurements, budget limitations usually constrain the number n of samples that one can acquire, leading to situations where the number p of variables is much larger than n. In this situation, classical statistical methods, founded on the assumption that n is large and p is fixed, fail both in theory and in practice. A successful approach to overcome this problem is to assume a parsimonious generative model characterized by a number k of free parameters, where k is much smaller than p.

In this dissertation we develop algorithms to fit low-dimensional generative models and extract relevant information from high-dimensional, multivariate time series. First, we define extensions of the well-known Scalar Shrinkage-Thresholding Operator, that we name Multidimensional and Generalized Shrinkage-Thresholding Operators, and show that these extensions arise in numerous algorithms for structured-sparse linear and non-linear regression. Using convex optimization techniques, we show that these operators, defined as the solutions to a class of convex, non-differentiable, optimization problems have an equivalent convex, low-dimensional reformulation. Our equivalence results shed light on the behavior of a general class of penalties that includes classical sparsity-inducing penalties such as the LASSO and the Group LASSO. In addition, our reformulation leads in some cases to new efficient algorithms for a variety of high-dimensional penalized estimation problems.

Second, we introduce two new classes of low-dimensional factor models that account for temporal shifts commonly occurring in multivariate time series. Our first contribution, called Order Preserving Factor Analysis, can be seen as an extension of the non-negative, sparse matrix factorization model to allow for order-preserving

temporal translations in the data. We develop an efficient descent algorithm to fit this model using techniques from convex and non-convex optimization. Our second contribution extends Principal Component Analysis to the analysis of observations suffering from arbitrary circular shifts, and we call it Misaligned Principal Component Analysis. We quantify the effect of the misalignments in the spectrum of the sample covariance matrix in the high-dimensional regime and develop simple algorithms to jointly estimate the principal components and the misalignment parameters.

All our algorithms are validated with both synthetic and real data. The real data is a high-dimensional longitudinal gene expression dataset obtained from blood samples of individuals inoculated by different types of viruses. Our results demonstrate the benefit of applying tailored, low-dimensional models to learn from high-dimensional multivariate time series.