Convex matrix sparsity for demixing with an application to graphical model structure estimation

Marina Vinyes

École Doctorale Paris-Est
Mathématiques & Sciences et Technologies
de l'Information et de la Communication

Thèse de doctorat
de l'Université Paris-Est

Domaine : Traitement du Signal et des Images

Présentée par

Marina Vinyes

pour obtenir le grade de

Docteur de l'Université Paris-Est

Convex matrix sparsity for demixing with an application to graphical model structure estimation

---

Soutenue publiquement le 27 novembre 2018 devant le jury composé de :

Guillaume Obozinski	École des Ponts ParisTech	Co-directeur de thèse
Nikos Komodakis	École des Ponts ParisTech	Directeur de thèse
Alexandre d'Aspremont	École Normale Supérieure	Rapporteur
Rene Vidal	Johns Hopkins University	Rapporteur
Claire Boyer	Sorbonne Université	Examinateur
Jean-Philippe Chancelier	École des Ponts ParisTech	Examinateur

Dedicated to my family, Silvia, Lola and Jordi, for their continuing support.

Abstract

The goal of machine learning is to learn a model from some data that will make accurate predictions on data that it has not seen before. In order to obtain a model that will generalize on new data, and avoid overfitting, we need to constrain the model, e.g., by using some a priori knowledge of the structure of the model. Classical approaches to constraining the model include regularization methods such as ridge regression or Lasso regularization. The latter induces sparsity in the solution. Parsimony, which is also called sparsity, has emerged as a fundamental concept in machine learning. Parsimonious models are appealing since they provide more interpretability and better generalization (avoid overfitting) through the reduced number of parameters.

Beyond general sparsity and in many cases, models are constrained structurally so they have a simple representation in terms of some fundamental elements, consisting for example of a collection of specific vectors, matrices or tensors. These fundamental elements are called atoms. In this context, atomic norms provide a general framework for estimating these sorts of models. The goal of this thesis is to use the framework of convex sparsity provided by atomic norms to study a form of matrix sparsity.

First, we develop an efficient algorithm based on Frank-Wolfe methods that is particularly adapted to problems with an atomic norm regularization. Then, we focus on the structure estimation of Gaussian graphical models, where the structure of the graph is encoded in the precision matrix and study the case with unobserved variables. We propose a convex formulation with an algorithmic approach and provide a theoretical result that states necessary conditions for recovering the desired structure.

Finally, we consider the problem of signal demixing into two or more components via the minimization of a sum of norms or gauges, encoding each a structural prior on the corresponding components to recover. In particular, we provide general exact recovery guarantees in the noiseless setting based on incoherence measures.

Résumé

En apprentissage automatique on a pour but d'apprendre un modèle, à partir de données, qui soit capable de faire des prédictions sur des nouvelles données (pas explorées auparavant). Pour obtenir un modèle qui puisse se généraliser sur les nouvelles données, et éviter le sur-apprentissage, nous devons restreindre le modèle. Ces restrictions sont généralement une connaissance a priori de la structure du modèle. Les premières approches considérées dans la littérature sont la régularisation de Tikhonov et plus tard le Lasso pour induire de la parcimonie dans la solution. La parcimonie fait partie d'un concept fondamental en apprentissage automatique. Les modèles parcimonieux sont attrayants car ils offrent plus d'interprétabilité et une meilleure généralisation (en évitant le sur-apprentissage) en induisant un nombre réduit de paramètres dans le modèle.

Au-delà de la parcimonie générale et dans de nombreux cas, les modèles sont structurellement contraints et ont une représentation simple de certains éléments fondamentaux, comme par exemple une collection de vecteurs, matrices ou tenseurs spécifiques. Ces éléments fondamentaux sont appelés atomes. Dans ce contexte, les normes atomiques fournissent un cadre général pour estimer ce type de modèles. périodes de modèles. Le but de cette thèse est d'utiliser le cadre de parcimonie convexe fourni par les normes atomiques pour étudier une forme de parcimonie matricielle.

Tout d'abord, nous développons un algorithme efficace basé sur les méthodes de Frank-Wolfe et qui est particulièrement adapté pour résoudre des problèmes convexes régularisés par une norme atomique. Nous nous concentrons ensuite sur l'estimation de la structure des modèles graphiques gaussiens, où la structure du modèle est encodée dans la matrice de précision et nous étudions le cas avec des variables manquantes. Nous proposons une formulation convexe avec une approche algorithmique et fournissons un résultat théorique qui énonce les conditions nécessaires pour récupérer la structure souhaitée.

Enfin, nous considérons le problème de démixage d'un signal en deux composantes ou plus via la minimisation d'une somme de normes ou de jauges, encodant chacune la structure a priori des composants à récupérer. En particulier, nous fournissons une garantie de récupération exacte dans le cadre sans bruit, basée sur des mesures d'incohérence.

Acknowledgements

I would like to thank first my advisor, Guillaume Obozinski for having me as a PhD student and who has been a great teacher to me.
I feel very lucky to have spent these years in the IMAGINE lab with all the amazing PhD students that contributed to a great ambience and that have now become friends. I am very thankfull to all of you: Laura, Maria, Raghudeep, Francisco, Spyros, Praveer, Shell, Mateusz, Sergey, Martin, Thibault, Alexandre, Loïc, Benjamin.
I would like to give a special thanks to my family and friends for their continuing support during those years and Yann without whom this adventure would not have been the same.

Contents

Chapter 1

Introduction

In machine learning we want to learn a model, that can be a set of parameters, from some data that will make accurate predictions on data that it has not seen before. The ability of model to be effective on data that it has not seen before is called generalization. In order to obtain a model that will generalize on new data, and avoid overfitting, we need to make assumptions on the model. This idea is called *inductive bias* and states that without constraining the class/structure of the models considered, the learned model will completely overfit the data and will perform poorly on new data. Inductive bias can be introduced in several ways. First, by restricting the complexity of the model directly by making hypothesis on the class of models we consider. Second, by encouraging simple solutions through regularization for example. A classical example of *inductive bias* that illustrates both forms is ridge regression as first proposed Tikhonov (1963) where we assume the model to be linear in the features and also add an ℓ_2 regularization to avoid too large parameters. Later in the literature, Tibshirani (1996) considered Lasso regularization (ℓ_1) to induce that only a few parameters are non zero.

Sparsity, also known as parsimony, has emerged as a fundamental concept in machine learning. It derives from Occam's razor principle that assumes that the simplest explanation tends to be the right one. Parsimonious models are appealing since they provide more interpretability and better generalization through the reduced number of parameters. Parsimony is particularly suited for problems in high dimension setting, when the number of parameters to estimate is of the order or larger than the number of samples. Beyond plain sparsity, some a priori knowledge on the structure of the model can be incorporated. It is known as structured sparsity. A first form of structured sparsity, usually employed for vectors, consists in constraints on the support (Yuan and Lin, 2006a; Jenatton et al., 2011a; Obozinski et al., 2011).

In applications, data comes in various forms such as images, videos, genetic microarrays and in some of these cases, data is not naturally represented by vectors but is inherently represented by a matrix or a tensor. When working with matrices, another form of sparsity arises: low rank. A matrix can also be considered from a linear operator (or linear transformation) point of view. The structure of the operator is linked to its singular value decomposition and a

natural new definition of sparsity in this context is to add a low rank a priori on the matrix. Since a low rank constraint is a non convex constraint, a convex surrogate that can be viewed as a counterpart of the ℓ_1 norm, namely the trace norm, has been introduced in the literature as a regularizer to induce low-rank matrix. Low rank models are widely used in machine learning going from classical techniques such as Principal Component Analysis (PCA) to other approaches such as multi-task learning (Obozinski et al., 2010; Argyriou et al., 2008) and matrix completion (Candès and Recht, 2009). PCA can suffer from noisy observations as well as from low interpretability since each principal component is a linear combination of all the original variables. Combining plain sparsity and low-rank can lead to more robust and interpretable solutions and appeared relevant in a number of models and formulations. There are different ways to combine these two types of sparsity, by either requiring that a matrix decomposes as the sum of a low rank plus a sparse matrix, by requiring that the matrix should be simultaneously low rank and sparse, or by constraining or inducing that the matrix is low-rank with structured factors. A sparse + low-rank decomposition (Chandrasekaran et al., 2011; Candès et al., 2011) has proven to lead to more robustness in PCA, known as *robust PCA*. Richard et al. (2013) introduced a convex nonsmooth regularizer encouraging multiple structural effects simultaneously like simultaneously sparse and low-rank matrices. A number of structured low rank models appear in the literature. In these models, a low-rank matrix has factors with additional structure, such as sparsity. Among instances of such models we can cite sparse PCA (Zou et al., 2006; d'Aspremont et al., 2008a) and subspace clustering (Vidal, 2011). Note that dictionary learning(Elad and Aharon, 2006; Mairal et al., 2014) and non-negative matrix factorization(Lee and Seung, 1999) correspond also to structure factorization models in which the factors have some specific structure, but the number of factors can be higher and the obtained matrix is therefore not necessarily low rank.

In many cases, models are constrained structurally so they have a simple representation in terms of some elementary pieces, consisting for example of a collection of specific vectors, matrices or tensors. These elementary pieces are called *atoms* and the mathematical object that induces sparse representations in terms of these atoms is called an *atomic norm.*. Atomic norms provide a general framework for convex sparsity. In a number of formulations, sparsity in enforced through regularization. In this thesis we contribute to this general approach and methodology by proposing new atomic norms, new algorithms, new uses of these norms in models and new theoretical results.

From a computational perspective, optimizing problems with sparse regularizations can be challenging. There has been a significant amount of research on structured convex regularizations and a number of algorithms have been proposed, in particular based on proximal methods. However proximal operators can be difficult to compute and may require to solve a complex optimization problem in itself. In the last five years there has been a regain of interest for a family of algorithm known as the Frank-Wolfe algorithm and variants which do not require to compute the proximal operator and can reveal to be powerful. In this thesis we develop efficient algorithms based on active-set strategies to optimize quadratic problems

regularized by atomic norms.

Probabilistic graphical models provide a language to construct structured models. An application in biology is to help infer the network of regulatory relationships among genes from data on their expression levels (Friedman, 2004). The graphical model encodes the structure of the problem considered but unfortunately very often the structure is not known or only partially known so that the structure of the model has to be learned too. This is called *structure learning*. For an undirected Gaussian graphical model, the graph structure and the parameters of the model are simultaneously encoded in the inverse covariance matrix. Two parameters of the model are conditionally independent given the others if and only if the corresponding entry on the inverse covariance matrix is zero. Hence, learning the structure of the graph boils down to learning a matrix (the inverse covariance matrix). The choice of structure of the underlying graphical model is often combinatorial. Two general classes of methods have been considered in the literature: greedy methods such as neighborhood selection (Meinshausen and Bühlmann, 2006), and convex relaxation with sparse regularization such as *graphical lasso* proposed by Banerjee et al. (2008). A major difficulty too often ignored in structure learning is the fact that if some variables are not observed, the marginal dependence graph over the observed variables will possibly be significantly more complex and no longer reflect the direct dependences that are potentially associated with causal effects. Unobserved variables are also called confounders and the problem of assesing causal effects of confounding factors is a subject of increasing interest in the past few years. In this thesis we focus on the problem of structure learning for Gaussian graphical models with unobserved variables. We use a convex sparse formulation based on an atomic norm that allows at the same time to regularize and leads to a convex relaxation of the problem. The approach boils down to approximate the empirical precision matrix by a superposition of components: a sparse matrix and low-rank sparse factors. We provide theoretical conditions for identifiability of the different components of the decomposition.

From a theoretical aspect, a more general question that arises is the problem of identifying a decomposition of multiple structured signals from an observation. This problem is known as the demixing problem. *Given a signal y that is a linear combination of signals $x_i^\star$ and some prior information on the characteristics or structure of the $x_i^\star$, can we identify the components $x_i^\star$ unambiguously?* We consider the problem of signal demixing into two or more components via the minimization of a sum of norms or gauges, encoding each a structural prior on the corresponding components to recover. The analysis is done in the context of atomic gauges. In particular, we provide general exact recovery guarantees in the noiseless setting based on incoherence measures.

1.1 Outline and contributions

The plan of this thesis is presented below. Chapter 2, Chapter 3 and Chapter 5 are introductory chapters while Chapter 4, Chapter 6 and Chapter 7 present our contributions.

Chapter 2 This chapter reviews useful concepts in optimization, and defines the concept of *gauge* also known as *Minkowski functionnal* that is a generalization of a norm. Next, the more recent concept of *atomic gauges* first introduced by Chandrasekaran et al. (2012) is explained and the properties of gauges and atomic gauges are presented. This will be essential throughout the thesis to understand how algorithms and optimality proofs of Chapter 6 and 7 are derived.

Chapter 3 This chapter describes the general framework of convex sparsity with atomic gauges. In particular we describe convex formulation for problems where the a priori structure is encoded as an atomic norm and review optimization algorithms in the literature that address similar formulations with a focus on the connections between different methods and their application to atomic norm regularization. Different methods are presented: proximal splitting methods, coordinate descend algorithms, Frank Wolfe and its variants and column generation algorithms, that sometimes appear in machine learning literature under the name of working set. A particular focus is done in describing active-set algorithm for quadratic programming which, in the case of simple constraints, turns out to be very efficient.

Chapter 4 This chapter is our first contribution and is based on our paper Vinyes and Obozinski (2017). We consider optimization problems that consist in minimizing a quadratic function under an atomic norm regularization or constraint. In the line of work on conditional gradient algorithms, we show that the fully corrective Frank-Wolfe (FCFW) algorithm - which is most naturally reformulated as a column generation algorithm in the regularized case - can be made particularly efficient for difficult problems in this family by solving the simplicial or conical subproblems produced by FCFW using a special instance of a classical *active set algorithm for quadratic programming* (Nocedal and Wright, 2006) that generalizes the min-norm point algorithm (Wolfe, 1976a).

Chapter 5 This chapter presents an overview of probabilistic graphical models, focusing in particular in undirected Gaussian graphical models. The edge structure of the graph defining an undirected graphical model describes precisely the structure of dependence between the variables in the graph. In many applications, the dependence structure is unknown and it is desirable to learn it from data, often because it is a preliminary step to be able to ascertain causal effects. This problem, known as structure learning, is a hard problem in general, but for Gaussian graphical models it is slightly easier because the structure of the graph is given by the sparsity pattern of the precision matrix. This Chapter reviews different methods of structure learning that exist in the literature for undirected and directed graphs. As it will be the object of 6, methods for learning the structure of Gaussian graphical models are reviewed in more detail.

Chapter 6 This chapter is also one of our contributions and is based on our paper Vinyes and Obozinski (2018). We focus on the case of gaussian graphical models with unobserved variables and propose a convex optimization formulation based on structured matrix sparsity to estimate the complete connectivity of the original complete graph including unobserved variables, given the knowledge of the number of missing variables, and a priori knowledge of their level of connectivity. Our formulation is supported by a theoretical result of identifiability of the latent dependence structure for sparse graphs in the infinite data limit.

Chapter 7 This chapter is another contribution of this thesis and aims to generalize the theoretical results of previous chapter. We consider the problem of signal demixing into two or more components via the minimization of a sum of norms or gauges, encoding each a structural prior on the corresponding components to recover. In particular, we provide general exact recovery guarantees in the noiseless setting based on *local cumulative coherence* measures that are related to the *cumulative coherence* measures introduced in Tropp (2004), for combinations of norms, that satisfy a decomposition property of the subgradient. In the case of demixing of two components, we provide finer recovery result applicable to general coercive gauges. Our general results subsume specific results from the literature for Basis Pursuit, Morphological Component Analysis, sparse+ low rank matrix decomposition and others.

Chapter 2

Atomic norms to induce structure

In many machine learning applications, and particularly for ill-posed problems, models are constrained structurally so they have a simple representation in terms of some fundamental elements, consisting for example of a collection of specific vectors, matrices or tensors. In this chapter we present the concept of gauge (extension of a norm) that can be used to impose structure on a problem. First, we review useful concepts in optimization and properties of gauges. This will be essential throughout the thesis to understand how we derive algorithms and optimality proofs of Chapter 6. Second, we introduce the concept of atomic gauges and explain how they leverage structure. Finally we present some examples of such gauges.

2.1 Concepts in convex optimization

In a purpose of of self-containedness, we present in this section important tools to study non-smooth convex optimization problems related to structured sparse methods. Most of them can be found in classical convex optimization books (Boyd and Vandenberghe, 2004; Bertsekas, 1999; Nocedal and Wright, 2006; Borwein and Lewis, 2006). Let us briefly remind the concept of smooth function. f is called smooth if it is differentiable and its gradient is L-Lipschitz

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall (x, y) \in \mathbb{R}^p \times \mathbb{R}^p. \tag{2.1}$$

Throughout this thesis, whenever not specified, x belongs to the ambient vector space $\mathbb{R}^p$.

2.1.1 Subgradients and polar gauges

The subgradient is an extension of the concept of gradient for non-differentiable convex functions. This concept will prove useful in our analysis since usually sparsity-inducing regularizers are not differentiable.

Definition 1. (subgradient) *Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be a convex function. A vector $z \in \mathbb{R}^p$ is called a* subgradient *of f at point $x_0 \in \mathbb{R}^p$ if for any $x \in \mathbb{R}^p$ we have*

$$f(x) \geq f(x_0) + \langle z, x - x_0 \rangle$$

The set of all subgradients of f at x_0 is called subdifferential and is written $\partial f(x_0)$.

Any subgradient z in $\partial f(x_0)$ defines a linear function $x \mapsto f(x_0) + \langle z, x - x_0 \rangle$ that is tangent to the graph of function f. Subgradients are a key object in nonsmooth optimization problems since they characterize optimality conditions as illustrated in the following proposition. The proof can be found on Chapter 3 of Bertsekas (2015).

Proposition 1. (subgradients at optimality) *For any convex function $f : \mathbb{R}^p \mapsto \mathbb{R}$, a point x^* in $\mathbb{R}^p$ is a global minimizer of f if and only if $0 \in \partial f(x^*)$.*

In particular, for a convex and differentiable function f, the subdifferential at point x_0 is a singleton that reduces to the gradient $\nabla f(x_0)$. The subgradient of a *gauge* has a nice characterization that will be extensively used in our analysis. Before introducing it, we need to introduce the concepts of *gauge* and *polar gauge* which are central objects throughout this thesis.

Gauges are interesting to consider since, unlike norms, they allow to introduce non-symmetric regularizers. Two enlightening cases are: the fact that a subspace constraint combined with a norm regularization can be formulated concisely as a gauge whose domain is the corresponding subspace; the fact that when working on cones (like cones of matrices), it is natural to define a regularizer only on the cone and which is therefore not necessarily symmetric.

We first introduce the concept of gauge on a convex set.

Definition 2. (gauge) *Let C be a bounded convex set containing the origin. The gauge of C is*

$$\gamma_C(x) := \inf\{t \mid x \in tC\}$$

In the next definition we introduce the concepts of *closed*, *finite* and *coercive* gauge.

Definition 3. (closed, finite, coercive gauge) *Let γ_C be a gauge.*

- *γ_C is a* closed gauge *if and only if the convex set C is closed.*
- *γ_C is a* finite gauge *if and only if $\forall x \in \mathbb{R}^p$, $\gamma_C(x)$ takes a finite value.*
- *γ_C is a* coercive gauge *if and only if $(\gamma_C(x) = 0) \Rightarrow (x = 0)$.*

All gauges considered in this thesis are closed gauges unless specified. A norm is a particular instance of a *gauge* when C is a bounded closed centrally symmetric set with non empty interior. In that case C is simply the unit ball of the norm. Figure 2.1 illustrates a norm (a gauge on a closed bounded symmetric set C) and a gauge on a closed bounded non-symmetric set C.

In the following proposition we present some elementary properties of gauges: the fact that the sum of two gauges is a gauge and the infimal convolution[1] of two gauges is also a gauge.

[1] The infimal convolution of two functions f and g is denoted by $f \square g$ and writes $f \square g(x) = \inf_y f(x-y) + g(y)$

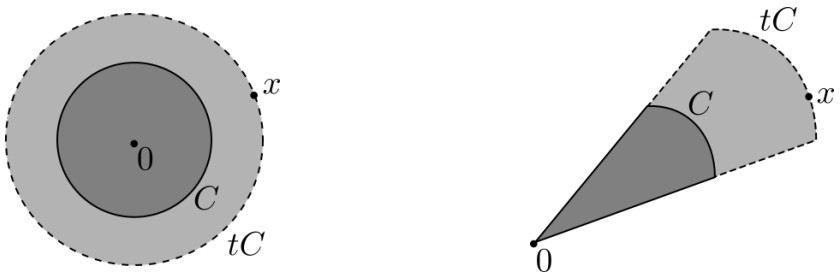


Figure 2.1: Illustration of a norm(left) and a gauge(right) defined form a convex set C.

Proposition 2. (some properties of gauges) *Let γ_{C_1} and γ_{C_2} be two gauges from convex sets C_1 and C_2.*

(i) the sum of two gauges is a gauge and $\gamma_{C_1} + \gamma_{C_2} = \gamma_{C \sharp D}$, where $C_1 \sharp C_2 := \bigcup_{\lambda \in [0,1]} \lambda C_1 \cap (1-\lambda) C_2$ is the inverse sum defined in Section 3, Part 1 of Rockafellar (1970)[2].

(ii) the infimal convolution of two gauges is a gauge and $\gamma_{C_1} \square \gamma_{C_2} = \gamma_{Conv(C_1 \cup C_2)}$

Proof. Let $x \in \mathbb{R}^p$, we have

$$\begin{aligned}(\gamma_{C_1} + \gamma_{C_2})(x) &= \inf_{s,t \in \mathbb{R}^+} \{s + t \mid x \in tC_1 \cap sC_2\} \\ &= \inf_{s,t \in \mathbb{R}^+} \{s + t \mid x \in (t+s) C_1 \sharp C_2\}\end{aligned}$$

For the infimal convolution we have

$$\begin{aligned}(\gamma_{C_1} \square \gamma_{C_2})(x) &= \inf_{z_1, z_2 \in \mathbb{R}^p} \{\gamma_{C_1}(x_1) + \gamma_{C_2}(x_2) \mid x = z_1 + z_2\} \\ &= \inf_{s,t \in \mathbb{R}^+} \{s + t \mid x = sx_1 + tx_2,\ \ x_1 \in C_1,\ \ x_2 \in C_2\} \\ &= \inf_{s+t \in \mathbb{R}^+} \{s + t \mid x \in (s+t)\mathrm{Conv}(C_1 \cup C_2)\} \\ &= \gamma_{\mathrm{Conv}(C_1 \cup C_2)}(x)\end{aligned}$$

where Conv designates the convex hull of a set. □

In a number of cases (cf section 2.2.3 about atomic norm examples) it is much easier to handle the polar gauge. The concept of polar gauge is presented in the following definition. The same concept applied to norms is called dual norm.

Definition 4. (polar gauge) *Let $\gamma : \mathbb{R}^p \mapsto \mathbb{R}^+$ be a gauge. The polar gauge is denoted γ° and is defined as*

$$\gamma^\circ(y) := \sup_{\substack{x \in \mathbb{R}^p \\ \gamma(x) \leq 1}} \langle x, y \rangle,$$

for all $y \in \mathbb{R}^p$.

[2]If C_1, C_2 are convex cones containing the origin $C_1 \sharp C_2 = C_1 \cap C_2$ (Section 3, Part 1 of Rockafellar (1970))

The subdifferential of a gauge has a handy characterization involving the polar gauge stated in the following lemma.

Lemma 1. (subdifferential of a gauge) *Let $\gamma : \mathbb{R}^p \mapsto \mathbb{R}^+$ be a gauge. A characterization of the subdifferential of γ at the point x_0 in $\mathbb{R}^p$ is*

$$\partial\gamma(x_0) = \{z \in \mathbb{R}^p; \quad \langle z, x_0\rangle = \gamma(x_0) \quad \textit{and} \quad \gamma^\circ(z) \leq 1\}. \tag{2.2}$$

Proof. First we define the set $\mathcal{G}(x_0) \coloneqq \{z \in \mathbb{R}^p; \quad \langle z, x_0\rangle = \gamma(x_0) \quad \text{and} \quad \gamma^\circ(z) \leq 1\}$. We will prove that $\mathcal{G}(x_0)$ equals $\partial\gamma(x_0)$. If $z \in \mathcal{G}(x_0)$ then

$$\gamma(x_0) + \langle z, x - x_0\rangle = \langle z, x\rangle \leq \gamma(x)\gamma^\circ(z) \leq \gamma(x)$$

where the first inequality comes from the definition of polar gauge.
To prove the other direction let us take $z \in \partial\gamma(x_0)$, then for any x,

$$\langle z, x\rangle - \gamma(x) \leq \langle z, x_0\rangle - \gamma(x_0).$$

By taking the supremum over all x we obtain

$$\gamma^*(z) \leq \langle z, x_0\rangle - \gamma(x_0), \tag{2.3}$$

where $\gamma^*(z)$ is the fenchel conjugate (definition 5) of γ and equals 0 if $\gamma^\circ(z) \leq 1$ and $+\infty$ otherwise, accordind to lemma 4. Since right hand side of 2.3 cannot be $+\infty$ for any z, we must have $\gamma^\circ(z) \leq 1$. As a consequence, equation 2.3 gives us $\gamma(x_0) \leq \langle z, x_0\rangle$. On the other hand

$$\langle z, x_0\rangle \leq \gamma(x_0)\gamma^\circ(z) \leq \gamma(x_0)$$

where first inequality comes from the definition of polar gauge. □

2.1.2 Fenchel conjugate and duality gap

Duality is a central concept in optimization. A convex optimization problem may be viewed from either of two perspectives, a primal problem or a dual problem. Fenchel conjugation is a transform that applies to any function and can be used to transform some optimization problems, mainly when the problem is composed by a sum of two convex functions, into a corresponding dual problem , which can sometimes be simpler to solve.

Definition 5. (Fenchel conjugate) *Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be a function, the* Fenchel conjugate *f^* of f is*

$$f^*(y) = \sup_{x \in \mathbb{R}^p} \langle x, y\rangle - f(x)$$

for all $y \in \mathbb{R}^p$.

The Fenchel conjugate of a function f is always convex, regardless of the nature of function f, as a pointwise supremum of linear functions. Another useful property that comes directly from Definition 5 is the Fenchel-Young inequality:

Lemma 2. (Fenchel-Young inequality) *For any function $f : \mathbb{R}^p \mapsto \mathbb{R}$ and its Fenchel conjugate $f^* : \mathbb{R}^p \mapsto \mathbb{R}$,* Fenchel–Young *inequality holds for every x, α in $\mathbb{R}^p$*

$$\langle x, \alpha \rangle \leq f(x) + f^*(\alpha).$$

When f is convex, equality holds if and only if $\alpha \in \partial f(x)$ [3]*.*

Proof. If equality holds, by applying the definition 5 on the equality we get that for all $y \in \mathbb{R}^p$, $\langle x, \alpha \rangle \leq f(x) + \langle y, \alpha \rangle - f(y)$, which is a characterization of the subgradient of a convex function and $\alpha \in \partial f(x)$.

Now, let $\alpha \in \partial f(x)$. Fenchel-Young inequality holds, we only need to proove $\langle x, \alpha \rangle \geq f(x) + f^*(\alpha)$. By applying the characterization of the subgradient of a convex function, we have for all $y \in \mathbb{R}^p$, $f(y) \geq f(x) + \langle \alpha, y - x \rangle$. Taking the suppremum on all y gives us the desired inequality $\langle \alpha, x \rangle \geq f(x) + f^*(x)$. □

Another important property, stated in the next lemma, is that a closed convex function equals its biconjugate. The proof can be found in Theorem 12.2, page 104 of (Rockafellar, 1970).

Lemma 3. (biconjugate) *A function f equals its biconjugate, i.e. $f^{**} = f$ if and only if f is a closed convex function.*

The Fenchel conjugate of a gauge γ is directly related to the polar gauge. It is in fact the indicator function of the convex set $\{y \in \mathbb{R}^p \ : \ \gamma^\circ(y) \leq 1\}$. The indicator function of a set C is denoted as ι_C and equals 0 on C and $+\infty$ otherwise.

Lemma 4. (Fenchel conjugate of a gauge) *The* Fenchel conjugate *of a gauge $\gamma : \mathbb{R}^p \mapsto \mathbb{R}$ is*

$$\gamma^*(y) = \begin{cases} 0 & \text{if} \quad \gamma^\circ(y) \leq 1 \\ +\infty & \text{otherwise} \end{cases} = \iota_{\gamma^\circ}(y)$$

for all $y \in \mathbb{R}^p$.

Proof. By definition 5, $\gamma^*(y) = \sup_{x \in \mathbb{R}^p} \langle x, y \rangle - \gamma(x)$. We distinguish two cases: either $\gamma^\circ(y) \leq 1$, then $\langle x, y \rangle \leq \gamma(x)$ for all $x \in \mathbb{R}^p$ and equality holds if $x = 0$; or $\gamma^\circ(y) > 1$ and there exists an x with $\gamma(x) \leq 1$, and $\langle x, y \rangle > 1$. Therefore, for any $t > 0$

$$\gamma^*(y) \geq \langle tx, y \rangle - t\gamma(x) = t\left(\langle x, y \rangle - \gamma(x)\right)$$

and $\gamma^*(y) \to \infty$ when $t \to \infty$ □

In this work we will focus on optimization problems that are a sum of a smooth and a non-smooth convex function. Now that all technical concepts have been introduced, we make explicit a dual problem in the following lemma.

[3] Note that it may happen that the subgradient is empty, in which case equality is never reached.

Lemma 5. (dual problems) *Let $g : \mathbb{R}^p \mapsto \mathbb{R}$ and $h : \mathbb{R}^p \mapsto \mathbb{R}$ be two functions, and $X \in \mathbb{R}^{p\times n}$ a linear operator. The following two problems*

$$P(x) := \arg\min_{x\in\mathbb{R}^p} g(X^\top x) + h(x),$$

$$D(\alpha) := \arg\max_{\alpha\in\mathbb{R}^p} -g^*(\alpha) - h^*(-X\alpha),$$

are duals of each other, where g^* and h^* are the Fenchel conjugates of f and g respectively.

During the optimization, sequences of primal variables x are available, and duality gaps give an upper bound on the difference between the objective value of the current solution and the optimal value, that is $P(x) - P(x^*)$, where x^* is an optimum of the primal problem. Therefore, computing duality gaps is a meticulous way of checking the accuracy of the current solution. Given any primal variable x and dual variable α in $\mathbb{R}^p$, the Fenchel duality gap is defined as the difference between primal and dual objectives, that is $P(x) - D(\alpha)$ with the notations of lemma 5. The duality gap can be decomposed in two terms,

$$\begin{aligned} P(x) - D(\alpha) &= g(X^\top x) + h(x) + g^*(\alpha) + h^*(-X\alpha) && (2.4)\\ &= \underbrace{g(X^\top x) + g^*(\alpha) - \langle x, \alpha\rangle}_{\geq 0} + \underbrace{h(x) + h^*(-X\alpha) + \langle x, \alpha\rangle}_{\geq 0} && (2.5)\end{aligned}$$

where the positiveness of the two terms is given by Fenchel-Young inequality. A duality gap is always positive and for convex problems, primal and dual objectives are equal at an optimum, hence the duality tends to zero while approaching an optimum. Fenchel duality gap defines an upper bound on $P(x) - P(x^*)$,

$$P(x) - D(\alpha) \geq P(x) - P(x^*) \geq 0.$$

Assuming that Fenchel conjugates are easy to compute, we still need to choose a "good" dual variable α such that $P(x) - D(\alpha)$ is as small as possible. An appropriate choice would be $\alpha := \nabla g(X^\top x)$ as it would set to zero the expression $g(X^\top x) + g^*(\alpha) - \langle x, \alpha\rangle$, and simplify the duality gap expression to

$$P(x) - D(\alpha) = h(x) + h^*(-\nabla g(X^\top x)) - \left\langle x, \nabla g(X^\top x)\right\rangle. \tag{2.6}$$

In this thesis, we are interested in the specific case where $h := \gamma$ is a gauge. Hence, $h^* := \iota_{\gamma^\circ}$ is its fenchel conjugate, and we scale the dual variable to ensure the constraints imposed by the indicator function, that is $\{\alpha \ : \ \gamma^\circ(\alpha) \leq 1\}$, are satisfied. We define a dual variable as

$$\hat{\alpha} := \min\left(1, \frac{1}{\gamma^\circ\left(\nabla g(X^\top x)\right)}\right) \nabla g(X^\top x),$$

and derive the corresponding duality gap

$$P(x) - D(\hat{\alpha}) = g(X^\top x) + g^*(\hat{\alpha}) + h(x). \tag{2.7}$$

2.2 Atomic norms for leveraging structure

In many machine learning applications, and particularly for ill-posed problems, models are constrained structurally so they have a simple representation in terms of some fundamental elements, consisting for example of a collection of specific vectors, matrices or tensors. Examples of such elements include sparse vectors for many sparsity inducing norms, rank-one matrices for the trace norm or low-rank tensors as used in the *nuclear tensor norm* (Liu et al., 2013). We call *atoms* these elements and *atomic set* $\mathcal{A}$ their (possibly uncountable) collection. In this section we remind usefull concepts and results of atomic norms that can be found in classical literature (Rockafellar, 1970; Chandrasekaran et al., 2012). We describe a model as "simple" if it can be written as a nonnegative combination of a "few" elements from an atomic set. In other words, the model is "simple" if it is a sparse combination of atoms. Parsimonious models are useful in machine learning for three main reasons: they lead to a better generalization of the model (avoid overfitting); they give interpretability through atom selection (extension of variable selection) and in some cases they are computationally cheaper both to learn at to perform predictions. The computational advantage depends on how hard is to find the atoms since the problem can be NP-hard.

2.2.1 Definition from a collection of atoms

We define now atomic gauges. Let $\mathcal{A}$ be a subset of $\mathbb{R}^p$, defined as a collection of atoms. These atoms can be sparse vectors, rank-one matrices and multiple other choices. The penalty of an element x in $\mathbb{R}^p$ is the minimum sum of non-negative weights c_i such that x writes as a linear combination of atoms $(a_i)_i \in \mathcal{A}$, $x = \sum_i c_i a_i$. Therefore, an atomic gauge is defined from its set of atoms $\mathcal{A}$ as the norm defined on the convex envelope of $\mathcal{A}$, denoted $C_{\mathcal{A}}$. The formal definition of atomic gauge is stated below.

Definition 6. (atomic gauge) *$\mathcal{A}$ is bounded and closed, and provided its convex hull $C_{\mathcal{A}}$ has non empty interior, we can define an atomic gauge $\gamma_{\mathcal{A}}$ as the gauge of $C_{\mathcal{A}}$.*

$$\gamma_{\mathcal{A}}(x) := \inf\{t \mid x \in t\mathit{Conv}(\mathcal{A})\}$$

When $C_{\mathcal{A}}$ is also centrally symmetric, $\gamma_{\mathcal{A}}$ is in fact an atomic norm and $C_{\mathcal{A}}$ is its unit ball. Chandrasekaran et al. (2012) show that the atomic gauge induced by the atomic set $\mathcal{A}$ is indeed a gauge and can be rewritten in a simple form. This characterization is stated in the next lemma,

Lemma 6. *Let $\gamma_{\mathcal{A}}$ be an atomic gauge induced by atomic set $\mathcal{A}$. In a finite dimensional space,*

$$\gamma_{\mathcal{A}}(x) := \inf\{\sum_{a\in\mathcal{A}} c_a \mid \sum_{a\in\mathcal{A}} c_a a = x,\ c_a \geq 0,\ a \in \mathcal{A}\}.$$

Proof. By applying Carathéodory's theorem, stated just below, we know that any point x in $\mathbb{R}^p$ writes a a convex combination of at most $p+1$ atoms in $\mathcal{A}$. Thus, we can write

$$\gamma_{\mathcal{A}}(x) := \inf\{t \mid x = t\sum_{a\in\mathcal{A}} w_a a \quad \text{s.t.} \quad \sum_a w_a = 1\}$$

where all the sums are in fact finite sums. By making the simple change of variable $c_a := tw_a$ we get the result.

Theorem 1. *(theorem of Carathéodory) If a point x of $\mathbb{R}^d$ lies in the convex hull of a set P, then x can be written as the convex combination of at most $d+1$ points in P.*

□

2.2.2 Polar gauge and subgradient

In some cases polar gauges of atomic gauges have a simple expression and can be much easily computed than the gauge itself. In the following lemma we state the a characterization of the polar of an atomic gauge.

Lemma 7. (polar of an atomic gauge) *The polar gauge of an atomic gauge induced by the atomic set $\mathcal{A}$ is*

$$\gamma_{\mathcal{A}}^{\circ}(x) = \sup_{a \in \mathcal{A}} \langle x, a \rangle .$$

Proof. Following the definition 4 of polar gauge

$$\gamma_{\mathcal{A}}^{\circ}(x) = \sup_{y \in \mathbb{R}^p;\ \gamma_{\mathcal{A}}(y) \le 1} \langle x, y \rangle = \sup_{\substack{c_a \ge 0;\ a \in \mathcal{A} \\ \sum_{a \in \mathcal{A}} c_a \le 1}} \sum_{a \in \mathcal{A}} c_a \langle x, a \rangle = \sup_{a \in \mathcal{A}} \langle x, a \rangle$$

where in the second equality we use the characterization of an atomic gauge of lemma 6 and the last equality uses the fact that maximum value of a linear function over a convex set occurs at an extreme point of the region, i.e. an atom in this case. □

Subsequently, we obtain the following characterization of the subgradient of an atomic gauge,

Lemma 8. (subdifferential of an atomic gauge) *Let $\gamma_{\mathcal{A}}$ be the atomic gauge induced by atomic set $\mathcal{A}$. A characterization of the subdifferential of Ω at the point x_0 in $\mathbb{R}^p$ is*

$$\partial \gamma_{\mathcal{A}}(x_0) = \{ z \in \mathbb{R}^p; \quad \langle z, x_0 \rangle = \gamma_{\mathcal{A}}(x_0) \quad \textit{and} \quad \forall a \in \mathcal{A},\ \langle x, a \rangle \le 1 \} . \tag{2.8}$$

2.2.3 Examples of gauges

In this section we present some examples of gauges.

Indicator of cone

Indicator of a convex cone, like cones[4] of matrices, is a gauge. Let C be a convex cone, $\gamma_C(x) := \inf\{t \mid x \in tC\}$. We distinguish two cases: either $x \in C$, and we have $x \in tC$ for any $t > 0$, taking $t \to 0$ we get $\gamma_C(x) = 0$; or $x \notin C$, and by definition of a cone there is no $t \ge 0$ such that $x \notin tC$, so $\gamma_C(x) = \infty$. For instance, the indicator function of the cone of positive semidefinite matrices is a gauge.

[4] A set C is a convex cone if it is convex and a cone, i.e., $\forall x_1, x_2 \in C$ and $\theta_1, \theta_2 \ge 0, \quad \theta_1 x_1 + \theta_2 x_2 \in C$.

Atomic gauges of union of atomic sets

For a number of atomic gauges we have $\mathcal{A} = \bigcup_{j=1}^{J} C_j$ where C_j are convex sets. As a consequence $\gamma_{\mathcal{A}}^{\circ}(s) = \max_j \gamma_{C_j}^{\circ}$ and $\gamma_{\mathcal{A}} = \gamma_{C_1} \square \ldots \square \gamma_{C_J}$ where $\square$ denotes the infimal convolution[5] with $f \square g(x) = \inf_y f(x-y) + g(y)$. We thus have

$$\gamma_{\mathcal{A}}(x) = \inf\{ \gamma_{C_1}(z_1) + \ldots + \gamma_{C_J}(z_J) \mid z_1 + \ldots + z_J = x\}.$$

.

Lasso and Group Lasso

The Lasso is a natural example of atomic norm, whose atoms are

$$\mathcal{A} := (\pm e_i)_{i \in [\![p]\!]},$$

where the $(e_i)_{i \in [\![p]\!]}$ is the canonical basis of $\mathbb{R}^p$. The Lasso polar norm is defined as $\Omega_{\text{Lasso}}^{\circ}(s) = \max_{i \in [\![p]\!]} |s_i|$. More generally, given $\mathcal{G}$ a partition of $[\![p]\!]$ and fixed positive weights δ_G (usually $\sqrt{|G|}$) for each set $G \in \mathcal{G}$, the Group Lasso norm is the atomic norm whose atoms are the vectors of Euclidean norm δ_G^{-1} and support in $\mathcal{G}$,

$$\mathcal{A} := \bigcup_{G \in \mathcal{G}} \{u \in \mathbb{R}^p \mid supp(u) \subset G,\ \|u\|_2 \leq \delta_G^{-1}\}.$$

The Group Lasso polar norm is defined as $\Omega_{\text{GL}}^{\circ}(s) = \max_{G \in \mathcal{G}} \delta_G^{-1} \|s_G\|_2$.

Trace norm

The trace norm is an atomic norm induced by the set of atoms

$$\mathcal{A}_{\text{tr}} := \left\{uv^{\top},\ \|u\|_2 = \|v\|_2 = 1\right\}$$

consisting of rank-one matrices. The polar norm is the nuclear norm. Note that when the matrix is positive semidefinite then the trace norm is just the trace. Hence, the trace norm plus the indicator cone of the positive semidefinite matrices defines a gauge which is the atomic gauge associated with the rank of a symmetric positive semidefinite matrix.

"Non-atomic" gauges

From a technical point of view all gauges are atomic, with the atoms being all the elements of the generating set of the gauge. What we mean by "non-atomic" gauges are gauges with a non sparse atomic set, and consequently their atomic structure cannot be exploited efficiently. The key property that we exploit in this thesis is related to conditional gradient algorithms: the gauges that we like to view as atomic gauges, are the ones whose polar gauge can be computed

[5]The infimal convolution is clearly commutative and associative.

efficiently, which corresponds to the LMO[6] in conditional gradient algorithm as we will explain in Chapter 3.

Some useful norms in machine learning do not have a sparse atomic set. In particular norms inducing multiple sparsity patterns at the same time like the ones introduced in Richard et al. (2013). Another example of a norm with non sparse atomic set is an extention of the ℓ_1/ℓ_2-norm introduced by (Jenatton et al., 2011b). The proposed norm induces supports that arise as intersections of a sub-collection of groups defining the norm. Given a collection of sets $\mathcal{B}$ covering $[\![p]\!]$ and which can overlap, and fixed positive weights w_B for each set $B \in \mathcal{B}$, the norm is defined as $\Omega(s) := \sum_{B \in \mathcal{B}} w_B \|s_B\|^2$.

In (Obozinski and Bach, 2012) authors show that all the norms that are the best relaxation of combinatorial penalties are naturally defined as "atomic" norms. In particular, authors show that there is implicitly a combinatorial function associated with the norms of (Jenatton et al., 2011b) and that the norm used (Jenatton et al., 2011b) is not the tightest relaxation, and that there is one that is better, and which is defined as an "atomic" norm.

[6] Linear Minimisation Oracle

Chapter 3

Optimization for convex sparsity

In machine learning we want to approach a distribution but we only have a finite set of examples. In particular in high dimension setting when the number of parameters to estimate is of the order or larger than the number of samples there is no unique solution. We call these problems ill-posed problems. These ambiguities in the solution can be reduced by incorporating some a priori knowledge on the structure of the objects to estimate. This a priori information can be applied as a regularization as first proposed Tikhonov (1963) for ridge regression or later by Tibshirani (1996) for inducing sparsity in the solution. A general framework for convex sparsity is provided by atomic norms. We review a convex formulation for problems where the a priori structure is encoded as an atomic norm and review optimization algorithms in the literature that address similar formulations with a focus on the connections between different methods and their application to atomic norm regularization. We briefly present proximal splitting methods and coordinate descend algorithms. We describe Frank Wolfe and its variants for a constrained problem with a note on how to extend these methods to the regularized problem. Finally we describe column generation algorithms, that sometimes appear in machine learning literature under the name of working set. We focus on active-set algorithm for quadratic programming where, in the case of simple constraints turns out to be very efficient.

3.1 Convex formulation

In a number of problems in machine learning, we seek for a "simple explanation" of the data leading to better interpretation and better generalization. This approach is particularly important in high dimensional setting where the number of variables p is larger than the number of samples n, $p \gg n$ leading to an ill-posed problem. This is usually achieved by imposing a priori knowledge on the structure of the objects to estimate (Chandrasekaran et al., 2010; Obozinski et al., 2011; Richard et al., 2014; Argyriou et al., 2012; Obozinski and Bach, 2012). Imposing the desired structure explicitly often leads to an intractable optimization problem, but in many cases we can define a convex formulation that yields a useful solution. The approach corresponds to minimizing some smooth convex function $f : \mathbb{R}^p \mapsto \mathbb{R}$ which is

typically an empirical risk in machine learning or another data fitting term, and a regularization term $\psi : \mathbb{R}^p \mapsto \mathbb{R}^+$, which is usually non-differentiable, that promotes the desired structure. The optimisation problem writes

$$\arg\min_{x \in \mathbb{R}^p} f(x) + \psi(x). \tag{3.1}$$

Throughout this thesis we mainly focus on regularized problem where the desired structure can be encoded in an atomic gauge $\gamma_{\mathcal{A}}$, $\psi := \gamma_{\mathcal{A}}$. A constrained formulation where the sparsity is enforced via a constraint is also possible

$$\arg\min_{x \in \mathbb{R}^p} f(x) \quad \text{s.t} \quad \psi(x) \leq \rho, \tag{3.2}$$

Problems (3.1) and (3.2) are non differentiable optimization problems, and so could be resolved by subgradient descent. However subgradient methods are usually very slow. Moreover in many cases ψ is a simple function in the sense that its structure is well understood and can be exploited to compute its *proximal operator*.

3.2 Proximal splitting methods

Proximal splitting methods arise in the context of a minimization of a sum of two convex functions: a smooth function (with Lipshitz gradient constant L) and a convex non-smooth function with a "simple" structure. What we call "simple" structure is the fact that the *proximal operator* can be easily computed. The *proximal operator* was introduced by Moreau (1962)

Definition 7. (proximal operator) *The proximal operator of a proper closed convex function* $\psi : \mathbb{R}^p \mapsto \mathbb{R}$ *is the function denoted* Prox_ψ *and defined as*

$$\text{Prox}_\psi(x) := \arg\min_{y \in \mathbb{R}^p} \left\{ \frac{1}{2}\|x - y\|^2 + \psi(y) \right\}.$$

Note that $\text{Prox}_\psi(x)$ is well defined since is the solution to a strongly convex probem. If ψ^* is the Fenchel conjugate of ψ, then an important and well known result is Moreau's identity (Moreau, 1965), which says that

$$\forall x,\ \text{Prox}_\psi(x) + \text{Prox}_{\psi^*}(x) = x.$$

This shows that computing the proximal operator of a function ψ or its conjugate is equally hard.

Let us consider problem in (3.1). The main proximal splitting algorithm is the forward-backward algorithm (also known as proximal gradient algorithm), presented in Algorithm 1 and consits of two alternating steps:

- *the forward step*: a gradient step with stepsize η : $\tilde{x}^{t+1} \leftarrow x^t - \eta \nabla f(x^t)$
- *the backward step*: a proximal step: $x^{t+1} \leftarrow \text{Prox}_{\eta\psi}(\tilde{x}^{t+1})$.

Algorithm 1 forward-backward

Initialization: $x^0 = 0$, $t = 0$
repeat
  $x^{t+1} \leftarrow \text{Prox}_{\eta\psi}(x^t - \eta\nabla f(x^t))$
until convergence

For any $0 < \eta < 2/L$ the algorithm converges. See Combettes and Pesquet (2011) for a review on proximal operators and proximal splitting methods.

A well-known example is when ψ is the indicator of a convex set C where the algorithm reduces to projected gradient descent (Bertsekas (1999), Chapter 2). Obozinski and Bach (2012) present methods to compute proximal operators for submodular functions.

For some optimization problems instead of performing an update of the whole vector of parameters we only update a subset of coordinates.

3.3 Coordinate descent

Coordinate descend methods optimize, either exatly or approximately, at each iteration the objective with respect to a single variable at a time while others are kept fixed. They all extend naturally to block coordinate descend where the optimization step is done on a block of variables while keeping the others fixed.

Coordinate descent algorithms have many variants. The update applied to each variable in the optimization step can take different forms of approximate updates such as: one or a few gradient descent steps; one or a few projected gradient descent steps; one or a few proximal gradient steps; acceleration steps etc. We focus on presenting proximal coordinate descent.

Let us consider problem in (3.1). When ψ is separable, meaning that it can be written

$$\psi(x) = \sum_{i=1}^{p} \psi_i(x_i),$$

and if for each ψ_i a proximal operator is easily computable, coordinate descent is particularly adapted in this case. The algorithm is presented in Algorithm 2.

Algorithm 2 Coordinate proximal descend

repeat
  Select a coordinate $i \in [\![p]\!]$
  $x_i^{t+1} \leftarrow \text{Prox}_{\eta^t\psi_i}(x_i^t - \eta^t[\nabla f(x^t)]_i)$
until stopping criterion

In a number of cases, the regularization term is not separable but it writes as an infimal convolution (Jalali et al., 2010; Jacob et al., 2009; Richard et al., 2014) and the optimization problem can be reformulated into a problem where the non-differentiable term is separable.

For instance when the regularization term is an atomic gauge that is an infimal convolution of the form

$$\gamma_{\mathcal{A}}(x) = \inf\{ \gamma_{C_1}(z_1) + \ldots + \gamma_{C_J}(z_J) \mid z_1 + \ldots + z_J = x\}, \tag{3.3}$$

the formulation of the optimization problem is particularly amenable to coordinate descent. Indeed, problem (3.1) can be reformulated as

$$\min_{z_1,\ldots,z_J} f(z_1 + \ldots + z_J) + \gamma_1(z_1) + \ldots + \gamma_J(z_J),$$

where the non-differentiable term is now separable. Note that this formulation is not necessarily strongly convex, even if f is strongly convex, which can result in slow convergence.

Friedman et al. (2010) introduce fast coordinate descent algorithms that apply to a broad set of problems: linear regression, logistic regression, and multinomial regression problems with lasso regularization (ℓ_1), ridge regression (ℓ_2) or mixtures of the two (Zou and Hastie, 2005). They can be applied to large datasets and they appear to be much faster than other methods. Convergence proofs of block coordinate descent in convex problems that are a sum of a smooth function and a separable function are established in Tseng (2001) for the cyclic scheme. Convergence of randomized schemes has been studied in Nesterov (2012) and Karimi et al. (2016). See Wright (2015) for a review on coordinate descend algorithms.

3.4 Frank-Wolfe methods

The Frank-Wolfe algorithm, also known as *conditional gradient*, was initially proposed by Frank and Wolfe (1956) for solving quadratic programming problems with linear constraints. They apply in the context where we can easily solve the Linear Minimization Oracle (LMO), a linear problem on a convex set of constraints defined as

$$\mathrm{LMO}_C(x) := \mathrm{argmin}_{a \in C} \langle a, x \rangle,$$

where C is a convex set of constraints. The principle of the algorithms is to build a sequence of approximations of the solution of the problem as a convex combination of extreme points of the constraint set C. In the past years it has captured the interest of machine learning community to solve the constrained problem. Subsequently, we describe the application of Frank-Wolfe algorithm to the constrained problem (3.2) and its application to the regularized problem (3.1).

3.4.1 Applying Frank-Wolfe to the constrained problem

Note that in the constrained formulation 3.2 the constraint set is simply the convex hull of $\mathcal{A}$. Hence, the extreme points of the constraint set correspond to atoms of $\mathcal{A}$, an approximate solution x take the form of a convex combination of atoms, that is $x = \sum_{i=1}^{t} c_i a_i$ with $\sum_{i=1}^{t} c_i = 1$.

This procedure guarantees a feasible sequence. At each iteration, the algorithm considers the linearization of the objective function around the current point x^t, and moves towards a minimizer a^{t+1} of this linear function. This minimizer is the extreme point of the constraint set obtained by the LMO at $-\nabla f(x^t)$,

$$a^{t+1} := \mathrm{LMO}_{\mathcal{A}}(-\nabla f(x^t)).$$

$\{a^{t+1} - x^t\}$ is called Frank-Wolfe direction or *forward* direction. The Frank-Wolfe algorithm is described in algorithm 3, where $\eta^t \in [0, 1]$ is a scalar stepsize and $x^0 = 0$. It can be set to $\frac{1}{1+t}$ or found by line search.

Algorithm 3 Classical Frank-Wolfe

1: **repeat**
2: Compute $a^{t+1} := \mathrm{argmax}_{a \in \mathcal{A}} \langle a, -\nabla f(x^t) \rangle$
3: $x^{t+1} = (1 - \eta^t)\, x^t + \eta^t a^{t+1}$
4: **until** $\langle -\nabla f(x^t), a^{t+1} - x^t \rangle \leq \epsilon$

Note that we get a bound on the duality gap for free. Convexity of f implies that the linear tangent of f at a given x, $y \mapsto f(x) + \langle -\nabla f(x), y - x \rangle$ lies below $f(y)$. Thus for all $y \in \mathrm{Conv}(\mathcal{A})$,

$$f(y) \geq f(x^t) + \langle -\nabla f(x^t), y - x^t \rangle$$

and by minimizing both sides over $y \in \mathrm{Conv}(\mathcal{A})$ we get an $f(x^*) \geq f(x) + \langle -\nabla f(x^t), a^{t+1} - x^t \rangle$. Rearranging the inequality we get an upper bound on the duality gap $f(x^t) - f(x^*)$. It is in fact a special case of the Fenchel duality gap described in Chapter 2.

Other variants of Franke-Wolfe (FW) algorithms have been proposed, notably, FW with away steps (AFW) introduced in Wolfe (1970), pairwise FW (PWFW) (Lacoste-Julien and Jaggi, 2015) and fully corrective Frank-Wolfe (FCFW). We summarize hereafter the form of the different updates for AFW, PWFW and FCFW. The active set of atoms $\mathcal{A}^t$ at time t is recursively defined by $\mathcal{A}^{t+1} = \widetilde{\mathcal{A}}^t \cup \{a^{t+1}\}$ with $\widetilde{\mathcal{A}}^t$ the set of *active* atoms of $\mathcal{A}^t$ at the end of iteration t, i.e. the ones that contributed with a non-zero coefficient in the expansion of x^t.

AFW makes use of a *backward* direction also called *away atom*, that is the active atom of largest projection on the gradient direction and formally defined as

Definition 8. (backward direction) *Let x^t be the approximate solution at iteration t, written as a convex combination of atoms and $\widetilde{\mathcal{A}}^t \subset \mathcal{A}$ the set of active atoms. An* away atom *is defined as*

$$a_B^{t+1} := \arg\max_{a \in \widetilde{\mathcal{A}}^t} \langle a, \nabla f(x^t) \rangle$$

and the backward *direction is $a_B^{t+1} - x^t$.*

AFW is stated in Algorithm 4. At each iteration we choose between progressing on the *forward* direction

$$x^{t+1} = (1 - \eta^t)\, x^t + \eta^t a^{t+1}$$

or the *backward* direction

$$x^{t+1} = (1 + \tilde{\eta}^t)\, x^t - \tilde{\eta}^t a_B^{t+1}$$

by taking the one that leads to more progress, i.e. the one more correlated with $-\nabla f(x^t)$. η^t and $\tilde{\eta}$ are chosen by line search in $[0, 1]$ with an additional constraint on $\tilde{\eta}^t$ to make sure that x^{t+1} is feasible.

Algorithm 4 Away steps Frank-Wolfe

repeat
  Compute $a^{t+1} := \operatorname{argmax}_{a \in \mathcal{A}} \langle a, -\nabla f(x^t) \rangle$
  Compute $a_B^{t+1} = \operatorname{argmax}_{a \in \widetilde{\mathcal{A}}^t} \langle a, \nabla f(x^t) \rangle$
  if $\langle -\nabla f(x^t), a^{t+1} - x^t \rangle \geq \langle -\nabla f(x^t), a_B^{t+1} - x^t \rangle$
    $x^{t+1} = (1 - \eta^t)\, x^t + \eta^t a^{t+1}$
  else
    $x^{t+1} = (1 + \tilde{\eta}^t)\, x^t - \tilde{\eta}^t a_B^{t+1}$ with $\tilde{\eta}^t$ such that x^{t+1} remains feasible
until $\langle -\nabla f(x^t), a^{t+1} - x^t \rangle \leq \epsilon$

The idea in PWFW is to move by transferring weight from the away atom a_B^{t+1} to the FW atom a^{t+1}:

$$x_{\text{PWFW}}^{t+1} = x^t + \eta^t\, (a^{t+1} - a_B^{t+1}),$$

where $\eta^t \in [0, c_B^t]$, with $c_B^t \geq 0$ the weight attributed to atom a_B^{t+1} at iteration t, and η^t is found by line search. PWFW is stated in Algorithm 5.The optimal step sizes $\eta^t \in \mathbb{R}$ for FW and PWFW are easily obtained in closed form when f is quadratic. When we have a general f we can use Armijo's line search.

Algorithm 5 Pairwise Frank-Wolfe

repeat
  Compute $a^{t+1} := \operatorname{argmax}_{a \in \mathcal{A}} \langle a, -\nabla f(x^t) \rangle$
  Compute $a_B^{t+1} = \operatorname{argmax}_{a \in \widetilde{\mathcal{A}}^t} \langle a, \nabla f(x^t) \rangle$
  $x^{t+1} = x^t + \eta^t\, (a^{t+1} - a_B^{t+1})$
until $\langle -\nabla f(x^t), a^{t+1} - x^t \rangle \leq \epsilon$

In FCFW, all weights are reoptimized at each iteration:

$$x_{\text{FCFW}}^{t+1} = \operatorname{argmin}_x f(x) \quad \text{s.t.} \quad x \in \text{Convhull}(\mathcal{A}^{t+1}).$$

Algorithm 6 describes FCFW. If $\mathcal{A}^t = \{a_1, ..., a_{k_t}\}$, where $k_t \leq t$ is the number of atoms in $\mathcal{A}^t$, the subproblem that has to be solved at each iteration t rewrites

$$\min_{c \geq 0} f\Big(\sum_{i=1}^{k_t} c_i a_i\Big) \quad \text{s.t.} \quad \sum_{i=1}^{k_t} c_i = 1. \tag{3.4}$$

Algorithm 6 Fully corrective Frank-Wolfe

1: **repeat**
2: Compute $a^{t+1} := \operatorname{argmax}_{a \in \mathcal{A}} \langle a, -\nabla f(x^t) \rangle$
3: $\mathcal{A}^{t+1} \leftarrow \widetilde{\mathcal{A}}^t \cup \{a^{t+1}\}$
4: $x^{t+1} \leftarrow \operatorname{argmin}_x f(x) \quad \text{s.t.} \quad x \in \text{Convhull}(\mathcal{A}^{t+1})$.
5: **until** $\langle -\nabla f(x^t), a^{t+1} - x^t \rangle \leq \epsilon$

Lacoste-Julien and Jaggi (2015) show that PWFW and FCFW converge linearly for strongly convex objectives when $\mathcal{A}$ is finite. Locatello et al. (2017a) show sublinear convergence results on general smooth and convex objectives for a non-empty bounded set of atoms. Locatello et al. (2017b) extend this convergence result to any conic hull of a generic atom set.

3.4.2 Applying Frank-Wolfe to the regularized problem

Beyond constrained optimization problems, the basic conditional gradient algorithm (corresponding to plain FW when f is quadratic) has been generalized to solve problems of the form $\min_x f(x) + \psi(x)$ where the set constraint $C_{\mathcal{A}}$ is replaced by a proper convex function ψ for which the subgradient of $\psi^\star$ can be computed efficiently (Bredies et al., 2009; Yu et al., 2014). Bach (2015) shows that the obtained algorithm can be interpreted as a dual mirror descent algorithm. Yu et al. (2014); Bach (2015) and Nesterov et al. (2015) prove sublinear convergence rates for these algorithms. Corresponding generalizations of PWFW and FCFW are however not obvious. As exploited in Yu et al. (2014); Harchaoui et al. (2015), if $\psi = h \circ \gamma_{\mathcal{A}}$, with h a nondecreasing convex function and $\gamma_{\mathcal{A}}$ an atomic norm, and if an upper bound ρ can be specified a priori on $\gamma_{\mathcal{A}}(x^\star)$ for $x^\star$ a solution of the problem, it can reformulated as

$$\min_{x,\tau} f(x) + \tau \quad \text{s.t.} \quad \gamma_{\mathcal{A}}(x) \leq \tau, \quad \tau \leq \rho, \tag{3.5}$$

and it is natural to apply the different variants of Frank-Wolfe on the variable (x, τ). Note that ρ can be chosen arbitrarily large as long as $\gamma_{\mathcal{A}}(x^\star) \leq \rho$, we can choose for instance $\rho = \gamma_{\mathcal{A}}(x_0)$. FW are part of column generation algorithms that we describe in the next section. In fact we can derive the Frank-Wolfe algorithm for regularized problems from column generation perspective in a more natural way that does not need to introduce the constant ρ.

3.5 Column generation

Column generation(CG) methods are a family of methods that are well known in the optimization literature. Some of the first *working set algorithms* (although called active set then) in the literature on sparsity inducing norms were introduced in Obozinski et al. (2006) and Roth and Fischer (2008). The design of increasingly more sophisticated methods to prune the set of possibly active variables at the optimum has been the object of a significant amount of research until this day (see Ndiaye et al., 2017; Bach et al., 2012b, and reference therein).

Column generation algorithms proceed by solving a sequence of small subproblems of the master problem, referred to as restricted master problems. After computing a solution to the restricted master problem, global optimality conditions are checked. If not satisfied, the inner approximation of the feasible set is improved by some rule that has to be defined.

To explain the precise form of CG for regularized form we first need to discuss optimality conditions.

3.5.1 Optimality conditions and dual problem

We consider the case of atomic gauge regularization, when $\psi := \gamma_{\mathcal{A}}$. With the technical results introduced in Chapter 2, at Equation (2.8), we can derive the optimality conditions for optimization problem (3.1). The next theorem states first order optimality conditions for problem (3.1).

Proposition 3. *Let f be a convex differentiable function and $\gamma_{\mathcal{A}}$ an atomic norm induced by the set of atoms $\mathcal{A}$. Then $x_0 \in \mathbb{R}^p$ is an optimum of problem (3.1) if and only if*

$$-\nabla f(x_0) \in \partial \gamma_{\mathcal{A}}(x_0).$$

Proof. Let us introduce $F := f + \gamma_{\mathcal{A}}$. By Proposition 1, $x_0 \in \mathbb{R}^p$ is an optimum of problem (3.1) if and only if $0 \in \nabla F(x_0)$. □

Let us denote x_0 an optimum of problem (3.1) and $z := -\nabla f(x_0)$. By further using the characterization of the subgradient of an atomic norm of Equation (2.8), the optimality conditions become

$$\langle z, x_0 \rangle = \gamma_{\mathcal{A}}(x_0) \quad \text{and} \quad \forall z \in \mathbb{R}^p \;\; \forall a \in \mathcal{A}, \;\; \langle x, a \rangle \leq 1. \tag{3.6}$$

It is useful to derive the dual of problem (3.1), which is

$$\arg\min_{x \in \mathbb{R}^p} f^*(-s) \quad s.t. \quad \langle s, a \rangle \leq 1, \;\; \forall a \in \mathcal{A}, \tag{3.7}$$

where f^* is the Fenchel conjugate of f.

3.5.2 Deriving the algorithm

Focusing on the formulation regularized by an atomic gauge, that is problem (3.1), the column generation conceptual algorithm follows three steps.

- (restricted master problem) An approximation of the original problem is constructed, and solved, wherein the original set of atoms $\mathcal{A}$ is replaced by a subset $\mathcal{A}^t$. We compute the solution of the problem restricted to a subset of atoms $\mathcal{A}^t$,

$$\arg\min_{x\in\mathbb{R}^p} f(x) + \gamma_{\mathcal{A}^t}(x), \tag{3.8}$$

  where $\gamma_{\mathcal{A}^t}$ is the atomic norm induced by the restricted set of atoms $\mathcal{A}^t$. As shown in Figure 3.5.2 $\gamma_{\mathcal{A}^t}$ is a polyhedral function whose graph is the convex cone generated by rays $a^t - x$. $\gamma_{\mathcal{A}^t}$ is an inner approximation of $\gamma_{\mathcal{A}}$.

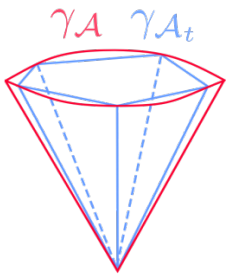


Figure 3.1: Inner approximation of $\gamma_{\mathcal{A}}$ by $\gamma_{\mathcal{A}^t}$

- (optimality conditions) We check optimality conditions for the master problem by using the characterization of Equation (3.6)

$$\forall a \in \mathcal{A}, \ \langle x, a\rangle \leq 1.$$

  If satisfied, the solution has been found and the algorithm stops.

- (column generation problem). If optimality conditions are not satisfied, an auxiliary problem is used to guide the search for a solution to the original problem. Its solution or *column*, here a new atom, is used to improve the inner approximation of the atomic set. A new atom enters current approximation of the atomic set set $\mathcal{A}^t$ following some rule. For instance, an atom a^{t+1} that violates optimality condition, i.e.

$$\left\langle x, a^{t+1}\right\rangle > 1,$$

  is added to the current approximation of the atomic set. At this step it is also possible to remove from $\mathcal{A}^t$ atoms that have zero weight on the current solution x. This option is called *column dropping* and convergence of the algorithm is still guaranteed (Larsson et al., 2015).

The algorithm we just described corresponds to FW applied to the regularized problem. Frank-Wolfe and its variants are an instance of column generation method where the column generation problem is the linear minimization oracle and the restricted master problem is a linear step size update in the case of classical FW and the exact minimization of the objective function over a restricted set of atoms in the case of fully corrective Frank-Wolfe.

3.5.3 Dual problem and cutting planes

Column generation algorithms correspond to cutting plane algorithms in the dual. Cutting planes have extensively been studied in the literature, we refer the reader to Boyd and Vandenberghe (2004) for detailed explanations and Franc et al. (2011) for a review of its use in machine learning. The principle of cutting plane algorithms is to solve a sequence of constrained optimization problems that are relaxations of the original problem, where the constraints introduced are gradually tightening the relaxation in the neighborhood of the optimum. The new constraint introduced at each iteration is called a *cut* since it cuts the previous relaxed constraint set in order to reduce it. A new cut is typically determined as a constraint of the original problem which is violated by a current solution s^t to the relaxed problem. Such a new constraint is called a *deep cut*.

Dual problem of optimization problem (3.1) is of the form

$$\min_{s \in C_{\mathcal{A}}^\circ} f^*(s)$$

where $C_{\mathcal{A}}^\circ = \{s \mid \langle s, a \rangle \leq 1,\ a \in \mathcal{A}\}$ and f^* is the Fenchel conjugate of f. A most violated constraint by a dual variable s can be computed as the inequality $\langle s, a \rangle \leq 1$ for the atom a which is a *direction conjugate* to s, that is a solution to $\max_{a \in \mathcal{A}} \langle s, a \rangle$. Indeed, this yields an atom a such that $\langle a, s \rangle$ is maximal. After t iterations the relaxed problem to solve in the dual is of the form

$$\min_s\ f^*(-s) \quad \text{s.t.} \quad \langle a_i, s \rangle \leq 1,\ \forall i \in [\![t]\!], \tag{3.9}$$

for $\mathcal{A}^t := (a_i)_{i \in [\![t]\!]}$ a sequence of atoms of $\mathcal{A}$. Note that for all $a \in \mathcal{A}$, $\{s \mid \langle a_i, s \rangle = 1\}$ is the hyperplane tangent to $C_{\mathcal{A}}^\circ$ at all points in $\arg\max\{\langle a_i, s \rangle \mid s \in C_{\mathcal{A}}^\circ\}$.

3.5.4 Active set methods for quadratic programming

Active-set algorithms are also an instance of column generation methods. Active-set methods are usually applied in the context of general convex quadratic problems with linear constraints and efficient implementations have been studied (see Nocedal and Wright (2006), Chapter 16 and Forsgren et al. (2015)). Before introducing active-set algorithm we review optimality conditions for inequality constrained quadratic problems. Then we present its application to the well known algorithm of min-norm point (Wolfe, 1976b) and we clarify how it can be

applied to the setup of atomic norm regularization.

Consider the following convex quadratic problem

$$\begin{aligned} \min_x \quad & \frac{1}{2} x^\top H x + g^\top x \\ \text{s.t.} \quad & Ax = b \\ & Cx \geq d \end{aligned} \tag{3.10}$$

where H is positive semidefinite. We can derive a set of optimality conditions by applying KKT[1] conditions as stated in the following proposition (proof in Nocedal and Wright (2006), Chapter 16, Theorem 16.4).

Proposition 4. *(optimality conditions) Consider the convex quadratic problem (3.10). If there exists a pair* (x^*, λ^*) *verifying the conditions*

$$Hx^* + g - \sum_i \lambda_i^* c_i = 0 \tag{3.10 a}$$

$$Ax^* = b \tag{3.10 b}$$

$$Cx^* \geq d \tag{3.10 c}$$

$$\lambda_i^* \geq 0 \quad \forall i \quad s.t. \quad c_i^\top x^* = d_i, \tag{3.10 d}$$

then, x^* *is a global solution of* (3.10)*.* λ_i^* *are the Lagrangian multipliers.*

We now introduce active set method for quadratic programming. We describe the *primal* active-set method where the iterates remain primal feasible while infeasibilities of the dual inequalities tend to zero. The goal is to predict the optimal active-set $S^* \subset [\![p]\!]$ that only contains the constraints that are satisfied with equality at the solution of the problem. Active-set algorithm solves a sequence of equality constrained quadratic programs

$$\begin{aligned} \min_x \quad & \frac{1}{2} x^\top H x + g^\top x \\ \text{s.t.} \quad & Ax = b \\ & c_i^\top x_i = d \quad \forall i \in S. \end{aligned} \tag{3.11}$$

where S is a subset of indexes.

We start at a feasible point of main problem (3.10) with an empty set S of indexes. S is called the *active set* and its complementary is called the *working set*. At each iteration we update the active-set S by performing the following steps:

- We compute $\tilde{x}$, the solution to the restricted problem (3.11).

[1] Karush–Kuhn–Tucker

- If the solution $\tilde{x}$ is primal feasible, we perform a *full-step* and update $\{x \leftarrow \tilde{x}\}$. Then, we check dual feasibility condition (3.10 d). If dual feasibility is also satisfied we terminate and if not we remove one of the blocking constraints from S (an index i such that $\lambda_i^* < 0$).
- If the solution $\tilde{x}$ is not primal feasible, we compute the most blocking constraint i^* (index i such that $\{c_i^\top x_i - d_i\}$ is minimized). Then we update x^{t+1} as the closest point to $\tilde{x}$ in the segment $[x^t, \tilde{x}]$ such that it remains primal feasible. Finally we add the blocking constraint i^* to S (we remove it from the working set), which is called a *drop-step*.

The previous algorithm terminates in a finite number of iterations because it strictly decreases the quadratic cost function at each iteration. Appart from *primal* active-set methods, other classes of active-set algorithms have been proposed. Forsgren et al. (2016) describes in detail *dual* and *primal-dual* versions of active-set.

Example of min-norm point

As an example, we describe the well known algorithm of min-norm point presented by Wolfe (1976b) as an instance of active-set algorithm. Minimum norm point algorithm is used for finding the minimum-norm point in the convex hull of a given finite set of points $\mathcal{P} := \{p_1, p_2, ..., p_k\}$ in $\mathbb{R}^p$. The optimization problem writes

$$\min \frac{1}{2}\|x\|_2^2 \quad \text{s.t.} \quad x \in C_{\mathcal{P}},$$

where $C_{\mathcal{P}}$ is the convex hull of $\mathcal{P}$

$$C_{\mathcal{P}} := \{x \in \mathbb{R}^p | x = \sum_{i=1}^{k} \eta_i p_i \ \text{ s.t. } \ \sum_{i=1}^{k} \eta_i = 1 \ \text{ and } \ \eta \geq 0\}.$$

The principle of minimum-norm-point algorithm is to solve a sequence of constrained optimization problems that are relaxations of the original problem, where $C_{\mathcal{P}}$ is replaced by an inner approximation, a simplex of a subset of $\mathcal{P}$. Updating such a simplex requires a solution of a linear optimization problem over the convex hull $C_{\mathcal{P}}$. In most cases we do not want to evaluate the functions at all points as there can be exponential number of points compared to the dimensions of the problem. The algorithm is applicable if the LMO

$$\text{LMO}_{\mathcal{P}}(x) := \arg\min_{p_i \in \mathcal{P}} \langle x, p_i \rangle$$

can be solved efficiently, which can be done for flow constraints or generally for submodular polytope constraints. We describe the min-norm point in Algorithm 7.

Application to regularization by atomic gauges

In this part we anticipate somewhat the result present in next chapter. We clarify why convex quadratic problems with linear constraints are relevant in the context of regularization by

Algorithm 7 Minimum-norm point algorithm

Initialization: $x^0 \coloneqq \sum_{j\in J_0} \eta_j p_j$ feasible point, $J_0 \subset \{1,\dots,k\}$
for $t = 1, ...$
$\tilde{\eta} \leftarrow \frac{1}{2} \arg\min_x \| \sum_{j\in J_t} \eta_i p_i \|_2^2 \quad \text{s.t.} \quad 1^\top \eta_{J_t} = 1$
  if $\tilde{\eta}_{J_t} \geq 0$,
    $\eta^{t+1} \leftarrow \tilde{\eta}$ ▷ full-step
    $\mathbf{x}^{t+1} \leftarrow \sum_{j\in J_t} \eta_i p_i$
    $p_{k^*} \leftarrow \mathrm{LMO}_{\mathcal{P}}(x^{t+1})$
    if ${x^{t+1}}^\top p_{k^*} \geq \|x^{t+1}\|_2^2$, **then stop,** ▷ optimality check
    else $J_{t+1} \leftarrow J_t \bigcup \{k^*\}$ **end**
  else
    $\tau \leftarrow \arg\max\{\tau \mid \tilde{\eta}_{J_t} + \tau(x^t_{J_t} - \tilde{\eta}_{J_t}) > 0\}$
    $K \leftarrow \arg\max\{i \mid \tilde{\eta}_i + \tau(x^t_i - \tilde{\eta}_i) = 0\}$
    $J_{t+1} \leftarrow J_t \backslash \{K\}$ ▷ drop-step
    $\eta^{t+1} \leftarrow \tilde{\eta}_{J_t} + \tau(x^t_{J_t} - \tilde{\eta}_{J_t})$
end

atomic gauges. Let us consider a fixed set $\mathcal{A}$ with a finite number of atoms. What we explain in this section does not apply for general(possibly infinite) $\mathcal{A}$. The application to that case will be discussed in the next chapter. Provided that f is quadratic, the regularized problem (3.1) can be rewritten

$$\min f\left(\sum_{a\in\mathcal{A}} c_a a\right) + \sum_{a\in\mathcal{A}} c_a \quad \text{s.t.} \quad \forall a \in \mathcal{A},\ c_a \geq 0 \tag{3.12}$$

by applying a simple change of variable $x = \sum_{a\in\mathcal{A}} c_a a$. Hence we obtain a quadratic problem with positivity constraints. More explicitly, if A denotes the matrix where the columns are the atoms[2] $a \in \mathcal{A}$ and parameters (Q, b) define the quadratic function f, i.e. $f(x) \coloneqq 1/2 x^\top Q x + b^\top x$, the problem (3.12) is reformulated as

$$\min \frac{1}{2} c^\top A^\top Q A c + (b+1)^\top c \quad \mathbf{s}.t. \quad c \geq 0, \tag{3.13}$$

which is a convex quadratic problems with positivity constraints.

[2] if atoms are matrices or tensors we can vectorize them

Chapter 4

Fast column generation for atomic norm regularization

This Chapter is based on our paper Vinyes and Obozinski (2017). We consider optimization problems that consist in minimizing a quadratic function under an atomic norm regularization or constraint. In the line of work on conditional gradient algorithms, we show that the fully corrective Frank-Wolfe (FCFW) algorithm - which is most naturally reformulated as a column generation algorithm in the regularized case - can be made particularly efficient for difficult problems in this family by solving the simplicial or conical subproblems produced by FCFW using a special instance of a classical active set algorithm for quadratic programming (Nocedal and Wright, 2006) that generalizes the min-norm point algorithm (Wolfe, 1976a). Our experiments show that the algorithm takes advantages of warm-starts and of the sparsity induced by the norm, displays fast linear convergence, and clearly outperforms the state-of-the-art, for both complex and classical norms, including the standard group Lasso.The code for experiments is available at `https://github.com/vinyesm/fcgan`.

4.1 Introduction and related work

A number of problems in machine learning and structured optimization involve either structured convex constraint sets that are defined as the intersection of a number of simple convex sets or dually, norms of sets that are defined as convex hull of either extreme points or of a collection of sets. A broad class of convex regularizers that can be used to encode a priori knowledge on the structure of the objects to estimate have been described as *atomic norms* and *atomic gauges* by Chandrasekaran et al. (2012). The concept of atomic norm has found several applications to design sparsity inducing norms for vectors (Jacob et al., 2009; Obozinski et al., 2011), matrices (Richard et al., 2014; Foygel et al., 2012) and tensors (Tomioka and Suzuki, 2013; Liu et al., 2013; Wimalawarne et al., 2014).

A number of these norms remain difficult to use in practice because it is in general not possible to compute the associated *proximal operator* or even the norm itself at a reasonable

cost. However, the dual norm which is defined as a supremum of dot products with the atoms that define the norm can often be computed efficiently because of the structure of the set of atoms. Also a number of atomic norms are actually naturally defined as *infimal convolution* of other norms (Jacob et al., 2009; Tomioka and Suzuki, 2013; Liu et al., 2013) and this structure has been used to design either block-coordinate descent approaches or dual ADMM optimization schemes (Tomioka and Suzuki, 2013) involving latent variables associated with the elementary norms convolved.

In this chapter, we propose to solve problems regularized or constrained by atomic norms using a fully corrective Frank-Wolfe algorithm—which can be reformulated as simple column generation algorithm in the regularized case—combined with a dedicated active-set algorithm for quadratic programming. Our experiments show that we achieve state-of-the-art performance. We also include a formal proof of the correspondance between the column generation algorithm and Fully Corrective Frank-Wolfe.

After a review of the concept of atomic norms, as well some illustrations, we present a number of the main algorithmic approaches that have been proposed. We then present the scheme we propose and finally some experiments on synthetic and real datasets.

4.1.1 Notations

$[\![p]\!]$ denotes the set $\{1, \dots, p\}$. If $x \in \mathbb{R}^p$, x_G denotes the subvector of x whose entries are indexed by a set $G \in [\![p]\!]$. Given a function ψ, ψ^* denotes its Fenchel conjugate $\psi^*(s) := \max_x \langle s, x\rangle - \psi(x)$. $\|M\|_{\text{tr}}$ denotes the trace norm of the matrix M defined as the ℓ_1-norm of its singular values.

4.2 Atomic norms

In many machine learning applications, and particularly for ill-posed problems, models are constrained structurally so they have a simple representation in terms of some fundamental elements. Examples of such elements include sparse vectors for many sparsity inducing norms, rank-one matrices for the trace norm or low-rank tensors as used in the *nuclear tensor norm* (Liu et al., 2013). We call *atoms* these elements and *atomic set* $\mathcal{A}$ their (possibly infinite) collection. Assuming $\mathcal{A}$ is bounded and centrally symmetric, and provided its convex hull $C_{\mathcal{A}}$ has non empty interior, we can define an atomic norm $\gamma_{\mathcal{A}}$ as the norm of unit ball $C_{\mathcal{A}}$. It can be shown that (in a finite dimensional space) $\gamma_{\mathcal{A}}(x) := \inf\{\sum_{a\in\mathcal{A}} c_a \mid \sum_{a\in A} c_a a = x,\ c_a \geq 0,\ a \in \mathcal{A}\}$. The polar norm or dual norm is defined as: $\gamma^\circ_{\mathcal{A}}(s) := \sup_{a\in\mathcal{A}} \langle s, a\rangle$. If $\mathcal{A}$ is not symmetric, or if $C_{\mathcal{A}}$ is empty, as long as $\mathcal{A}$ contains the origin and is closed, $\gamma_{\mathcal{A}}$ can still be defined as a *gauge* instead of a norm and the theory and algorithms presented in this chapter still apply. We restrict the discussion to norms for simplicity. For a reference on gauges, see Rockafellar (1970).

We consider in this chapter formulations in which an atomic norm is used as a regularizer,

and which lead to an optimization problem of the form

$$\min_{x \in \mathbb{R}^p} f(x) + \gamma_{\mathcal{A}}(x), \tag{4.1}$$

where f is a *quadratic* function. The case where f is more generally twice differentiable is obviously of interest, but beyond the scope of this work.

4.2.1 Examples of atomic norms

Lasso. The Lasso is a natural example of atomic norm, whose atoms are the $(\pm e_i)_{i \in [\![p]\!]}$, where the $(e_i)_{i \in [\![p]\!]}$ is the canonical basis of $\mathbb{R}^p$. The Lasso polar norm is defined as $\Omega^\circ_{\text{Lasso}}(s) = \max_{i \in [\![p]\!]} |s_i|$.

Latent group lasso (LGL)

The norms introduced in Jacob et al. (2009) are a strong motivating example. For instance Obozinski and Bach (2012) show that a broad family of tight relaxations for structured sparsity can be written in LGL form. Given a collection of sets $\mathcal{B}$ covering $[\![p]\!]$ and which can overlap, and fixed positive weights δ_B for each set $B \in \mathcal{B}$, the atoms of LGL norm are the vectors of norm δ_B^{-1} and support in $\mathcal{B}$. The polar LGL norm is defined as $\Omega^\circ_{\text{LGL}}(s) = \max_{B \in \mathcal{B}} \delta_B^{-1} \|s_B\|_2$. In the particular case where $\mathcal{B}$ form a partition of $[\![p]\!]$ we recover the group Lasso norm. Maurer and Pontil (2012) consider a generalization to a broader family of atomic norms with dual norms of the form $\sup_{M \in \mathcal{M}} \|Ms\|_2$, where $\mathcal{M}$ is a collection of operators. Matrix counterparts of the latent group Lasso norms are the *latent group trace norms* (Tomioka and Suzuki, 2013; Wimalawarne et al., 2014).

Additive decompositions

There has been interest in the literature for additive matrix decompositions (Agarwal et al., 2012), the most classical example being "sparse+low rank decompositions" which have been proposed for robust PCA and multitask learning (Candès et al., 2011; Chandrasekaran et al., 2011). This formulation leads to a problem of the form $\min_{L,S} f(L+S) + \mu\|L\|_{\text{tr}} + \lambda\|S\|_1$, which under the form $\min_M f(M) + \gamma_{\mathcal{A}}(M)$ with $\gamma_{\mathcal{A}}$ the atomic norm where $\mathcal{A} \subset \mathbb{R}^{p_1 \times p_2}$ is defined as

$$\begin{aligned} \mathcal{A} &:= \lambda\, \mathcal{A}_1 \cup \mu\, \mathcal{A}_{\text{tr}}, \qquad \text{where} \\ \mathcal{A}_1 &:= \left\{ \pm\, e_i e_j^\top,\ (i,j) \in [\![p_1]\!] \times [\![p_2]\!] \right\}, \\ \mathcal{A}_{\text{tr}} &:= \left\{ uv^\top,\ \|u\|_2 = \|v\|_2 = 1 \right\}. \end{aligned}$$

As a consequence, $C^\circ_{\mathcal{A}} = \frac{1}{\lambda} C^\circ_1 \cap \frac{1}{\mu} C^\circ_{\text{tr}}$ with C°_1 a unit ℓ_∞ ball and C°_{tr} a unit spectral norm ball.

Convex sparse SVD and PCA

A third example are the norms introduced in Richard et al. (2014), including the (k, q)-trace norm for which

$$\mathcal{A} := \bigcup\{\mathcal{A}_{I,J} \mid (I, J) \subset [\![p_1]\!] \times [\![p_2]\!],\ |I| = k, |J| = q\},$$
$$\text{with} \qquad \mathcal{A}_{I,J} := \{uv^\top \in \mathcal{A}_{\text{tr}} \mid \|u\|_0 \le k,\ \|v\|_0 \le q\},$$

and the sparse-PCA norm[1] for which

$$\mathcal{A} := \bigcup\{\mathcal{A}_{I,\geq} \mid I \subset [\![p_1]\!],\ |I| = k\},$$

with $\mathcal{A}_{I,\geq} := \{uu^\top \mid u \in \mathcal{A}_I\}$, and $\mathcal{A}_I$ defined like $\mathcal{A}_B$ for LGL.

Beyond these examples a number of structured convex optimization problems encountered in machine learning and operations research that involve combinatorial or structured tasks such as finding permutations or alignments, convex relaxation of structured matrix factorization problems (Bach et al., 2008; Ding et al., 2010), Procrustes analysis, etc, involve difficult convex constraint sets such as elliptope, the Birkhoff polytope, the set of doubly nonnegative matrices that are naturally written (themselves or their polar) as intersections of simpler sets such as the p.s.d. cone, the positive orthant, simplices, hypercubes, etc, and which lead to optimization problems whose duals are regularized by associated atomic norms.

4.3 Previous approaches

Conditional gradient algorithms

For many[2] of these norms, it is assumed that an efficient algorithm is available to compute $\operatorname{argmax}_{a\in\mathcal{A}}\langle a, s\rangle$. For the case of the constrained problem

$$\min_x f(x) \quad \text{s.t.} \quad x \in C_{\mathcal{A}}, \tag{4.2}$$

this has motivated a number of authors to suggest variants of the *conditional gradient algorithm*, also known as the Frank-Wolfe algorithm when the objective is quadratic, as a tool of choice to solve problems with atomic norm constraints. Indeed, the principle of conditional gradient algorithms is to build a sequence of approximations to the solution of the problem as convex combinations of extreme points of the constraint sets, which here correspond to atoms, so that the expansion take the form $x = \sum_{i=1}^t c_i a_i$ with $\sum_{i=1}^t c_i = 1$. This procedure guarantees a feasible sequence. At each iteration a new atom, also called Frank-Wolfe direction or *forward* direction, is added in the expansion. This atom is the extreme point of the constraint set defined by $a^{t+1} := \operatorname{argmax}_{a\in\mathcal{A}}\langle a, -\nabla f(x^t)\rangle$. The Frank-Wolfe (FW) algorithm writes

$$x_{\text{FW}}^{t+1} = (1 - \eta^t)\, x^t + \eta^t a^{t+1},$$

[1] In fact this is not a *norm* but only a *gauge*.

[2] This is not true for the norms introduced in Richard et al. (2014) whose dual are NP-hard to compute, but for which reasonable heuristic algorithms or relaxations are available.

where $\eta^t \in [0,1]$ is a scalar stepsize and $x^0 = 0$. It can be set to $\frac{1}{1+t}$ or found by line search.

Other variants of FW algorithms have been proposed, notably, FW with away steps (which we do not describe here), pairwise FW (PWFW) and fully corrective Frank-Wolfe (FCFW). We refer the reader to Lacoste-Julien and Jaggi (2015) for a detailed presentation and summarize hereafter the form of the different updates for PWFW and FCFW. The active set of atoms $\mathcal{A}^t$ at time t is recursively defined by $\mathcal{A}^{t+1} = \widetilde{\mathcal{A}}^t \cup \{a^{t+1}\}$ with $\widetilde{\mathcal{A}}^t$ the set of *active* atoms of $\mathcal{A}^t$ at the end of iteration t, i.e. the ones that contributed with a non-zero coefficient in the expansion of x^t.

PWFW makes use of a *backward* direction also called *away atom*, and defined as $a_B^{t+1} = \operatorname{argmax}_{a \in \widetilde{\mathcal{A}}^t} \langle a, \nabla f(x^t) \rangle$, i.e. it is the active atom of largest projection on the gradient direction. The idea in PWFW is to move by transferring weight from the away atom a_B^{t+1} to the FW atom a^{t+1}:

$$x_{\text{PWFW}}^{t+1} = x^t + \eta_p^t \, (a^{t+1} - a_B^{t+1}),$$

where $\eta^t \in [0, c_B^t]$, with $c_B^t \geq 0$ the weight attributed to atom a_B^{t+1} at iteration t, and η^t is found by line search. The optimal step sizes $\eta^t \in \mathbb{R}$ for FW and PWFW are easily obtained in closed form when f is quadratic.

In FCFW, all weights are reoptimized at each iteration:

$$x_{\text{FCFW}}^{t+1} = \operatorname{argmin}_x f(x) \quad \text{s.t.} \quad x \in \text{Convhull}(\mathcal{A}^{t+1}).$$

If $\mathcal{A}^t = \{a_1, ..., a_{k_t}\}$, where $k_t \leq t$ is the number of atoms in $\mathcal{A}^t$, the subproblem that has to be solved at each iteration t of FCFW rewrites

$$\min_{c \geq 0} f\Big(\sum_{i=1}^{k_t} c_i a_i \Big) \quad \text{s.t.} \quad \sum_{i=1}^{k_t} c_i = 1. \tag{4.3}$$

Lacoste-Julien and Jaggi (2015) show that PWFW and FCFW converge linearly for strongly convex objectives when $\mathcal{A}$ is finite.

Rao et al. (2015) propose a variant of FCFW to solve (4.2) for f smooth and specifically for atomic norm constraints, with an enhancing "backward step" which applies hard-thresholding to the coefficients c^t. To solve (4.3) they use a projected gradient algorithm.

Beyond constrained optimization problems, the basic conditional gradient algorithm (corresponding to plain FW when f is quadratic) has been generalized to solve problems of the form $\min_x f(x) + \psi(x)$ where the set constraint $C_{\mathcal{A}}$ is replaced by a proper convex function ψ for which the subgradient of $\psi^\star$ can be computed efficiently (Bredies et al., 2009; Yu et al., 2014). Bach (2015) shows that the obtained algorithm can be interpreted as a dual mirror descent algorithm. Yu et al. (2014); Bach (2015) and Nesterov et al. (2015) prove sublinear convergence rates for these algorithms. Corresponding generalizations of PWFW and FCFW are however not obvious. As exploited in Yu et al. (2014); Harchaoui et al. (2015), if $\psi = h \circ \gamma_{\mathcal{A}}$,

with h a nondecreasing convex function and $\gamma_{\mathcal{A}}$ an atomic norm, and if an upper bound ρ can be specified a priori on $\gamma_{\mathcal{A}}(x^\star)$ for $x^\star$ a solution of the problem, it can reformulated as

$$\min_{x,\tau} f(x) + h(\tau) \quad \text{s.t.} \quad \gamma_{\mathcal{A}}(x) \leq \tau, \quad \tau \leq \rho, \tag{4.4}$$

and it is natural to apply the different variant of Frank-Wolfe on the variable (x, τ), because the FW direction is easy to compute (see Section 4.4.1).

Proximal block-coordinate descent

In the context where they are applicable, proximal gradient methods provide an appealing alternative to Frank-Wolfe algorithms. However, the former require to be able to compute efficiently the *proximal operator* of the norm $\gamma_{\mathcal{A}}$ appearing in the objective, which is typically more difficult to compute than the Frank-Wolfe direction.

For a number of atomic norms, we have $\mathcal{A} = \bigcup_{j=1}^{J} C_j$ where C_j are convex sets. As a consequence the polar norm takes the form $\gamma_{\mathcal{A}}^\circ(s) = \max_j \gamma_{C_j}^\circ$, with γ_{C_i} the atomic norm (or gauge) associated with the set C_i, and it is a standard result that

$$\gamma_{\mathcal{A}}(x) = \inf\{ \gamma_{C_1}(z_1) + \ldots + \gamma_{C_J}(z_J) \mid z_1 + \ldots + z_J = x\}.$$

Technically, $\gamma_{\mathcal{A}}$ is called the *infimal convolution* of the norms $(\gamma_{C_i})_i$ (see Rockafellar, 1970). In fact most of the norms that we presented in section 4.2.1 are of this form, including LGL norms, latent group trace norms, norms arising from additive decomposition (obviously by construction), and the norms for sparse SVD and sparse PCA.

For all these norms, problem (4.1) can be reformulated as

$$\min_{z_1,\ldots,z_J} f(z_1 + \ldots + z_J) + \gamma_{C_1}(z_1) + \ldots + \gamma_{C_J}(z_J).$$

Since the objective is then a sum of a smooth and of a separable function, randomized proximal block-coordinate descent algorithm are typical candidates. These algorithms have attracted a lot of attention in the recent literature (see Hong et al., 2013, and reference therein) and have been applied successfully to a number of formulations involving convex sparsity inducing regularizers (Friedman et al., 2010; Shalev-Shwartz and Tewari, 2011; Gu et al., 2016), where they achieve state-of-the-art performance. Such BCD algorithms where the ones proposed for the norms proposed in Jacob et al. (2009) and Richard et al. (2014).

Unfortunately these algorithms are slow in general even if f is strongly convex because of the composition with the linear mapping $(z_1, \ldots, z_J) \mapsto z_1 + \ldots + z_J$. Intuitively if the atoms of the different norms are similar, then the formulation is badly conditioned. If they are different or essentially decorrelated, BCD remains one of the most efficient algorithms (Shalev-Shwartz and Tewari, 2011; Gu et al., 2016).

4.4 Pivoting Frank Wolfe

After reviewing the form of the corrective step of FCFW and reformulating FCFW in the regularized case as a column generation algorithm, we introduce active-set algorithms to solve efficiently sequences of corrective steps.

4.4.1 Simplicial and conical subproblems

We focus on the sequence of subproblems that need to be solved at the corrective step of FCFW. Let $k_t := |\mathcal{A}^t|$ be the number of selected atoms at iteration t, and $A^t \in \mathbb{R}^{p \times k_t}$, the matrix whose columns are the atoms $\mathcal{A}^t$, then, for the constrained problem (4.2), the subproblem is the *simplicial* problem:

$$\min_c f(A^t c) \quad \text{s.t.} \quad c \in \Delta^{k_t}, \tag{4.5}$$

with $\Delta^k := \{c \in \mathbb{R}^k_+ \mid \sum_{i=1}^k c_i = 1\}$ the canonical simplex. The regularized problem (4.1) can be reformulated as the constrained optimization problem (4.4) on a truncated cone, provided the truncation level ρ is an upper bound of the value of $\gamma_{\mathcal{A}}$ at the optimum. Actually, if ρ is sufficiently large, several Frank-Wolfe algorithms do not depend any longer on the value of ρ and can be interpreted as algorithms in which whole extreme rays of the cone $\{(x;\tau) \mid \gamma_{\mathcal{A}}(x) \leq \tau\}$ enter the active set via the linear minimization oracle, and where the original cone is locally approximated from inside by the simplicial cone obtained as their conical hull. In particular in the case of FCFW, the subproblem considered at the t-th iteration takes the form of the *conical* problem

$$\min_c f(A^t c) + \sum_i c_i \quad \text{s.t.} \quad c \geq 0, \tag{4.6}$$

which is simply a Lasso problem with positivity constraints when f is quadratic. The fact that problem (4.1) can be solved by as sequence of problems of the form (4.6) is shown in Harchaoui et al. (2015, Sec. 5), who argue that this leads to an algorithm no worse and possibly better. We formally show that the simple column generating scheme presented as Algorithm 8 is in fact exactly equivalent to FCFW applied to the truncated cone formulation as soon as ρ is large enough:

Proposition 5. *If f is assumed lower bounded by 0 and if $\rho > f(0)$, or more generally if the level sets of $x \mapsto f(x) + \gamma_{\mathcal{A}}(x)$ are bounded and ρ is sufficiently large, then the sequence $(\bar{x}^t)_t$ produced by the FCFW algorithm applied to the truncated cone constrained problem (4.4) and initialized at $(\bar{x}^0;\tau^0) = (0;0)$ is the same as the sequence $(x^t)_t$ produced by Algorithm 8 initialized with $x^0 = 0$, with equivalent sequences of subproblems, active sets and decomposition coefficients.*

See the appendix for a proof.

Figure 4.4.1 illustrates Algorithm 8, where the atomic gauge $\gamma_{\mathcal{A}}$ is inner approximated by a gauge on a subset of atoms $\gamma_{\mathcal{A}^t}$. As discussed as well in the appendix, a variant of Algorithm 8 without pruning of the atoms with zero coefficients (at step 7) is derived very naturally as the dual of a cutting plane algorithm.

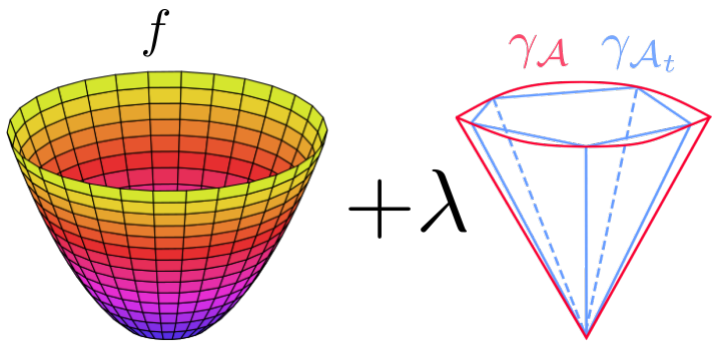


Figure 4.1: Illustration of column generation Algorithm 8

Algorithm 8 Column generation

Require: f convex differentiable, tolerance ϵ
Initialization: $x^0 = 0$, $A^0 = \varnothing$, $k_0 = 0$, $t = 1$
repeat
  $a_t \leftarrow \arg\max_{a \in \mathcal{A}} \langle -\nabla f(x^{t-1}), a \rangle$
  $A^t \leftarrow [A^{t-1}, a_t]$
  $c^t \leftarrow \arg\min_{c \geq 0} f\big(A^t c\big) + \|c\|_1$
  $I \leftarrow \{i \mid c_i^t > 0\}$,
  $c^t \leftarrow c_I^t$
  $A^t \leftarrow A_{\cdot I}^t$
  $x^t \leftarrow A^t c^t$
  $t \leftarrow t + 1$
until $\max_{a \in \mathcal{A}} \langle -\nabla f(x^{t-1}), a \rangle \leq \epsilon$

4.4.2 Leveraging active-set algorithms for quadratic programming

Problems (4.5) and (4.6) can efficiently be solved by a number of algorithms. In particular, an appropriate variant LARS algorithm solves both problem in a finite number of iterations and it is fast if the solution in sparse, in spite of the fact that it solves exactly a sequence of linear systems. Interior point algorithms can always be used, and are often considered to be a natural choice to solve this step in the literature. For larger scale problems, and if f has Lipschitz gradients (which is obviously the case for a quadratic function), the forward-backward proximal algorithm can be used as well, since the projection on the simplex for (4.5) and the asymmetric soft-thresholding for (4.6) can be computed efficiently. For the constrained case, this is the algorithm used by Rao et al. (2015).

In our case, we need to solve a sequence of problems of the form (4.5) or (4.6), that differ each from the previous one by the addition of a single atom. So being able to use *warm-starts* is key! If the simplicial problems remains of small size, and if the corresponding Hessians can be computed efficiently, using second order algorithms is likely to outperform first order methods.

But the LARS and interior point methods cannot take advantage of warm-starts. Thus, when f is quadratic, we propose to use *active set algorithms for convex quadratic programming* (Nocedal and Wright, 2006; Forsgren et al., 2015). In particular, following[3] Bach (2013, Chap. 7.12), we propose to apply the active-set algorithm of Nocedal and Wright (2006, Chap. 16.5) to iteratively solve (4.5) and (4.6). This algorithm takes the very simple[4] form of Algorithm 9. In fact, as noted in Bach (2013, Chap. 9.2), this algorithm is a generalization of the famous *min-norm point algorithm* (Wolfe, 1976a), the latter being recovered when the Hessian is the identity.

Algorithm 9 is illustrated in Figure 4.2. The obtained iterates always remain in the positive orthant (i.e. primal feasible). Each update of c in Algorithm 9 is called a *pivot*, which is either *full-step* or *drop-step*. Given a collection of active atoms indexed by a set J, the solution d of the non-constrained quadratic program restricted to this set of atoms and obtained by removing the positivity constraints is computed (line 4). If d lies in the positive orthant, we set $c = d$, and we say that we perform a *full-step*. In that case, the index of an atom that must become active (if any), based on gradients, is added to J. If $d \notin \mathbb{R}_+^{|J|}$, a *drop-step* is performed: c is updated as the intersection between segment $[c_{\text{old}}, d]$ and the positive orthant, and the index i such that $c_i = 0$ is dropped from J (line 13). The algorithm stops if after a full-step, no new index is added in J.

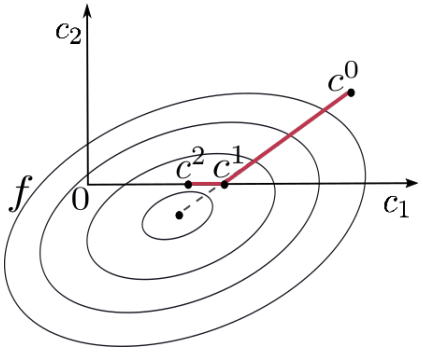


Figure 4.2: Illustration of Algorithm 9. Here, it converges after a *drop-step* (variable c_2 is dropped) leading to c^1 followed by a *full-step* (along c_1) leading to c^2.

4.4.3 Connection with cutting plane algorithms

It is well known that the Frank-Wolfe algorithm is an instance of a column generation algorithm (Forsgren et al., 2015). We explain in this section how Algorithm 1 is naturally

[3] Bach (2013) proposed to use this active-set algorithm to optimize convex objectives involving the Lovász extension of a submodular function.

[4] Despite the fact that, in the context of a simplicial algorithms, the polyhedral constraints sets of (4.5) and (4.6) as convex hulls, the algorithm of Nocedal and Wright (2006, Chap. 16.5) actually exploits their structure as intersections of half-spaces, and thus the active constraints of the algorithm actually correspond counter-intuitively to dropped atoms.

Algorithm 9 $[c, J]$=Active-set(H, b, c_0, J_0)

Solves: $P := \min_c c^\top Hc + b^\top c, \quad \text{s.t.} \quad c \geq 0$
Initialization: $c = c_0$, $J = J_0$,
repeat
  $d \leftarrow {H_{J,J}}^{-1} b_J$
  if $d \geq 0$,
    $c \leftarrow d$ ▷ full-step
    $g \leftarrow Hc + b$
    $k \leftarrow \arg\min_{i \in J_0 \setminus J} g_i$
    if $g_k \geq 0$, **then break, else** $J \leftarrow J \bigcup \{k\}$ **end**
  else
    $i^* \leftarrow \arg\min_i \frac{c_i}{c_i - d_i}$ s.t. $c_i - d_i > 0$, $d_i < 0$
    $\tau \leftarrow \frac{c_{i^*}}{c_{i^*} - d_{i^*}}$
    $J \leftarrow J \setminus \{i^*\}$ ▷ drop-step
    $c \leftarrow c + \tau(d - c)$
  end
until $g_{J_0 \setminus J} \geq 0$
return c, J

derived as such.

Column generation algorithms correspond to cutting plane algorithms in the dual. The principle of the latter algorithms is to solve a sequence of constrained optimization problems that are relaxations of the original problem, where the constraints introduced are gradually tightening the relaxation around the optimum. The new constraint introduced at each iteration is called a *cut* since it cuts the previous relaxed constraint set in order to reduce it. A new cut is typically determined as a constraint of the original problem which is violated by a current solution s^t to the relaxed problem. Such as new constraint is called a *deep cut*. For problems of the form $\min_{s \in C_\mathcal{A}^\circ} f^*(s)$ and given that $C_\mathcal{A}^\circ = \{s \mid \langle s, a \rangle \leq 1,\ a \in \mathcal{A}\}$, a most violated constraint by a dual variable s can be computed as the inequality $\langle s, a \rangle \leq 1$ for the atom a which is a *conjugate direction* to s, that is a solution to $\max_{a \in \mathcal{A}} \langle s, a \rangle$. Indeed, this yields an atom a such that $\langle a, s \rangle$ is maximal.

After t iterations the relaxed problem to solve in the dual is of the form

$$\min_s \ f^*(-s) \quad \text{s.t.} \quad \langle a_i, s \rangle \leq 1,\ \forall i \in [\![t]\!], \tag{4.7}$$

for $\mathcal{A}^t := (a_i)_{i \in [\![t]\!]}$ a sequence of atoms of $\mathcal{A}$.

It is immediate to check that the corresponding primal algorithm is a version of Algorithm 1 in which all atoms are stored. The classical constrained version of Frank-Wolfe correspond a cutting plane algorithm in the dual problem regularized by the dual norm, where this regularization is reformulated as a conic constraint like in formulation (4) in the main paper.

4.5 Convergence and computational cost

In this section, we discuss first the convergence of the algorithm and the number of pivots needed for convergence, and the the cost of each pivot.

Algorithm 9 is an instance of min-norm point (MNP) with a general quadratic instead of Euclidean distance, but the algorithm is affine invariant, so the convergence is the same. MNP is known to be finitely convergent. The positive orthant in dimension k_t has at most 2^{k_t} faces which is a naive bound on the number of pivots in the active-set at iteration t of FCFW. But, Lacoste-Julien and Jaggi (2015) prove that MNP is linearly convergent. In practice, the solution is most of the time either strictly inside the orthant or in one of the $k-1$ dimensional faces in which case it is in fact found in just 1 or respectively 2 iterations! The number of pivots per call is illustrated in Figure 4.6.2 upper left.

Let $s = \max_{a \in \mathcal{A}} \|a\|_0$ be the sparsity of the atoms, k the number of active atoms at iteration t and $H^t = A^{t\top} Q A^t$ the Hessian of the quadratic problem in the active set, where Q is the Hessian of the quadratic function f.

The cost of one pivot is the cost of computing the Hessian H^t and its inverse, which is $\mathcal{O}\big(\min(k^2 s^2, kps + k^2 s)\big)$ for building the Hessian and an extra $\mathcal{O}\big(k^3\big)$ for the inversion. In the active-set with warm starts we only add or remove one atom at a time. We can take advantage of this to efficiently update the Hessian H^t and its inverse with rank one updates. The computational cost for updating the Hessian is $\mathcal{O}\big(\min(ks^2, ps + ks)\big)$ when an atom is added and $\mathcal{O}(k)$ when removing an atom. The additional cost to update $(H^t)^{-1}$ is then just $\mathcal{O}(k^2)$ in both cases. See the appendix for more details on the rank one updates.

4.6 Experiments

In this section, we report experiments that illustrate the computational efficiency of the proposed algorithm. We consider linear regression problems of the form of (4.1) with $f(w) = 1/2\|Xw - y\|^2$, where X is a design matrix and $\gamma_{\mathcal{A}}$ the LGL or the sparse-PCA norms described in Section 4.2. We also considered the constrained version for LGL, $\min_x f(x)$ s.t. $\Omega_{\text{LGL}}(w) \leq \rho$, in section 4.6.2.

Section 4.6.1 compares the performance of our proposed algorithm with state-of-the-art algorithms for the group Lasso. Section 4.6.2 presents comparisons with the variants of Frank-Wolfe and with COGEnT on problem involving the latent group Lasso. Section 4.6.3 provides a comparison with a version of FCFW relying on interior-point solver on larger scale problems. Sections 4.6.3 and 4.6.4 provide comparisons with randomized block proximal coordinate descent algorithms. Most experiments are on simulated data to control characteristics of the experiments, except in section 4.6.3.

4.6.1 Classical group Lasso

We consider an example with group Lasso regularization with groups of size 10,

$$\mathcal{B} = \big\{\{1,\dots,10\},\{11,\dots,20\},\dots\big\}.$$

We choose the support of the parameter $w_0 \in \mathbb{R}^{1000}$ of the model to be $\{1,\dots,50\}$ and all non zero coefficients are set to 2. We generate $n = 200$ examples $(y_i)_{i=1,..,n}$ from $y = x^\top w + \varepsilon$. Block Coordinate Descent (BCD) algorithms are the standard method for this problems but they suffer slow convergence when the design matrix is highly correlated. In this experiment we choose a highly correlated design matrix (with singular values in $\{1, 0.9^2, .., 0.9^{2(p-2)}, 0.9^{2(p-1)}\}$) to highlight the advantages of our algorithm for the harder instances. We compared our algorithm to our own implementation of BCD and an enhanced BCD from Qin et al. (2013) (hyb-BCD). Figure 4.3 shows that we outperform both methods.

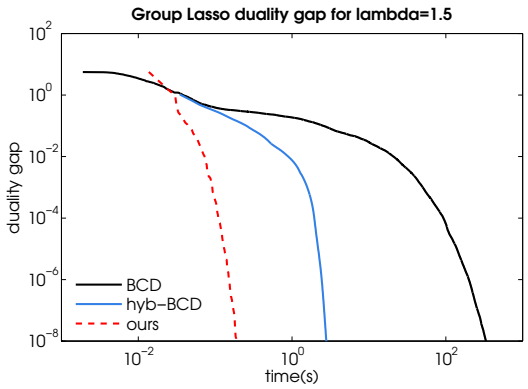


Figure 4.3: Experiment for classical group Lasso. *log-log* plot of progress of the duality gap.

4.6.2 k-chain latent group Lasso

We consider a toy example involving latent group Lasso regularization where the groups are chains of continuous indices of length $k = 8$, that is where the collection of group is $\mathcal{B} = \big\{\{1,\dots,k\},\{2,\dots,k+1\},\dots,\{p-k+1,\dots,p\}\big\}$. We choose the support of the parameter w_0 of the model to be $\{1,\dots,10\}$. Hence, three overlapping chains are needed to retrieve the support of w_0. We generate $n = 300$ examples $(y_i)_{i=1,..,n}$ from $y = x^\top w + \varepsilon$ where x is a standard Gaussian vector and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$. The noise level is chosen to be $\sigma = 0.1$. In upper Figure 4.6.2 we show a time comparison of our algorithm on the regularized problem. We implemented Algorithm 1 and three Frank-Wolfe versions: simple FW, FW with line search (FW-ls) and pairwise FW (FW-pw). We compare also with a regularized version of the *forward-backward greedy* algorithm from Rao et al. (2015)(CoGEnT). In the bottom plot of Figure 4.6.2 we show a comparison on the constrained problem. All codes are in MATLAB and we used Rao et al.'s code for the *forward-backward greedy* algorithm.

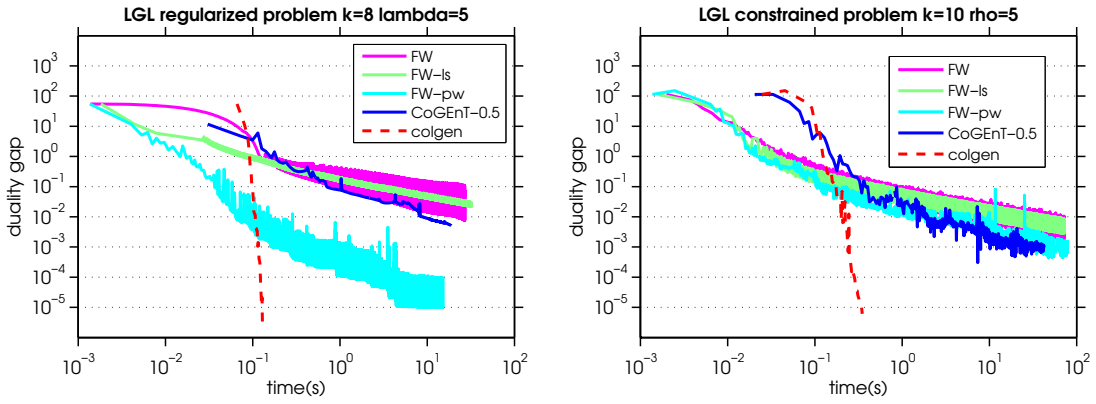


Figure 4.4: Experiments for k-chain group Lasso, where X is a generated random design matrix.*log-log* plot of progress of the duality gap during computation time. CoGEnT truncation parameter is set to $\eta = 0.5$.

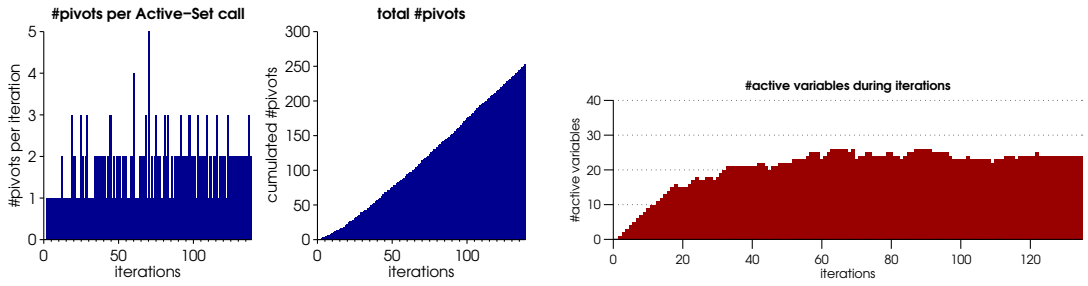


Figure 4.5: Experiment for the k-chain group Lasso. Left: number of pivots, i.e., drop/full step in active set. Middle the number of pivots per active set call and right plot shows the total number of pivots during iterations. Right: evolution of the number of active atoms in our algorithm.

Figure 4.6.2 illustrates complexity and memory usage of our algorithm for the same experiment. Top plots show that each call to the active-set algorithm has low cost. Indeed less than two pivots in average, i.e. drop or full steps, are needed to converge. This is clearly due to the use of warm starts. Bottom plot shows the number of active atoms during iterations.

4.6.3 Hierarchical sparsity

In high-dimensional linear models that involve interaction terms, statisticians usually favor variable selection obeying certain logical hierarchical constraints. In this section we consider a quadratic model (linear + interaction terms) of the form

$$y = \sum_{i=1}^{p} \beta_i x_i + \sum_{i \neq j} \beta_{ij} x_i x_j.$$

Strong and weak hierarchical sparsity are usually distinguished (see Bien et al., 2013, and reference therein). The Weak Hierarchical (WH) sparsity constraints are that if an interaction is selected, then at least one of its associated main effects is selected, i.e., $\beta_{ij} \neq 0 \Rightarrow \beta_i \neq 0$ or $\beta_j \neq 0$. We use the latent overlapping group Lasso formulation proposed in Yan and Bien (2015) to formulate our problem. The corresponding collection of groups $\mathcal{B}$ thus contains the singletons $\{i\}$ and contains for all pairs $\{i, j\}$ the sets $\{i, \{i, j\}\}$ and $\{j, \{i, j\}\}$ (coupling respectively the selection of β_{ij} with that of β_i or that of β_j). We focussed on WH sparsity which is more challenging here because of the group overlaps, but the approach applies also to the counterpart for strong hierarchical constraints.

Simulated data We consider a quadratic problem with $p = 50$ main features, which entails that we have $p \times (p-1)/2 = 1225$ potential interaction terms and simulate $n = 1000$ samples. We choose the parameter β to have 10% of the interaction terms β_{ij} equal to 1 and the rest equal to zero. In order to respect the WH structure, the minimal number of necessary unary terms β_i possible given the WH constrains are included in the model with $\beta_i = 0.5$. We compare our algorithm with FCFW combined with an interior point solver (FCFW-ip) instead of the active-set subroutine, and with a degraded version of our algorithm not using warm starts. Figure 4.6 shows that FCFW-ip becomes slower than our algorithm only beyond 200 seconds. A plausible explanation is that at the begining the subproblems being solved are small and time is dominated by the search of the new direction; when the size of the problem grows, the active-set with warm start is faster, meaning that the active-set exploits the structure of positivity constraints better than IP, which has to invert bigger matrices. Full corrections of FCFW-ip call the `quadprog` function of MATLAB, which is an optimized C++ routine, whereas our implementation is done in MATLAB. An optimized C implementation of our active-set algorithm, in particular leveraging the rank one updates on the inverse Hessian described in sections 4.5 should provide an additional significant speedup.

California housing data set We apply the previous hierarchical mode to the California housing data (Pace and Barry, 1997). The data contains 8 variables, so with interaction terms the intial model contains 36 variables. To make the selection problem more challenging, following She and Jiang (2014), we add 20 main nuisance variables, generated as standard Gaussian random variables corresponding to 370 additional noisy interaction terms. We compare our algorithm to the greedy Forward-Backward algorithm with a truncation parameter $\eta = 0.5$ and with Block Coordinate Descent (BCD). Table 4.1 shows running time for different levels of regularization λ. $\lambda = 10^{-3}$ is the value selected by 10-fold cross validation on the validation risk. Figure 4.7 shows the running time for the different algorithms.

4.6.4 Sparse PCA

We compare our method to the block proximal gradient descent (BCD) described in Richard et al. (2014). We generate a sparse covariance matrix $\Sigma^\star$ of size 150×150 obtained as the sum of five overlapping rank one blocks $\mathbf{1}\mathbf{1}^\top$ of size $k \times k$ with $k = 10$. We generate a noisy covariance with a noise level $\sigma = 0.3$. We consider an ℓ_2 loss and a regularization by the gauge

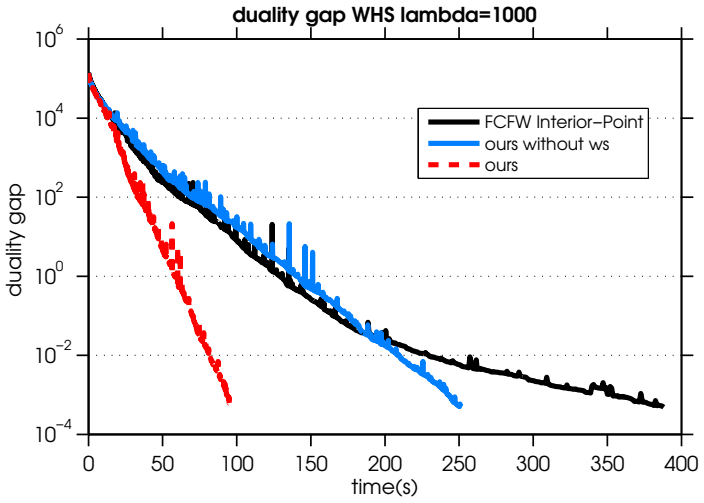


Figure 4.6: Experiments on simulated data for WH sparsity. *log*-plot of progress of the duality gap as a function time in seconds.

Table 4.1: Computation time in seconds needed to reach a duality gap of 10^{-3} on California housing data set. Time is not reported when larger than 10^3 seconds.

λ	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
BCD	-	-	585	73	5
CoGEnT	-	-	1300	14	0.2
ours	27	1.4	0.4	0.06	0.02

$\gamma_{\mathcal{A}_{k,\geq}}$ described in Section 4.2 with $k = 10$. The regularization parameter is λ. Figure 4.9 shows a time comparison with BCD.

4.7 Discussion

In this chapter, we have shown that to minimize a quadratic function with an *atomic norm* regularization or constraint, the fully corrective Frank-Wolfe algorithm, which in the regularized case corresponds exactly to a very simple column generating algorithm that is not well known, is particularly efficient given that sparsity make the computation of the reduced Hessian relatively cheap. In particular, the corrective step is solved very efficiently with a simple active-set methods for quadratic programming. The proposed algorithm takes advantage of warm-starts, and empirically outperforms other Frank-Wolfe schemes, block-coordinate descent (when applicable) and the algorithm of Rao et al. (2015). Its performance could be

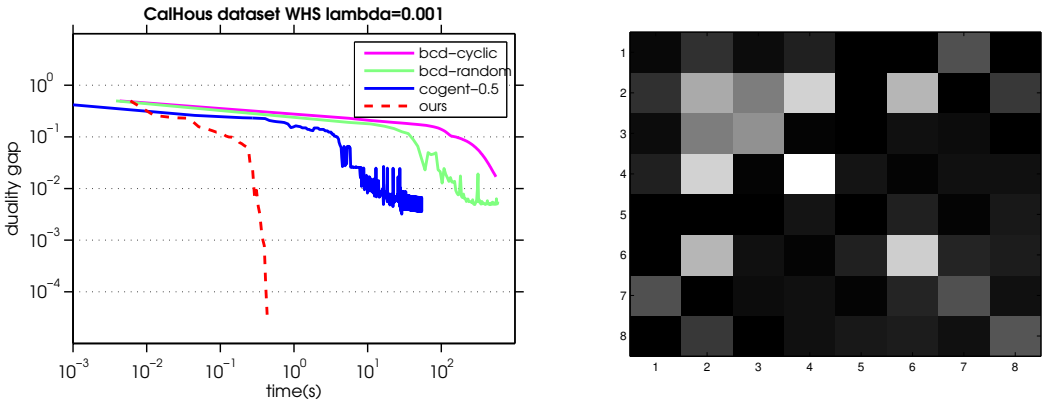


Figure 4.7: Experiments on California House data set. Left: *log-log* plot of progress of the duality gap during computation time. Right: the obtained interaction matrix between the 8 variables.

enhanced by low-rank updates of the inverse Hessian. In future work we intend to generalize the algorithm to smooth loss functions using sequential quadratic programming.

In this chapter we have focused in the general atomic norm case, with possibly uncountable number of atoms. It is worth noting that for specific instances of atomic norms as Lasso, Elastic Net or group Lasso other efficient algorithms exist. The idea of using active set algorithm to speed up an inner loop solver is well known from the prior literature (Roth and Fischer, 2008; Kowalski et al., 2011). In Bach et al. (2012a), authors show that active set method provide significant speedup for problems regularized by sparsity-inducing norms. In You et al. (2016) the active set algorithm is combined with LARS where several variables are added at each round of active set. In our case, since there is a continuum of atoms, we can only add one atom at a time because adding k atoms most violating the optimality conditions would result in adding k very similar atoms. Indeed, since there is a continuum of atoms, the second most violating atom is infinitesimally close to the previous one. In our case, using LARS in the inner loop would not accelerate the algorithm since we add only one variable at a time which is redundant with what LARS does. It would be interesting doing research in this direction in order to be able to add several atoms at a time but this goes beyond the scope of this work.

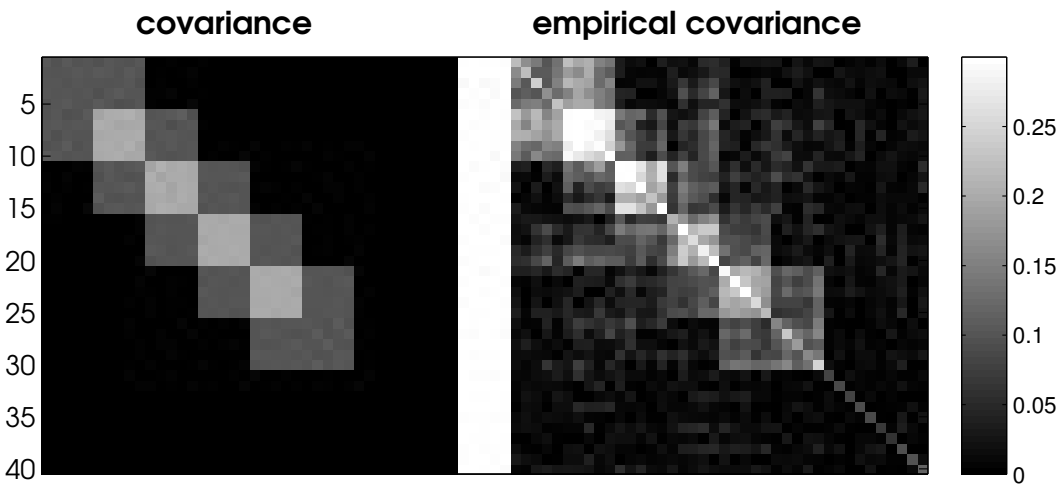


Figure 4.8: Zoom on 40 first variables of true covariance(left) and empirical covariance(right) for a noise level $\sigma = 0.3$.

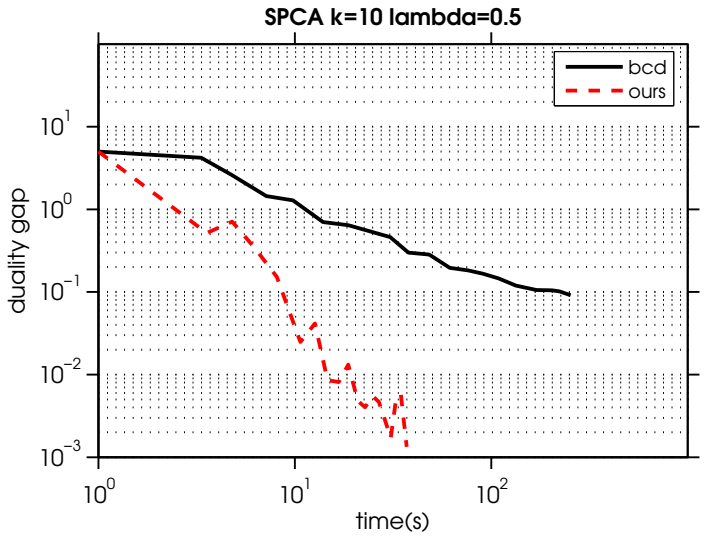


Figure 4.9: Experiments on sparse PCA. *log-log* plot of progress of the duality gap.

Chapter 5

Learning structure in probabilistic graphical models

Probabilistic graphical models are a powerful tool for modeling multivariate distributions in statistical learning. They capture interaction between variables, such as conditional independence and are useful to provide meaningful information about interactions. Graphical models have been applied to a large number of fields, including bioinformatics, social science, image processing, among others. However, structure learning for graphical models remains an open challenge, since the space of all possible structures is exponentially large. In this chapter we present an overview of graphical models, focusing in particular in undirected Gaussian graphical models. We briefly review the framework of exponential families. We also present the different methods of structure learning that exist in the literature for undirected and directed graphs. Finally, as it will be the object of the next chapter, we review in more detail methods for learning the structure of Gaussian graphical models.

5.1 Graphical models

A graphical model represents a family of distributions (Lauritzen, 1996). Each model is associated to a graph $G = (V, E)$, where the vertex set V indexes the variables and the edge set E reflects stochastic dependencies among random variables X_v, $v \in V$. The random variables X_v can either be discrete or continuous. The edge set E encodes allowed conditional dependencies among the variables

Definition 9. (conditional independence) *Two sets of variables X_A and X_B are said to be conditionally independent given a set of variables $X_C := \{X_u \mid u \in C\}$ if and only if*

$$p(x_A, x_B|x_C) = p(x_A|x_C)p(x_B|x_C).$$

It is denoted by the expression $X_A \perp\!\!\!\perp X_B|X_C$.

We distinguish two classes of graphical models: undirected graphical models, also known as Markov random fields, and directed graphical models such as a Directed Acyclic Graphs(DAG).

5.1.1 Undirected graphical models

In an undirected graph the conditional independence of sets of variables X_A and X_B given the set of variables X_S is encoded by the fact that every path between any node in A and any node in B contains a node from S. Such a path is called a *blocked* path. Conditional independence structure determined by a graph is called *global Markov property* and defined below.

Definition 10. (global Markov property) *We say that probability distribution p satisfies the* global Markov property *with respect to graph G if and only if for all A, B, S ⊂ V disjoint subsets:*

$$(\textit{all paths between } A \textit{ and } B \textit{ are blocked by } S) \quad \Rightarrow \quad (X_A \perp\!\!\!\perp X_B | X_S).$$

A fundamental result in graphical models is the Hammersley–Clifford theorem that states equivalence between conditional independence and factorization of the probability distribution. The proof can be found in Koller and Friedman (2009).

Theorem 2. (Hammersley–Clifford). *Let $X = X_1, \ldots, X_n$ be a multivariate random variable and p a positive distribution, i.e. $p(x) > 0$ for all x. The distribution p satisfies the conditional independence structure captured by $G = (V, E)$ if and only if it factorizes over the maximal cliques of G as*

$$p(x) = \prod_{C \in cliques(G)} \phi_C(x_C), \tag{5.1}$$

where each ϕ_C is called potential function*, and x_C denotes the subvector of x indexed by C.*

Thus, the graphical model G is a family of distributions satisfying the factorization (5.1).

5.1.2 Directed graphical models

In a directed graphical models, also known as Bayesian network, edges are oriented. If there is an edge $E_{i,j}$ from v_i to v_j , then v_i is a parent of node v_j and v_j is a child of node v_i. π_i denotes the set of parents of vertex v_i. If there is no cycle, we call it a Directed Acyclic Graph (DAG). The Markov property for DAG is equivalent to the existence of a factorization of the joint distributions into child-given-parents conditional distributions, stated in the next theorem.

Theorem 3 (Factorisation DAG)**.** *The probability distribution p satisfies the conditional independence structure captured by $G = (V, E)$ if and only if*

$$\forall x, \quad p(x) = \prod_{i=1}^{n} p(x_i | x_{\pi_i}). \tag{5.2}$$

5.2 Exponential families

An exponential family is a set of probability distributions of a certain form that unifies many of the most important and widely-used statistical models such as the Normal, Binomial, Poisson, and Gamma into one framework. A probability distribution on a set $\mathcal{X}$ (typically a finite set of values or $\mathbb{R}^p$) is part of an exponential family if it can be written of the form

$$p(x;\theta)\,d\mu(x) = h(x)\exp\left\{b(\theta)^T\phi(x) - \tilde{A}(\theta)\right\}d\mu(x),$$

where:

- $h(x)$ is the ancillary statistic, a statistic whose sampling distribution does not depend on the parameters of the model.
- $h(x)d\mu(x)$ is the reference measure or base measure (in many cases it is equal to 1).
- $\phi(x) : \mathcal{X} \mapsto \mathbb{R}^p$ the *sufficient statistic*, also called feature vector, where p is some fixed integer. Sufficient statistics summarize the relevant information in a sample about the desired parameter.
- θ is the parameter of the model.
- $\eta = b(\theta)$ is the *canonical parameter* of the model and weights the sufficient statistic. When b is the identity, the family is called a *flat exponential family*, and a *curved exponential family* otherwise.
- $\tilde{A}(\theta) = A(\eta)$ is the *log-partition function*

  $$A(\eta) = \log\int_{\mathcal{X}} h(x)\exp\left\{\eta^T\phi(x)\right\}d\mu(x),$$

  and it ensures that we obtain a probability distribution, i.e. $1 = \int_{\mathcal{X}} p(x|\eta)d\mu(x)$. The set of admissible parameters is $\{\eta|A(\eta) < \infty\}$ and is called *domain*.

The different terms are illustrated in the next examples.

Multinomial model Let X be a random variable on $\mathcal{X} = \{0,1\}^K$. X follows a multinomial distribution of parameter $\pi \in [0,1]^K$. The probability distribution writes $p(x;\pi) = \prod_{k=1}^K \pi_k^{x_k}$ and the canonical parameterization is given by $p(x;\eta) = h(x)\exp(\eta^T\phi(x) - A(\eta))$ where

$$\eta = (\log\pi_1, \log\pi_2, \cdots, \log\pi_K)^T, \quad \phi(x) = x \quad \text{and} \quad A(\eta) = \log\Big(\sum_{k=1}^K \exp(\eta_k)\Big).$$

Next we use the canonical parameterization of an exponential familiy in the context of Gaussian graphical models to shows the link between the sparsity pattern of the inverse covariance matrix and the structure of the graphical model.

5.2.1 Gaussian graphical models and zeros of the precision matrix

When $X = (X_1, X_2, ..., X_p)$ follows a multivariate Gaussian distribution with mean $\mu \in \mathbb{R}^p$ and covariance $\Sigma \in \mathbb{R}^{p \times p}$, $\Sigma \succeq 0$, the density is defined by

$$p(x; \Sigma, \mu) = \frac{1}{(2\pi)^{p/2}(\det\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

Using the canonical parameterization, the probability distribution writes

$$p(x; \Lambda, \eta) = \exp\left[\eta^\top x - \frac{1}{2}x^\top \Lambda x - A(\eta, \Lambda)\right],$$

where canonical parameters are $\eta = \Sigma^{-1}\mu$ and $\Lambda = \Sigma^{-1}$. The inverse of the covariance matrix is also called *precision matrix* or *concentration matrix*. The log-partition is $A(\eta, \Lambda) = -\frac{1}{2}\eta^\top \Lambda^{-1}\eta + \frac{p}{2}\log 2\pi - \frac{1}{2}\log\det(\Lambda)$. More precisely, the distribution is proportional to $\exp\left(\eta^\top x - \frac{1}{2}x^\top \Lambda x\right)$ and

$$\exp\left[\eta^\top x - \frac{1}{2}x^\top \Lambda x\right] = \prod_{i=1}^{p} \exp(\eta_i x_i) \prod_{i=1}^{p}\prod_{j=1}^{p} \exp\left[-\frac{1}{2}x_i \Lambda_{ij} x_j\right],$$

which proves that non zero coefficients in Λ correspond to edges in the underlying graphical model.

5.3 Learning structure of graphical models

The problem of structure selection states as follows: given a collection $\{x^{(1)}, ..., x^{(n)}\}$ of n i.i.d. samples from a graphical model, we want to retrieve the unknown underlying graph structure. This problem is NP-hard and has been solved only under special assumptions on the graphical model structure. Concretely, we find two main topics of interest graphical models with discrete variables and Gaussian graphical models for continuous variables. We review different methods for undirected graphs and directed graphs. The special case of undirected Gaussian graphical model, where there is a direct link the nonzeros of the inverse of the covariance matrix and edges of the graph, is detailed in Section 5.4. The content of this section is based on the two reviews Zhou (2011) and Drton and Maathuis (2017).

5.3.1 Undirected graphical models structure learning

We distinguish greedy methods and convex optimization methods.

In greedy algorithms we learn the structure of the graph by sequentially adding nodes and edges to the graph while trying to maximizes a likelihood, information criterion (such as BIC the Bayesian Information Criterion) or a structure criterion (such as MDL the Minimum Description Length). Greedy methods involve local search to perform edge addition/deletion. In some simple structure cases such as trees, exact maximization is possible (Chow and Liu, 1968). Greedy methods are adapted to decomposable graphs such as Gaussian models or

discrete models where maximum likelihood estimator admits closed form solution.

More recently, greedy search has been applied in a framework of neighborhood selection.This avoids the need for iterative computation of MLEs when dealing with nondecomposable graphs. Neighborhood selection has been proposed for learning discrete graphical model (Jalali et al., 2011; Ray et al., 2015). Meinshausen and Bühlmann (2006) propose an approach for Gaussian graphical models that is explained in Section 5.4.

We can also consider convex optimization based algorithms by applying ℓ_1-regularization to the joint distribution. The first works to explore this in undirected graphical models over discrete variables are Lee et al. (2007); Ravikumar et al. (2009) and Dahinden et al. (2007). Most of the work has considered the special case of pairwise undirected graphical models with discrete variables. It has also been applied to Gaussian graphical models (Yuan and Lin, 2006b; Banerjee et al., 2008).

5.3.2 Directed graphical models structure learning

Structure learning problems for directed graphical models is out of the scope of this thesis but for completness purposes we briefly review main methods used for structure learning in directed graphical models. Contrarily to general undirected graphical models where one must cope with the normalization constant, in DAGs log-Likelihood separates into a set of independent problems. Thus it is possible to perform many operations (ie. computing the probability of a vector, computing marginals) exactly or approximately in DAG models in polynomial time whereas it is intractable for general undirected models.

In search and score methods, we use some criterion to assess the quality of a particular structure (such as the BIC or validation set likelihood), and we optimize this criterion by using a local search on the space of DAGs (Lam and Bacchus, 1993; Heckerman et al., 1995). Usually a greedy local search is performed where at each iteration an edge is added/removed. Another class of methods are constraint-based methods that prune the set of possible edges by selecting pairs of variables that satisfy a conditional independence hypothesis test(Geiger et al., 1990; Spirtes and Glymour, 1991). In practice, conditional independencies need to be tested based on data. Standard tests are available for multivariate Gaussian and multinomial data. The disadvantages of the constraint-based methods and the search and score methods have led to the development of hybrid methods. In hybrid methods, constraint-based reasoning is used to prune the set of edges to consider within a search and score method. This can lead to an enormous reduction in the number of possible graphs to search over.

Another line of work involves ℓ_1-regularization. In DAGs, a not necessarily topological ordering of the nodes can always be defined according to edge distribution (Kahn, 1962). Identifying this ordering is known to be a challenging problem (Cook, 1985). Because the graph must be acyclic, we can not simply regress each node on all other nodes. Subsequently, we need to consider searching through the space of topological orderings, or directly searching

through the space of directed acyclic graphs. Previous work on structure learning in DAG models with ℓ_1-regularization has considered the case of a known ordering (Huang et al., 2006; Li and Yang, 2005; Levina et al., 2008) and more recently Champion et al. (2018) propose a method without assuming known ordering.

5.4 Inverse covariance estimation in Gaussian graphical models

We now turn to the case of fitting Gaussian graphical models. In undirected Gaussian graphical models, the problem of structure learning reduces to the estimation of the precision matrix and is also known as inverse covariance selection. We describe the two main classes of methods for graph selection that have been proposed in the literature: neighborhood selection and penalized likelihood. In order to obtain a parsimonious model, i.e. control the number of edges, it makes sense to impose an ℓ_1 penalty for the estimation of Λ. We also mention other types of regularization

Neighborhood selection

Meinshausen and Bühlmann (2006) propose a simple approach based on neighborhood selection. The conditional distribution of X_s given the other variables, denoted X_{-s}, is also Gaussian and is expressed as a combination of the other variables, that is

$$x_s = \beta^{s\top} x_{-s} + \varepsilon, \tag{5.3}$$

where $\beta^s \in \mathbb{R}^{p-1}$ is the parameter vector and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ is the noise. Thus, the support of the regression vector β^s is equal to the neighbours of variable s. Neighborhood selection estimates individually the neighborhood of each given variable $s = 1, ..., p$ as a Lasso regression problem

$$\hat{\beta}^s \in \arg\min_{\beta^s \in \mathbb{R}^{p-1}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (x_s^{(i)} - \beta^{s\top} x_{-s}^{(i)})^2 + \lambda \|\beta^s\|_1 \right\}. \tag{5.4}$$

Finally the neighborhood estimates are combined to obtain the final set of edges of the graph. Neighborhood selection is computationally attractive since many efficient Lasso implementations exist.

Graphical Lasso

Yuan and Lin (2006b) and Banerjee et al. (2008) proposed the *graphical lasso* (glasso) estimator, that is the minimization of the ℓ_1-penalized log-likelihood,

$$\hat{\Lambda} \in \arg\min_{\Lambda \succeq 0} \left\{ \log\det \Lambda - \mathrm{tr}\hat{\Sigma}\Lambda + \lambda \|\Lambda\|_1 \right\} \tag{5.5}$$

where $\hat{\Sigma}$ is the empirical covariance. Contrarily to neighborhood selection methods, the penalized likelihood formulation allows model selection and parameter estimation simultaneously. Yuan and Lin (2006b) used the interior-point algorithm to compute the estimator. Banerjee et al. (2008) propose an efficient semi-definite programming algorithm based on Nesterov's method for interior point optimization and also block coordinate descent method. A faster implementation using a pathwise-coordinate-descent approach to solve the modified lasso problems at each stage, and for a decreasing series of values of the regularization parameter λ was proposed by Friedman et al. (2008). Rothman et al. (2008) introduce the SPICE (Sparse Permutation Invariant Covariance Estimator) estimator where only the off-diagonal elements are penalized in (5.5) and propose an algorithm.

Score Matching loss

In undirected graphical models, the probability density function is known only up to a multiplicative normalization constant which is often intractable. Hyvärinen (2005) propose a score matching loss for estimating non-normalized statistical models. For Gaussian models, score matching loss takes a very simple quadratic form

$$f(\Lambda) = \frac{1}{2}\mathrm{tr}(\Lambda^2\hat{\Sigma}) - \mathrm{tr}(\Lambda).$$

Other penalties

A conditional independence graph is sometimes expected to have particular structure. In the context of graphs with 'hub' nodes with many neighbors, Tan et al. (2014) present a convex formulation that involves a row-column overlap norm penalty. Defazio and Caetano (2012) use a convex penalty adapted for a scale-free network in which the degree of connectivity of the nodes follows a power law distribution. Tao et al. (2017) impose an overlapping group structure on the concentration matrix.

Another useful problem is finding the structure of graphical models with unobserved variables. Chandrasekaran et al. (2010) propose a convex formulation to find the number of latent components and learn the structure of on the entire collection of variables. In the next chapter we present our contributions on learning graphical models with unobserved variables.

Chapter 6

Learning the effect of latent variables in Gaussian Graphical models with unobserved variables

This chapter is based on our paper Vinyes and Obozinski (2018). The edge structure of the graph defining an undirected graphical model describes precisely the structure of dependence between the variables in the graph. In many applications, the dependence structure is unknown and it is desirable to learn it from data, often because it is a preliminary step to be able to ascertain causal effects. This problem, known as structure learning, is a hard problem in general, but for Gaussian graphical models it is slightly easier because the structure of the graph is given by the sparsity pattern of the precision matrix of the joint distribution, and because independence coincides with decorrelation.

A major difficulty too often ignored in structure learning is the fact that if some variables are not observed, the marginal dependence graph over the observed variables will possibly be significantly more complex and no longer reflect the direct dependences that are potentially associated with causal effects. This is the problem of confounding variables. In this work, we consider a family of latent variable Gaussian graphical models (LVGGM) in which the graph of the joint distribution between observed and unobserved variables is sparse, and the unobserved variables are conditionally independent given the others. Prior work (Chandrasekaran et al., 2010) was able to recover the connectivity between observed variables, but could only identify the subspace spanned by unobserved variables, whereas we propose a convex optimization formulation based on structured matrix sparsity to estimate the complete connectivity of the original complete graph including unobserved variables, given the knowledge of the number of missing variables, and a priori knowledge of their level of connectivity. Our formulation is supported by a theoretical result of identifiability of the latent dependence structure for sparse graphs in the infinite data limit. We propose an algorithm leveraging recent active set methods, which performs well in the experiments we ran on synthetic data.

6.1 Introduction

Graphical models provide a sound theoretical framework to model a joint probability distribution with complex interdependences between a potentially large number of random variables, with applications in several fields including genomics and finance among others.

In the Gaussian Graphical Models (GGM) literature, a central problem is to estimate the inverse covariance matrix, also known as the *precision* or *concentration matrix*. The sparsity pattern of the concentration matrix in Gaussian models corresponds to the structure of the graph; more precisely, the nonzeros of the concentration matrix correspond to the edges of the underlying undirected graphical model, which encode pairs of variables that are conditionally dependent given all the others. Identifying the structure of the graph is important since the number of parameters of the model grows linearly with the number of edges in the graph.

The main formulation for edge selection in the GGM setting is based on ℓ_1-regularized maximum-likelihood (Yuan and Lin, 2007; d'Aspremont et al., 2008b; Friedman et al., 2008; Banerjee et al., 2008), for which several algorithms have been proposed. The ℓ_1 regularization provides convex formulation which induces the selection of some edges while implicitly removing others in the graph.

A serious practical difficulty is that applications in which all variables potentially relevant for the problem considered have been identified and measured are extremely rare. This entails the possible presence of *confounding variables*. More precisely, some of the relevant variables may be latent and induce correlations between observed variables that can be misleading and can only be explained correctly if the presence of the latent variables that produce confounding effects is explicitly modeled. More precisely, when latent variables are missing, the marginalized precision matrix may not be sparse even if the full precision matrix is sparse. Imposing sparsity on the complete model results in a marginal precision matrix of the Latent Variable Gaussian Graphical Model (LVGGM) that has a sparse plus low-rank structure. Chandrasekaran et al. (2010) consider a regularized maximum likelihood approach, using the ℓ_1-norm to recover the sparse component and the trace norm to recover the low-rank component and show that they consistently estimate the sparsity pattern of the sparse component and the number of latent variables. Their method identifies the low-rank structure corresponding to the effect of latent variables but, in general, it does not allow us to identify the covariance structure of each latent variable individually, or which observed variables are directly dependent on which unobserved ones.

In this work, we propose to impose more structure on the low rank matrix using a variant of the norms introduced in Richard et al. (2014) as a regularizer. This leads to formulations which yields estimates of the structure of the complete graphical model, and, in particular, make it possible to identify which observed variables are affected by which latent variables.

The paper is structured as follows: In Section 6.2 we review the relevant prior literature. In Section 6.3, we formulate the LVGGM estimation problem as a regularized convex problem that imposes a sparsity structure on the latent variables. In Section 6.5, we propose a convex

formulation with a quadratic loss function, and an algorithm to solve this problem efficiently. In Section 6.6, we show that different parts of the complete graph are identifiable by our convex formulation, under appropriate conditions. We finally present experimental results in Section 6.8.

6.2 Related Work

To construct an interpretable graph in high-dimensional regimes, many authors have proposed applying an ℓ_1 penalty to the parameter associated with each edge, in order to encourage sparsity. For instance such an approach is taken by Yuan and Lin (2007) and Banerjee et al. (2008) in the context of Gaussian graphical models. Later, Krishnamurthy et al. (2011) propose an algorithm to compute a full regularization path of solutions to this problem. The first works to explore ℓ_1 regularization in undirected graphical models over discrete variables are Lee et al. (2007); Ravikumar et al. (2009) and Dahinden et al. (2007). In another line of work, authors have considered ℓ_1-regularization for learning structure in directed acyclic graphs given an ordering of the variables (Huang et al., 2006; Li and Yang, 2005; Levina et al., 2008) and Schmidt et al. (2007); Champion et al. (2018) propose methods without assuming known ordering.

A conditional independence graph is sometimes expected to have particular structure. In the context of graphs with hub nodes, that is nodes with many neighbors, Tan et al. (2014) present a convex formulation that involves a row-column overlap norm penalty. Defazio and Caetano (2012) use a convex penalty adapted for a scale-free network in which the degree of connectivity of the nodes follows a power law distribution. Tao et al. (2017) impose an overlapping group structure on the concentration matrix.

Another useful problem, that is the focus of this paper, is finding the structure of Gaussian graphical models with unobserved variables. Chandrasekaran et al. (2010) introduced a convex formulation to find the number of latent components and learn the structure of on the entire collection of variables. Meng et al. (2014) also studied regularized maximum likelihood estimation and derive Frobenius norm error bounds in the highdimensional setting based on the restricted strong convexity. In order to speed up the estimation of the sparse plus low-rank components, Xu et al. (2017) propose a sparsity constrained maximum likelihood estimator based on matrix factorization, and an efficient alternating proximal gradient descent algorithm with hard thresholding to solve it. Hosseini and Lee (2016) present a bi-convex formulation to jointly learn both a network among observed variables and densely connected and overlapping groups of variables, revealing the existence of potential latent variables. These methods identify the low-rank structure corresponding to the effect of latent variables but it does not allow us to identify the structure of the full model. In this work, we propose to impose more structure on the low rank matrix in order to obtain a decomposition that gives the structure of the complete graphical model.

Notations

$[\![p]\!]$ denotes the set $\{1, ..., p\}$ and $\mathcal{G}_k^p$ denotes the set of subsets of k elements in $[\![p]\!]$. $|I|$ denotes the cardinality of a set I. If $v \in \mathbb{R}^p$ is a vector, Supp(v) denotes its support. If $M \in \mathbb{R}^{p\times p}$ is a matrix, $I \subset [\![n]\!]$, $M_{II} \in \mathbb{R}^{|I|\times|I|}$ is the submatrix obtained by selecting the rows and columns indexed by I in M. For a symmetric matrix M, $\lambda^+_{\max}(M)$ is the largest positive eigenvalue and zero if they are all nonpositive. If S is a set, $|S|$ denotes its cardinality.

6.3 Gaussian Graphical Models with Latent Variables

We consider a multivariate Gaussian variable $(X_O, X_H) \in \mathbb{R}^{p+h}$ where O and H are respectively the set of indices of observed variables, with $p = |O|$, and of latent variables, with $h = |H|$. We denote $\Sigma \in \mathbb{R}^{(p+h)\times(p+h)}$ the complete covariance matrix and $K = \Sigma^{-1}$ the complete *concentration matrix* or *precision matrix*. Let $\hat{\Sigma} \in \mathbb{R}^{(p+h)\times(p+h)}$ denote the empirical covariance matrix, based on a sample of size n. We only have access to the empirical marginal covariance matrix $\hat{\Sigma}_{OO}$. It is well known that the marginal concentration matrix on the observed variables can be computed from the full concentration matrix as

$$\Sigma^{-1}_{OO} = K_{OO} - K_{OH}K^{-1}_{HH}K_{HO}. \tag{6.1}$$

We assume that the original graphical model is sparse and that there is a small number of latent variables. This implies that K_{OO} is a sparse matrix and that $K_{OH}K^{-1}_{HH}K_{HO}$ is a low-rank matrix, of rank at most h. Note that Σ^{-1}_{OO} is typically not be sparse due to the addition of the term $K_{OH}K^{-1}_{HH}K_{HO}$. Figure 6.1 shows an example of an LVGGM structure where variables {1,2,3} are hidden variables and Figure 6.2(a) shows the structure of its corresponding complete concentration matrix K. Figure 6.2(b) shows an approximation of Σ^{-1}_{OO} as "sparse + low rank" matrix.

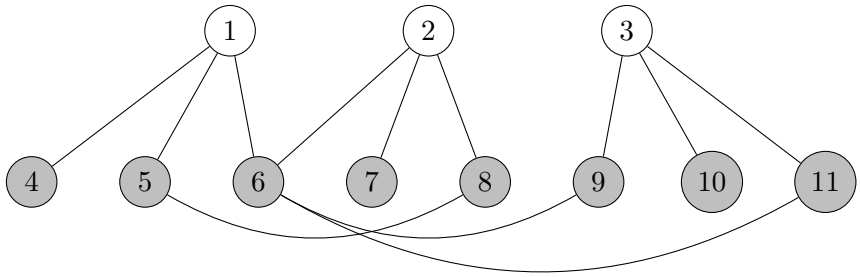
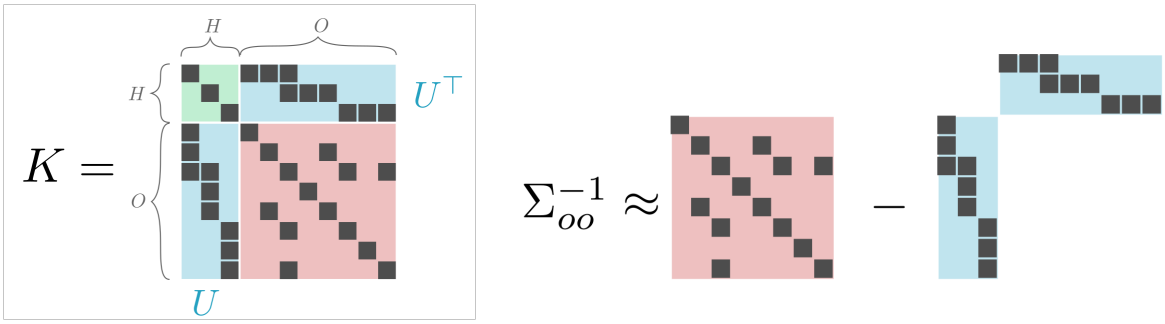


Figure 6.1: Example of an LVGGM structure where the variables $\{1, 2, 3\}$ are hidden variables

Chandrasekaran et al. (2010) show that under appropriate conditions, namely if K_{OO} is sufficiently sparse and $K_{OH}K^{-1}_{HH}K_{HO}$ is low rank and cannot be approximated by a sparse matrix, these two terms are identifiable and can be estimated, via an estimator of Σ^{-1}_{OO} of the form $S-L$, where S is sparse, L is low rank, and $S - L$, S and L are p.s.d. matrices in order to match the structure of (6.1), and guarantee that the estimate of the original matrix K is p.s.d. Moreover the authors show that S and L can be estimated via the following convex optimization problem:



(a) (b)

Figure 6.2: (a) Structure of complete concentration matrix K of graph in Figure 6.1. (b) approximation of Σ_{OO}^{-1} as "sparse + low rank"

$$\min_{S,L} f(S-L) + \lambda\left(\gamma\|S\|_1 + \mathrm{tr}(L)\right) \tag{6.2}$$
$$\text{s.t.} \quad S - L \succeq 0, \quad L \succeq 0,$$

where f is a convex loss function, and λ, γ are regularization parameters. The positivity constraint on S has been dropped since it is implied by $S - L \succeq 0$ and $L \succeq 0$. Typically, in GGM selection, f is the negative log-likelihood.

$$f_{ML}(M) := -\log\det(M) + \mathrm{tr}(M\hat{\Sigma}). \tag{6.3}$$

Two other natural losses, that have the advantage of being quadratic, are the second order Taylor expansion around the identity matrix of the log-likelihood f_T and the score matching loss f_{SM}, introduced by Hyvärinen (2005) and used for GGM estimation in Lin et al. (2016),

$$f_T(M) := \frac{1}{2}\|\hat{\Sigma}^{1/2} M \hat{\Sigma}^{1/2} - I\|_2^2 \tag{6.4}$$
$$f_{SM}(M) := \frac{1}{2}\mathrm{tr}(M^2\hat{\Sigma}) - \mathrm{tr}(M). \tag{6.5}$$

Chandrasekaran et al. (2010) show that under appropriate technical conditions, the regularized maximum log-likelihood formulation (6.2) provides estimates (S_n, L_n) that have respectively the same sparsity pattern and rank as K_{OO} and $K_{OH}K_{HH}^{-1}K_{HO}$. The obtained low rank component L_n retrieves the latent variable subspace.

Note first that, in general, K_{HH} and K_{OH} are not identifiable and cannot be estimated from L_n. Therefore the connectivity between the latent variables and the connectivity between latent and observed variables cannot be recovered. However, under the assumption that the sources are conditionally independent given observed nodes, K_{HH} is diagonal, and, when the

groups of observed variables associated with each latent variables are moreover disjoint, the columns of K_{OH} have disjoint support and are therefore orthogonal. This necessarily implies that they are proportional to the eigenvectors of $K_{OH}K_{HH}^{-1}K_{HO}$ as soon as the coefficients of the diagonal matrix K_{HH} are all distinct, by uniqueness of the SVD. In that case, they are thus identifiable, and it makes sense to estimate the columns of K_{OH} by the eigenvectors of the estimated L.

However, if the columns of K_{OH} are sparse, it would seem relevant to encode this in the model, as this is potentially a stronger prior than orthogonality. Moreover, it might be relevant to allow the groups of observed variables associated with each given latent variable to overlap.

In this work, assuming that the latent variables are independent, we propose a formulation allowing to estimate the columns of K_{HO} up to a constant, based on an assumption on its relative sparsity, that we encode as a prior using a matrix norm introduced by Richard et al. (2014).

6.4 Spsd-rank(k) and a convex surrogate

Richard et al. (2014) proposed matrix norms and gauges[1] that yield estimates for low-rank matrices whose factors are sparse. One variant, which is actually a gauge[2], specifically suited to the estimation of p.s.d. matrices, induces a decomposition into with sparse rank one p.s.d. factors. In this section, we introduce the k-spsd-rank of a p.s.d. matrix relate it to this gauge, which assumes that the sparsity of the factors is known and fixed. We then discuss a generalization for factors of different sparsity levels.

The following definition is a generalization of the rank for p.s.d. matrices,

Definition 11 (k-spsd-rank)**.** *For a p.s.d. matrix $Z \in \mathbb{R}^{p\times p}$ and for $k > 1$ we define its k-spsd-rank as the optimal value of the optimization problem:*

$$\begin{aligned} &\min \|c\|_0 \\ &s.t.\ \ Z = \sum_i c_i u_i u_i^\top,\ \ c_i \in \mathbb{R}^+,\ \ u_i \in \mathbb{R}^p : \|u_i\|_0 \leq k, \|u_i\|_2 = 1. \end{aligned}$$

Note that not all p.s.d. matrices admit such a decomposition, in which case the k-spsd-rank is by convention infinite. This is in particular the case for low-rank non sparse matrices like $11^\top$(see Richard et al. (2014) for a proof). A natural convex relaxation of the k-spsd-rank is based on the concept of *atomic norm* proposed in Chandrasekaran et al. (2012). *Atomic norms* are norms (or gauges) whose unit ball is the convex hull of a reduced set of elements of

[1] We will use the word gauge in the paper to mean *closed gauge*. We remind the reader that a closed gauge is simply a proper closed convex positively homogeneous function, and that a gauge γ which is symmetric ($\gamma(x) = \gamma(-x)$), takes finite values, and such that ($\gamma(x) = 0$) $\Rightarrow$ ($x = 0$) is a norm. Gauges are thus natural generalizations of norms, that share many properties including the triangle inequality and the same Fenchel duality theory. We refer the reader to Friedlander et al. (2014) or Rockafellar (1970) for a more detailed presentation of gauges.

[2] See Chandrasekaran et al. (2012) for a discussion.

the ambient space $\mathcal{A}$ called *atoms*. Here we consider the *atomic gauge* associated with the set $\mathcal{A} = \{uu^\top \mid \|u\|_2 \leq 1,\ \|u\|_0 \leq k\}$. In particular, it follows from basic results on *atomic norms* that we can write this one as follows

Definition 12 (Ω, convex relaxation of k-spsd-rank)**.** *For $Z \in \mathbb{R}^{p\times p}$,*

$$\begin{aligned}&\Omega(Z) := \min \|c\|_1\\ &s.t.\ \ Z = \sum_i c_i u_i u_i^\top,\ \ c_i \in \mathbb{R}^+,\ \ u_i \in \mathbb{R}^p : \|u_i\|_0 \leq k, \|u_i\|_2 = 1.\end{aligned}$$

Note that we can have $\Omega(Z) = +\infty$ even when Z is p.s.d., if Z cannot be decomposed in k-sparse, rank-1 p.s.d. factors, as it is the case for $11^\top$. The polar gauge of Ω is characterized as follows:

Lemma 9. *Let $Y \in \mathbb{R}^{p\times p}$ be a symmetric matrix. The polar gauge to Ω writes*

$$\Omega^\circ(Y) = \max_{I\in\mathcal{G}_k^p} \lambda^+_{max}(Y_{II}). \tag{6.6}$$

Unfortunately, the polar gauge Ω° is a priori NP-hard to compute, since it is the largest sparse eigenvalue associated with a sparse eigenvector with k non zero coefficients:

$$\min_u u^\top XX^\top u \quad \text{s.t.} \quad \|u\|_0 \leq k, \quad \|u\|_2 = 1,$$

which is known to be an NP-hard problem to solve (Moghaddam et al., 2006). However, a recent literature proposed quite a number of algorithms to solve sparse PCA approximately or heuristically, among others via convex relaxations (d'Aspremont et al., 2005, 2008a; Zhang et al., 2012), which can be leveraged to approximately solve the corresponding problems. Yuan and Zhang (2013) propose a Power Method type algorithm.

6.4.1 A variant for factors with different sparsity levels

Ω can be generalized to allow each rank one factor have a different sparsity level. A simple way to do this is to consider a gauge of the form

$$\begin{aligned}&\Omega_w(Z) := \inf \sum_i \sum_{k=1}^p w_k c_i^k\\ &\text{s.t.}\ \ Z = \sum_i \sum_{k=1}^p c_i^k u_i^k u_i^{k\top}, c_i^k \in \mathbb{R}^+, u_i^k \in \mathbb{R}^p : \|u_i^k\|_0 \leq k, \|u_i^k\|_2 = 1,\end{aligned}$$

where $k \mapsto w_k$ is an increasing function that penalizes each sparsity level k by w_k. Via a simple change of variable, we can rewrite Ω_w

$$\begin{aligned}&\Omega_w(Z) := \inf \sum_i \sum_{k=1}^p c_i^k\\ &\text{s.t.}\ \ Z = \sum_i \sum_{k=1}^p c_i^k u_i^k u_i^{k\top}, c_i^k \in \mathbb{R}^+, u_i^k \in \mathbb{R}^p : \|u_i^k\|_0 \leq k, \|u_i^k\|_2 = w_k,\end{aligned}$$

which shows that it is a standard atomic gauge in which the rank one atoms with k^2 non-zero coefficients have weight w_k. If we choose $w_k = 1$ for all k, then it can be shown that only the non-sparse atoms will appear in the expansion and so $\Omega_w(Z) = \mathrm{tr}(Z) + \iota_{\{Z \succeq 0\}}$. If $k \mapsto w_k$ accelerates quickly, the gauge will favor sparser factors, but since some p.s.d. matrices cannot be expressed as positive combinations of very sparse p.s.d. rank-one factors, the behavior of the gauge is not trivial for any weights of the form $w_k = k^m$, $m > 0$, even when m is large. Although a detailed analysis of Ω_w is beyond the scope of this work, we illustrate this generalization in the experiments.

6.5 Convex Formulation and Algorithm

We use Ω to impose structure on the low rank component and consider the following convex optimization problem,

$$\min_{S,L} f(S - L) + \lambda\big(\gamma\|S\|_1 + \Omega(L)\big) \quad \text{s.t.} \quad S - L \succeq 0. \tag{6.7}$$

Note that the nonnegativity constraint on L is no longer necessary since the gauge Ω only provides symmetric p.s.d. matrices, as a sum of p.s.d. rank-one matrices.

In order to rewrite our problem as a simple convex regularized by Ω, we drop[3] the nonegativity constraint on $S - L$ and consider the optimization problem

$$\min_{S,L} f(S - L) + \lambda\big(\gamma\|S\|_1 + \Omega(L)\big). \tag{6.8}$$

We propose the alternating optimization scheme presented in Algorithm 10. First, we update the sparse factor S by optimizing problem (6.8) with L fixed, then we update L by solving problem (6.8) with S fixed.

- to update the sparse factor S we apply a fixed number of soft-thresholding iterations, i.e several steps of iterative shrinkage-thresholding algorithm (ISTA). In the experiments we perform 10 soft-thresholding iterations when updating S
- to update the low rank factor L we apply an efficient algorithm for quadratic losses recently proposed by Vinyes and Obozinski (2017) called Fast Column Generation algorithm (FCG). This algorithm is well adapted to the quadratic losses f_T and f_{SM} introduced in Section 6.3

FCG consists in applying a Fully Corrective Frank Wolfe (Lacoste-Julien and Jaggi, 2015) to a regularized optimization problem. Frank Wolfe (FW) algorithm (Frank and Wolfe, 1956), also known as conditional gradient, is particularly well suited for solving quadratic programming problems with linear constraints. They apply in the context where we can easily

[3]It would be possible to still enforce $S - L \succeq 0$, with approach proposed in this paper using Lagrangian techniques with an increase of computational costs.

solve the Linear Minimization Oracle (LMO), a linear problem on a convex set of constraints $\mathcal{C}$ defined as

$$\mathrm{LMO}_{\mathcal{C}}(y) \coloneqq \arg\min_{z \in \mathcal{C}} \langle y, z \rangle . \tag{6.9}$$

In particular $\mathcal{C}$ can be the convex hull of a set of atoms $\mathcal{A}$. At each iteration FW selects a new atom a^t from $\mathcal{C}$ querying the LMO and computes the new iterate as a convex combination of a^t and the old iterate x^t. The convex update can be done by line search. FCFW, discussed in Lacoste-Julien and Jaggi (2015), is a variant of FW that consists in finding the convex combination of all previously selected atoms $(a^i)_{i<t}$. When using the algorithm proposed in Vinyes and Obozinski (2017) we need to compute the following LMO

$$\mathrm{LMO}_{\Omega}(M) \coloneqq \arg\max_{u} \; u^\top M u \quad \text{s.t.} \quad \|u\|_0 = k, \|u\|_2 = 1. \tag{6.10}$$

at each iteration, and subsequently use a working set algorithm to solve the fully corrective step.

We propose to use the Truncated Power Iteration (TPI) heuristic introduced by Yuan and Zhang (2013) to obtain an approximation to the oracle $\mathrm{LMO}_{\Omega}(M)$.

Algorithm 10 Alternate minimization

1: **Require:** f quadratic, maximum iterations T
2: **Initialization:** $S^0 = 0$, $L^0 = 0$, $t = 0$
3: **for** $t = 1..T$ **do**
4: Compute S^t applying a fixed number of `ISTA` iterations on problem (6.8) with L^{t-1} fixed
5: Compute L^t applying `FCG` on problem (6.8) with S^t fixed
6: **end for**
7: return S^t, L^t

6.6 Identifiability of S^* and of the sparse factors of L^*

For formulation (6.8) to yield good estimators, a necessary condition is that, if M is a marginal precision matrix with decomposition $M = S^* + L^*$ with $L^* = \sum_i s_i u^i u^{i\top}$, $\mathrm{Supp}(u^i) \subset I_i$ and $|I_i| = k$, this decomposition can be recovered from perfect knowledge of M (which corresponds to the case where we have an infinite amount of data with no noise). We therefore consider in this section the decomposition problem of a known precision matrix M. For the estimator obtained from (6.8) to provide reasonable estimates, a necessary condition is that it returns correct estimates in the limit of an infinite amount of data.

We will provide sufficient conditions on S^* and L^* so that if $M = S^* + L^*$ and $(\hat{S}, \hat{L})$ is an optimum of the problem

$$\min \gamma \|S\|_1 + \Omega(L) \quad \text{s.t.} \quad M = S + L, \tag{6.11}$$

then $\hat{S} = S^*$, $\hat{L} = L^*$ and the decompositions of $\hat{L}$ and L^* are the same. Our approach is based on the work of Chandrasekaran et al. (2011) but several of our results and proofs are tighter than the original analysis.

We will make the simplifying assumption that the sets I_i are disjoint, so that part of the analysis decomposes on each of the blocks $I_i \times I_i$ and on the complement of $\bigcup_i I_i \times I_i$.

Assumption 1. *Let $L^* = \sum_i s_i u^i u^{i\top}$, with $\mathrm{Supp}(u^i) = I_i$. We assume that the sets I_i are all disjoint and that $|I_i| = k$.*

In particular, this assumption entails implicitly that if $L^* = \sum_i L_i^*$ with L_i the component supported on block $I_i \times I_i$, then L_i^* is of rank one.

In order to be able to decompose M as $M = S^* + L^*$, we need to make assumptions on S^* and L^*. Indeed, there are a number of scenarios in which the possible decompositions of M into *psd rank-one* matrices and *sparse* parts may not be uniquely defined. For instance if the low-rank matrix is itself sparse, or the sparse part not sufficiently sparse, the decomposition might not be identifiable.
Two quantities are key: let $\overline{\tau}$ be an upper bound such that

$$\overline{\tau} \geq k \max_{i \in [\![r]\!]} \|u^i\|_\infty^2 \quad \text{and} \quad k_0 := \max_i \|S_{i\cdot}^*\|_0, \text{ where } \|S_{i\cdot}^*\|_0 := \big|\{j \mid S_{ij}^* \neq 0\}\big|.$$

On one side, k_0 measures the sparsity of S^*, it is the maximal degree of the graph on the observed variables. S^* will be sufficiently sparse if $k_0 \ll k$. On the other, $\overline{\tau} \geq 1$ measures the flatness (vs spikiness) of L^*: again L^* be sufficiently flat if $\overline{\tau} \ll k$.

The interpretation behind an assumption of the form $k_0 \ll k$ is that, in the precision matrix of the joint distribution over observed and latent variables, all the neighbors of a latent node i form a clique, and in this clique, each node has k neighbors. If $k_0 \ll k$, then the connections explained by this clique cannot be attributed to individual connections between observed nodes, and can only be attributed to the presence of a latent variable.

Second, the interaction strength of each hidden node i with its observed neighbors in the graph should be of a similar order of magnitude. Symmetrically, an assumption of the form $\overline{\tau} \ll k$ just imposes an upper bound on the interaction strength between a hidden node and its observed neighbors. Indeed, if latent node i had very strong interactions with j and j', in the marginalized graph the interaction between j and j' induced by i might be difficult to tell appart from a direct interaction between j and j'.

In the next theorems, we will either assume that $\alpha := k_0\sqrt{\frac{2\overline{\tau}}{k}}$, which combines both quantities, is small, or, that $k_0 \leq \frac{1}{7}\sqrt{k}$ and $\overline{\tau} \leq 2$.

To be able to position our general result w.r.t. to the literature, we first state a counterpart for the decomposition into a sparse and a (non necessarily) sparse rank-one p.s.d. matrix, which is very close but improves Corollary 3 of Chandrasekaran et al. (2011).

Theorem 4 (sparse + one rank-one block)**.** *Let $M = S^* + L^*$.*

Consider the optimization problem

$$\min \gamma \|S\|_1 + \mathrm{tr}(L) \quad s.t. \quad M = S + L, \quad L \succeq 0. \tag{6.12}$$

Under the assumption that L^* is p.s.d., rank one and symmetric, if, for the pair (S^*, L^*) the quantities k_0, p and $\overline{\tau}$ are such that $\alpha := k_0\sqrt{\frac{2\overline{\tau}}{p}}$ satisfies $\alpha + \frac{\alpha^2}{2k_0} < \frac{1}{3}$, where p is the ambient dimension, there exist values of γ, such that

$$\frac{\bar{\tau}}{p}\frac{1}{1-3\alpha} \leq \gamma < \frac{1}{k_0}\frac{1-k_0\overline{\tau}/p}{1+\alpha}, \tag{6.13}$$

(i.e. the interval is non empty), and, for any such value of γ, the pair (S^*, L^*) is the unique optimum of problem (6.11)*.*

The result we obtained here provides an improvement over the main result in Chandrasekaran et al. (2011) as stated in Corollary 3. Indeed, in our setting (a single rank one component), the quantities appearing in that result can be computed: $\deg_{\max}(S^*) = k_0$ and $\mathrm{inc}(L^*) = \sqrt{\frac{\overline{\tau}}{k}}$. Thus Corollary 3 of Chandrasekaran et al. (2011) requires $\alpha < \frac{\sqrt{2}}{12}$ when $\alpha < \frac{2}{7}$ is sufficient in our case, and even smaller values of α are allowed for sufficiently large k_0; also, the interval allowed for γ in Chandrasekaran et al. (2011) is, with our notations, $\left(2\sqrt{\frac{\overline{\tau}}{k}}(1-8\alpha/\sqrt{2})^{-1}, \frac{1}{k_0}(1-6k_0\sqrt{\frac{\overline{\tau}}{k}})\right)$, where both the upper bound and the lower bound have a dependence in $\sqrt{\frac{\overline{\tau}}{k}}$, while we obtain a dependance in $\frac{\overline{\tau}}{k}$. Given that Chandrasekaran et al. (2011) show that there always exist a value of γ that is valid under the assumption that $\alpha < \frac{\sqrt{2}}{12}$, this improvement might seem minor, but since γ depends on quantities that are not known in practice and need to found by trial and error, knowing that a larger interval is allowed might help finding a correct value of γ in practice. Note that this improvement is not due to the fact that we restricted ourselves to the rank one case, but to the use of sharper *incoherence measures* (see Definition 13) and improvements in the bounding scheme for the subgradients.

In fact, the possibility of choosing a value of γ which is an order of magnitude smaller is crucial for the theorem that we present next, and which extends this type of result to the recovery of several sparse p.s.d. rank one terms, using the gauge Ω.

Theorem 5 (sparse + multiple *sparse* rank-one blocks)**.** *Let $\alpha := k_0\sqrt{2\overline{\tau}/k}$ and let $\mu := (1-3\alpha)^{-1}$. Under Assumption 1, if $k_0 \leq \frac{1}{7}\sqrt{k}$, and if there exists $\kappa > 16\mu$ and $\underline{\tau}, \overline{\tau} > 0$ such that $\underline{\tau} + \overline{\tau} = 2$, with*

$$\kappa\overline{\tau}^2\frac{k_0}{k} < \underline{\tau} \leq 1 \quad \textit{and} \quad \forall j \in I_i, \quad \frac{\underline{\tau}}{k} \leq (u_j^i)^2 \leq \frac{\bar{\tau}}{k}, \tag{6.14}$$

then there exists a constant $C > 0$ such that if $k > Ck_0$, the pair (S^*, L^*) is the unique optimum of problem (6.11) *for a regularization parameter $\gamma := \mu\frac{\overline{\tau}}{k}$.*

Note that $\overline{\tau}$ is essentially the same upper bound as before, except that it is now tied with a lower bound $\underline{\tau}$; these constrained are however relaxed when C is sufficiently large, and $\underline{\tau}$ can then be chosen sufficiently small to allow for all lower bounds to hold.

6.6.1 An informal motivation for the tangent space based analysis

As first discussed in Chandrasekaran et al. (2011) and later in Negahban et al. (2012), specific subspaces play a natural role in the analysis of this type of decomposition problem.

Consider first a simple sparse + low-rank decomposition of a matrix $M = S^* + L^*$. If the decomposition is unique, then by definition there is no perturbation $(\Delta S, \Delta L)$ so that (a) $S^* + \Delta S$ has the same sparsity pattern as S^*, (b) $L^* + \Delta L$ is of rank r, and (c) $M = S^* + \Delta S + L^* + \Delta L$. Note that we then have $\Delta S + \Delta L = 0$. We continue this discussion informally to provide intuition. A particular case occurs is if this equality holds for an infinitesimal pair $(\Delta S, \Delta L)$, in which case ΔS and ΔL must each belong respectively to a certain tangent set: indeed, since $L^* + \Delta L$ belongs to the manifold of matrices of rank k, then in the limit of small ΔL, it belongs to the tangent space to the manifold of rank k matrices at L^*, a space which we will denote $\mathcal{T}_r(L^*)$; for S^* the assumption that S^* has s non zero coefficients is equivalently reformulated as the constraint that S belong the union of all the subspaces spanned by s elements of the canonical basis, which is a union of manifolds. In particular, if S^* has exactly s non zero coefficients, this fixes the support, which has to contain the support of ΔS. Since S^* is in a manifold which is simply a linear subspace, then ΔS must belong to that subspace as well, which we can denote $\mathcal{T}_s(S^*)$ and call the tangent space for S^*. To exclude the existence of non trivial pairs $(\Delta S, \Delta L)$ such that $\Delta S + \Delta L = 0$, it seems relevant to impose that $\mathcal{T}_s(S^*) \cap \mathcal{T}_r(L^*) = \{0\}$, i.e. the subspaces are in *direct sum*. If this equality holds, Chandrasekaran et al. (2011) say that the subspaces are *transverse*.

The previous discussion is non-rigorous because we reasoned informally about infinitesimal $(\Delta S, \Delta L)$. What Chandrasekaran et al. (2011) have shown is that if we solve $\min_{(S,L)} \|S\|_1 + \|L\|_{\text{tr}}$ s.t. $M = S + L$, then, for a solution $(\hat{S}, \hat{L})$, the first order optimality conditions of this optimization problem naturally decompose onto $\mathcal{T}_s(\hat{S})$, $\mathcal{T}_r(\hat{L})$ and their orthogonal complements. This type of decomposition of optimality condition on a tangent space and its complement motivated the introduction the term *decomposable norm* in Negahban et al. (2012).

In our case, L is not simply low rank, it is a sum of p.s.d. matrices L_i of rank r_i each with support in $I_i \times I_i$. We will therefore have to consider the tangent subspaces to the manifolds associated with each L_i.

6.6.2 Definition of tangent spaces and associated projections

For a symmetric sparse matrix S, let $\mathcal{T}_s(S)$ be the tangent space at S with respect to the set of symmetric sparse matrices:

$$\mathcal{T}_s(S) = \{M \in \mathbb{R}^{p \times p} \mid M = M^\top, \text{Supp}(M) \subset \text{Supp}(S)\}.$$

Next, let $\mathcal{T}_I(u)$ be the tangent space at $uu^\top$ to the manifold of rank one matrices, restricted to the space of matrices with support in $I \times I$. If we first define $\bar{\mathcal{T}}_I$, the subspace of matrices with support included in $I \times I$ with

$$\bar{\mathcal{T}}_I := \{M \in \mathbb{R}^{p \times p} \mid M = M^\top, \text{Supp}(M) \subset I \times I\},$$

then, as in Chandrasekaran et al. (2011), we can express concisely $\mathcal{T}_I(u)$ as

$$\mathcal{T}_I(u) := \{M \in \bar{\mathcal{T}}_I \mid M = uv^\top + vu^\top,\ v \in \mathbb{R}^p\}.$$

Let $\mathcal{T}_s^c(A)$ denote the orthogonal complement of $\mathcal{T}_s(A)$ in $\mathbb{R}^{p\times p}$ and $\mathcal{T}_I^c(u)$ denote the orthogonal complement[4] of $\mathcal{T}_I(u)$ in $\bar{\mathcal{T}}_I$.

The projections on the defined subspaces are respectively $\mathcal{P}_{\mathcal{T}_s(A)}(M) = M_{\mathrm{Supp}(A)}$ and $\mathcal{P}_{\mathcal{T}_I(u)}(M) = \mathcal{P}_u(M_{II})$ with

$$\mathcal{P}_u(M) := M - (I - uu^\top)M(I - uu^\top).$$

In order to simplify notations we introduce

$$\mathcal{T}_0 := \mathcal{T}_s(S^*), \quad \mathcal{T}_i := \mathcal{T}_{I_i}(u^i), \quad \bar{\mathcal{T}}_i := \bar{\mathcal{T}}_{I_i}, \quad \bar{\mathcal{T}}_{00} := \mathcal{T}_0 \cap \mathrm{span}\big((\bar{\mathcal{T}}_i)_{i\in[\![r]\!]}\big)^\perp.$$

6.6.3 First order optimality conditions

Since (6.11) is a convex optimization problem, its minima are characterized by first order subgradient conditions. The pair (S^*, L^*) with $L = \sum_i s_i u^i u^{i\top}$ is an optimum of (6.11) if and only if an only if there exists a dual Q satisfying first order optimality conditions

$$Q \in \gamma\partial\|.\|_1(S^*) \quad \text{and} \quad Q \in \partial\Omega(L^*).$$

With the introduced tangent spaces, we state the following proposition that provides sufficient conditions for the existence of a unique optimum of (6.11).

Proposition 6. *The pair (S^*, L^*) is the unique optimum of (6.11) if*

(T) $\forall i \in [\![r]\!], \quad \mathcal{T}_0 \cap \mathcal{T}_i = \{0\}$,

and there exists a dual $Q \in \mathbb{R}^{p\times p}$ such that:

(S.1) $\mathcal{P}_{\mathcal{T}_0}(Q) = \gamma\ sign(S^*)$

(S.2) $\|\mathcal{P}_{\mathcal{T}_0^c}(Q)\|_\infty < \gamma$

(L.1) $\forall i \in [\![r]\!], \quad \mathcal{P}_{\mathcal{T}_i}(Q) = u^i u^{i\top}$

(L.2) $\forall i \in [\![r]\!], \quad \lambda^+_{\max}\big(\mathcal{P}_{\mathcal{T}_i^c}(Q)\big) < 1$

(L.3) $\forall J \in \mathcal{G}_k^p \backslash \{I_1, \ldots, I_r\}, \quad \lambda^+_{\max}(Q_{JJ}) < 1$

Note that the optimality condition decompose on the subspaces of matrices with support in the sets $I_i \times I_i$ and in the remaining set of indices, the complement of $\bigcup_i I_i \times I_i$. Indeed, we can write $Q = \sum_{i=1}^n Q_{I_i I_i} + Q_{0,0}$ where $Q_{0,0}$ is the matrix whose non-zero coefficients are the coefficients of Q that are not indexed by any pair in $\bigcup_{i=1}^r I_i \times I_i$. If $Q \in \mathrm{span}(\mathcal{T}_0, \ldots, \mathcal{T}_r)$, then, we necessarily have $Q_{I_i I_i} \in \mathrm{span}(\mathcal{T}_0, \mathcal{T}_i)$ and if $\mathcal{T}_0 \cap \mathcal{T}_i = \{0\}$ then $Q_{I_i I_i}$ admits a unique decomposition $Q_{I_i I_i} = Q_i + Q_{i,0}$ with $Q_i \in \mathcal{T}_i$ and $Q_{i,0} \in \mathcal{T}_0 \cap \bar{\mathcal{T}}_i$.

[4] Note in particular that it is not the orthogonal complement in the entire space.

6.6.4 Transversality and incoherence conditions

Since we consider a convex formulation, transversality is not sufficient: we need more than an assumption that $\mathcal{T}_0 \cap \mathcal{T}_i = \{0\}$ for all i. In fact, it will be necessary to assume that $\mathcal{T}_0$ and $\mathcal{T}_i$ are not too far from being orthogonal subspaces, a property which is usually called *incoherence* (Tropp, 2004; Candès and Recht, 2009; Chandrasekaran et al., 2011). And furthermore, it will be necessary that elements of one subspace do not have a too large norm for the norm associated w.r.t. to another subspace.

Definition 13 (Incoherence measures)**.** *For i in $[\![r]\!]$, let*

$$\begin{aligned}
\zeta_{i\to 0} &= \max\{\|M\|_\infty \mid M \in \mathcal{T}_i,\ \|M\|_{\mathrm{op}} \le 1\},\\
\zeta_{0\to i} &= \max\{\|Z\|_{\mathrm{op}} \mid Z \in \mathcal{T}_0,\ \|Z\|_\infty \le 1\},\\
\zeta'_{i\to 0} &= \max\{\|\mathcal{P}_{\mathcal{T}_0}(M)\|_\infty \mid M \in \mathcal{T}_i,\ \|M\|_{\mathrm{op}} \le 1\},\\
\zeta'_{0\to i} &= \max\{\|\mathcal{P}_{\mathcal{T}_i}(Z)\|_{\mathrm{op}} \mid Z \in \mathcal{T}_0,\ \|Z\|_\infty \le 1\}.
\end{aligned}$$

Note that by definition $\zeta'_{i\to 0} \le \zeta_{i\to 0}$ and $\zeta'_{0\to i} \le 2\zeta_{0\to i}$. For this reason Chandrasekaran et al. (2011) only introduced quantities of the type $\zeta_{i\to j}$. However, given that they involve the projection of one subspace on another, the quantities $\zeta'_{i\to j}$ are the ones that really capture that the subspaces are incoherent, whereas $\zeta_{i\to j}$ is an measure of incoherence between a subspace and a norm. The quantity $\zeta'_{i\to j}$ can be much smaller than $\zeta_{i\to j}$, so the distinction is useful.

Lemma 10 (Bounds on ζ)**.**

$$\zeta'_{i\to 0} \le \zeta_{i\to 0} \le \sqrt{\frac{2\bar{\tau}}{k}}, \quad \zeta'_{0\to i} \le 2k_0\sqrt{\frac{k_0\bar{\tau}}{k}} \quad \textit{and} \quad \zeta'_{0\to i} \le \zeta_{0\to i} \le k_0.$$

We then have

Lemma 11 (Transversality)**.** *Let $\alpha := k_0\sqrt{\frac{2\bar{\tau}}{k}}$. If $\alpha < 1$, then, for all $i \in [\![r]\!]$, $\mathcal{T}_0 \cap \mathcal{T}_i = \{0\}$.*

6.7 Proofs of main theorems

We will first prove Theorem 5 and then use some of the intermediate results to prove the restricted case of Theorem 4. For proofs of the different lemmas and propositions we refer the reader to the supplementary material.

6.7.1 Proof of Theorem 5

Notice that the assumptions that $k_0 < \frac{1}{6}\sqrt{k}$ and that $\bar{\tau} \le 2$ together imply that we have $\alpha < 1/3$. In order to prove this theorem we aim to construct a dual $Q \in \mathrm{span}\{\mathcal{T}_0, \mathcal{T}_1, ..., \mathcal{T}_r\}$ satisfying **(S.1)**, **(S.2)**, **(L.1)**, **(L.2)** and **(L.3)** of Proposition 6. We can write any matrix

$Q \in \operatorname{span}(\mathcal{T}_0, \mathcal{T}_1, ..., \mathcal{T}_r)$ as $Q = \sum_{i=1}^{n} Q_{I_i I_i} + Q_{0,0}$ where $Q_{0,0}$ is the matrix whose non-zero coefficients are the coefficients of Q that are not indexed by any pair in $\cup_{i=1}^{r} I_i \times I_i$. But by Lemma 11, $\forall i \in [\![r]\!]$, $\mathcal{T}_0 \cap \mathcal{T}_i = \{0\}$, which entails that $Q_{I_i I_i}$ admits a unique decomposition $Q_{I_i I_i} = Q_i + Q_{i,0}$ with $Q_i \in \mathcal{T}_i$ and $Q_{i,0} \in \mathcal{T}_0$. Finally, given the difference of supports, $Q_{0,0}$ is clearly orthogonal to $\operatorname{span}\{\mathcal{T}_1, ..., \mathcal{T}_r\}$ which entails that $Q_{0,0} \in \mathcal{T}_0$. As a consequence, if we define $Q_0 \coloneqq Q_{0,0} + \sum_{i=1}^{r} Q_{i,0}$, then $Q = \sum_{i=0}^{r} Q_i$ provides the unique decomposition of Q such that $Q_i \in \mathcal{T}_i$ for all i.

In the next part of this proof, we consider a number of projectors and other linear transformations operating on the Q_is. Since some of these calculations are naturally written in matrix form, it is most natural to view the Q_is as vectors. For the sake of clarity, we therefore switch notations and write q_i for a vectorization of Q_i, and q for a vectorization of Q. We slightly abuse notation and still say that q_i belongs to $\mathcal{T}_i$, identify it with the corresponding matrix, etc. We also write $P_{\mathcal{T}_i}$ the matrix of the projector $\mathcal{P}_{\mathcal{T}_i}$ in the same basis as the one in which q_i is written.

With this change of notation, q is uniquely decomposed onto $\mathcal{T}_0 \oplus \mathcal{T}_1 \oplus ... \oplus \mathcal{T}_r$ and we can write

$$q = \sum_{i=0}^{r} (q_i^* + \varepsilon_i), \tag{6.15}$$

where $q_0^* = \gamma \operatorname{sign}(S^*)$, $q_i^* = u^i u^{i\top}$ for $i \in [\![r]\!]$ and $\varepsilon_i \in \mathcal{T}_i$ for $i \in \{0, 1, \dots, r\}$. Conditions **(S.1)** and **(L.1)** are satisfied if and only if $P_{\mathcal{T}_i} q = q_i^*$ for all $0 \le i \le r$, which is true if and only if $(\varepsilon_i)_{1 \le i \le r}$ solves the following system of equations:

$$\begin{cases} \varepsilon_0 + \sum_{i=1}^{r} P_{\mathcal{T}_0} q_i^* + P_{\mathcal{T}_0} \varepsilon_i = 0, \\ \\ P_{\mathcal{T}_i} q_0^* + P_{\mathcal{T}_i} \varepsilon_0 + \varepsilon_i = 0, \quad \forall i \in [\![r]\!]. \end{cases}$$

Denote $\varepsilon_{0,i} \coloneqq \mathcal{P}_{\bar{\mathcal{T}}_{I_i}}(\varepsilon_0)$ the projection of ε_0 on the set of matrices with support in $I_i \times I_i$. Note that we always have $P_{\mathcal{T}_i} \varepsilon_0 = P_{\mathcal{T}_i} \varepsilon_{0,i}$, because $\mathcal{T}_i$ is a subspace of $\bar{\mathcal{T}}_{I_i}$. Finally, note that we have $\varepsilon_0 = \sum_{i=1}^{r} \varepsilon_{0,i}$ because, by projecting the first equation above onto the subspace $\bar{\mathcal{T}}_{00}$ of matrices with zero entries on $\bigcup_{i=1}^{r} I_i \times I_i$., we get $\mathcal{P}_{\bar{\mathcal{T}}_{00}} \varepsilon_0 = 0$.

Since the sets I_i are disjoint, by projecting on the each of the spaces of matrices with support in $I_i \times I_i$ the previous system of equations, we get the equivalent set of systems:

$$\forall i \in [\![r]\!], \quad \begin{bmatrix} I & P_{\mathcal{T}_0} \\ P_{\mathcal{T}_i} & I \end{bmatrix} \begin{bmatrix} \varepsilon_{0,i} \\ \varepsilon_i \end{bmatrix} = \begin{bmatrix} \eta_0 \\ \eta_i \end{bmatrix} \quad \text{where} \quad \begin{bmatrix} \eta_0 \\ \eta_i \end{bmatrix} = \begin{bmatrix} -P_{\mathcal{T}_0} q_i^* \\ -P_{\mathcal{T}_i} q_0^* \end{bmatrix}, \tag{6.16}$$

The following lemma provides conditions for the invertibility of (6.16) and the form of the inverse matrix.

Lemma 12. *Let* $A \coloneqq \begin{bmatrix} I & P_{\mathcal{T}_0} \\ P_{\mathcal{T}_i} & I \end{bmatrix}$.
Then, with Definition 13, if $\zeta_{0 \to i} \zeta_{i \to 0} \le \alpha < 1$, *$A$ is invertible and its inverse is*

$$A^{-1} = \begin{bmatrix} I & -P_{\mathcal{T}_0} \\ -P_{\mathcal{T}_i} & I \end{bmatrix} \begin{bmatrix} (I - P_{\mathcal{T}_0} P_{\mathcal{T}_i})^{-1} & 0 \\ 0 & (I - P_{\mathcal{T}_i} P_{\mathcal{T}_0})^{-1} \end{bmatrix}.$$

Moreover, $\begin{cases} \forall v \in \mathcal{T}_i, & \|(I - P_{\mathcal{T}_i} P_{\mathcal{T}_0})^{-1} v\|_{\mathrm{op}} \leq \frac{1}{1-\alpha} \|v\|_{\mathrm{op}}, \\ \forall v \in \mathcal{T}_0, & \|(I - P_{\mathcal{T}_0} P_{\mathcal{T}_i})^{-1} v\|_{\infty} \leq \frac{1}{1-\alpha} \|v\|_{\infty}. \end{cases}$

But if we let $\alpha := k_0 \sqrt{\frac{2\bar{\tau}}{k}}$, then by Lemma 10, we have $1 - \zeta_{0\to i}\zeta_{i\to 0} \geq 1 - \alpha$ and the assumption that $k_0 < \frac{1}{6}\sqrt{k}$ entails that $\alpha < \frac{1}{3} < 1$, so, by the previous lemma, each of the systems in (6.16) has a unique solution, and the obtained $(\varepsilon_i)_{i \in [\![r]\!]}$ together with $\varepsilon_0 = \sum_{i=1}^{r} \varepsilon_{0,i}$ thus yield in (6.15) a value of q that satisfies conditions **(S.1)** and **(L.1)**.

We now prove that this value of q satisfies **(S.2)** and **(L.2)**, which requires to bound $\|P_{\mathcal{T}_0^c} q\|_{\infty}$ and $\Omega^{\circ}(P_{\mathcal{T}_i^c} q)$. Since $\Omega^{\circ}(P_{\mathcal{T}_i^c} q) \leq \|P_{\mathcal{T}_i^c} q\|_{\mathrm{op}}$, we bound this latter quantity.

Lemma 13 (Bounds on $\|P_{\mathcal{T}_0^c} q\|_{\infty}$ and $\|P_{\mathcal{T}_i^c} q\|_{\mathrm{op}}$)**.** *Assume $\zeta_{0\to i}\,\zeta_{i\to 0} \leq \alpha < 1$, and let q be defined by* (6.15)*, with $\varepsilon_0 = \sum_{i \in [\![r]\!]} \varepsilon_{0,i}$ and the pairs $(\varepsilon_{0,i}, \varepsilon_i)$ the unique solution of* (6.16)*. Then*

$$\|P_{\mathcal{T}_0^c} q\|_{\infty} \leq \max_{i \in [\![r]\!]} \|q_i^*\|_{\infty} + \zeta_{i\to 0} \|\varepsilon_i\|_{\mathrm{op}} \quad \textit{and} \quad \|P_{\mathcal{T}_i^c} q\|_{\mathrm{op}} \leq \|q_0^*\|_{\mathrm{op}} + \zeta_{0\to i} \|\varepsilon_0\|_{\infty}.$$

The following lemma provides upper bounds for the quantities $\|\varepsilon_0\|_{\infty}$ and $\|\varepsilon_i\|_{\mathrm{op}}$.

Lemma 14 (Bounds on ε_i)**.** *If $\zeta_{0\to i}\,\zeta_{i\to 0} \leq \alpha < 1$, and $(\varepsilon_i)_{i \in [\![r]\!]}$ be defined as in the previous lemma, then*

$$\|\varepsilon_0\|_{\infty} \leq \tfrac{1}{1-\alpha}\big(\frac{\bar{\tau}}{k} + \zeta'_{i\to 0} 2\gamma k_0\big) \qquad \textit{and} \qquad \|\varepsilon_i\|_{\mathrm{op}} \leq \tfrac{1}{1-\alpha}\big(2\gamma k_0 + \zeta'_{0\to i}\frac{\bar{\tau}}{k}\big).$$

Finally we obtain simplified bounds on $\|P_{\mathcal{T}_0^c} q\|_{\infty}$ and $\|P_{\mathcal{T}_i^c} q\|_{\mathrm{op}}$.

Lemma 15 (Simplified bounds on $\|P_{\mathcal{T}_0^c} q\|_{\infty}$ and $\|P_{\mathcal{T}_i^c} q\|_{\mathrm{op}}$)**.** *Let $\alpha := k_0 \sqrt{\frac{2\bar{\tau}}{k}}$. If $\alpha < 1$, for q as in Lemma 13, we have*

$$\|P_{\mathcal{T}_0^c} q\|_{\infty} \leq \frac{\bar{\tau}}{k} \frac{1 - \alpha + \alpha^2 \sqrt{2/k_0}}{1-\alpha} + \gamma \frac{2\alpha}{1-\alpha}, \quad \|P_{\mathcal{T}_i^c} q\|_{\mathrm{op}} \leq \gamma k_0 \frac{1+\alpha}{1-\alpha} + \frac{\bar{\tau}}{k} \frac{k_0}{1-\alpha}.$$

Note that the previous lemmas provide better bounds that the ones used in the proof of Theorem 2 from Chandrasekaran et al. (2011), which allows for the slightly sharper characterization:

Lemma 16. *Let $\alpha := k_0 \sqrt{\frac{2\bar{\tau}}{k}}$, if $\alpha + \frac{\alpha^2}{2k_0} < \frac{1}{3}$ then $\Gamma := \big[\frac{\bar{\tau}}{k}\frac{1}{1-3\alpha}, \frac{1}{k_0}\frac{1 - k_0\bar{\tau}/k}{1+\alpha}\big)$ is a non empty interval, and for any $\gamma \in \Gamma$, the dual matrix q defined in Lemma 13 satisfies conditions **(S.2)** and **(L.2)**.*

To conclude the proof of Theorem 5, note that the assumptions $k_0 \leq \frac{1}{7}\sqrt{k}$ and $\bar{\tau} \leq 2$ implies $\alpha + \frac{\alpha^2}{2k_0} < \frac{1}{3}$. Indeed it implies $\alpha < \frac{2}{7}$ and so $\alpha + \frac{\alpha^2}{2} < \frac{2}{7} + \frac{2}{49} = \frac{16}{49} < \frac{1}{3}$. As a consequence, Lemmas 12 and 16 apply. The last thing we need to prove is then that q satisfies condition **(L.3)**, which we prove in Appendix B.2 as

Proposition 7. *Under the assumptions of Theorem 5, $\forall J \in \mathcal{G}_k^p$, $\lambda_{\max}^{+}(Q_{JJ}) < 1$.*

6.7.2 Proof of Theorem 4

Note first that the optimization problem stated in the theorem is equivalent to

$$\min \gamma \|S\|_1 + \Omega_p(L) \quad \text{s.t.} \quad M = S + L,$$

with Ω_p the gauge associated with the p-spsd-rank.

Note that we have just removed the p.s.d. constraint and replaced the trace of L by its trace norm, which should be equivalent if the obtained matrix is p.s.d.

In order to prove this theorem we need to construct a dual $q \in \operatorname{span}\{\mathcal{T}_0, \mathcal{T}_1, ..., \mathcal{T}_r\}$ satisfying **(S.1)**, **(S.2)**, **(L.1)**, **(L.2)** of Proposition 6. Note that condition **(L.3)** is void in this context, since we are considering a unique low rank block of rank-one and with full support $[\![p]\!]$, and so it is trivially satisfied. But given the assumptions of the theorem, Lemma 16 applies immediately with $k = p$, which yields the result.

6.8 Experiments

We first perform experiments on relatively small synthetic graphs and then on a larger one.

6.8.1 First experiment

First, we consider three different LVGGM with $p = 45$ observed variables. In each case, we chose the restriction of the graph on observed variables to be a tree (with maximal degree ≤ 5), and the graph structure corresponds to latent variables that are independent given all observed variables. The interactions between latent variables and observed variables are chosen as follows :

- *model 1* has $h = 3$ latent variables; we split observed variables in three groups of size 15 and connect each group to a single latent variable.
- *model 2*: has $h = 3$ latents variables; we split observed variables in three groups of different sizes $(20, 15$ and $10)$ and connect each group to a single latent variable.
- *model 3*: has $h = 4$ latent variables; we select four overlapping groups of size 15 with 5 variables shared between each pair of consecutive groups (see Fig. B.5.(b)).

The scheme used to construct a sparse precision matrix K for a given graph is described in Appendix B.4. For each mode, we draw $50p$ random vectors from the corresponding p dimensional multivariate normal distribution and compute the associated marginal empirical covariance matrix from these observations.

We then estimate the original concentration matrix K by minimizing the score matching loss regularized either in ℓ_1-norm and Ω-gauge as in (6.8) or with the ℓ_1-norm+trace-norm ($\ell_1 + \mathrm{tr}$), as proposed by Chandrasekaran et al. (2010). As discussed in Section 6.3, for the $\ell_1 + \mathrm{tr}$ regularization, the sources are a priori only identified up to a rotation matrix. However, under the assumption that the sources are conditionally independent given observed nodes,

K_{HH} is diagonal, and when the groups of observed variables associated with each latent variables are disjoint, the columns of K_{OH} are orthogonal, and are thus proportional to the eigenvectors of $K_{OH}K_{HH}^{-1}K_{HO}$ as soon as the coefficients of the diagonal matrix K_{HH} are all distinct, by uniqueness of the SVD. They are thus identifiable, and it makes sense to estimate the columns of K_{OH} by the eigenvectors of the estimated matrix L. Obviously, for model 3, we cannot hope to recover K_{OH} with this estimator.

Figure B.5 shows the different estimated concentration matrices obtained, for the choice of hyperparameters γ and λ, that produced matrices S with the correct sparsity level and L with the correct rank.

For models 1 and 2, the size of the blocks is fixed. For model 3, we use the gauge Ω_w introduced in Section 6.4.1 which estimates as well the size of the different blocks, based on prior specified via the vector of weights w, which penalizes differently different block sizes. We use $w_k = \sqrt{k}$ which we found performs reasonably well empirically. The result show clearly that even for models 1 and 3, where, in theory the different columns of K_{OH} could be estimated with an SVD based on the formulation of Chandrasekaran et al. (2010), these columns are not so well estimated and their support would not be estimated correctly by thresholding the absolute value of the estimated coefficients (with perhaps the exception of the smallest component in model 3).

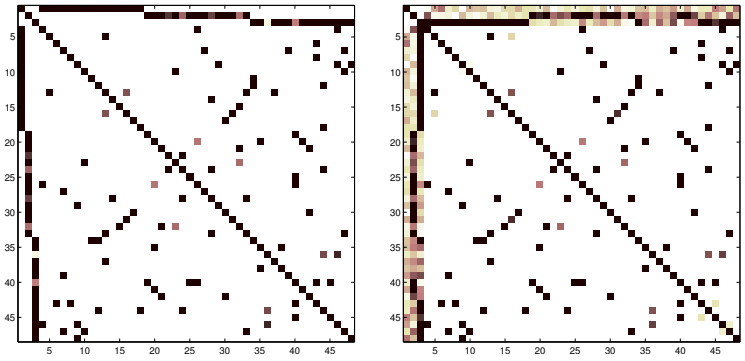
These results show empirically that the proposed formulation performs well beyond the regime for which we provide theoretical guarantees in Section 6.6: first, the experiments are in a finite data setting, so in a sense with noise; then the settings considered are of relatively low dimension with ratio k_0/k and $k_0/\sqrt{k}$ larger than in the theoretical analysis; and we obtained also convincing results for the case where blocks overlap (model 3), or the size of the blocks is estimated as well (model 2).

6.8.2 Second experiment

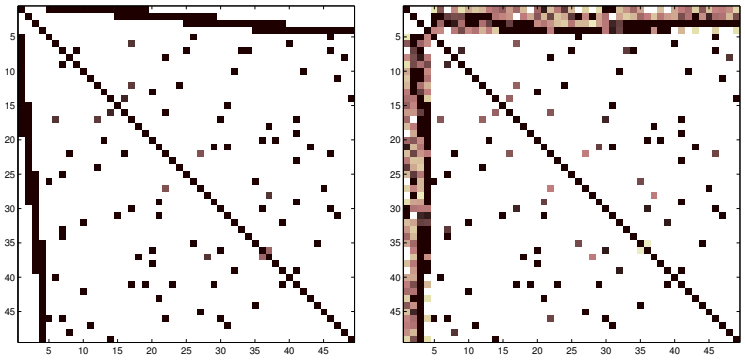
We consider a graph which is somewhat larger, with 160 nodes, corresponding to an empirical covariance matrix which is 12 times larger than the previous ones. In this case, the part of the graph corresponding to the observed variables is drawn from an Erdös-Rényi model, where each edge has a fixed appearance probability $p_s = 0.01$. We add 4 latent variables connected to non overlapping groups of 35 observed variables and we generate 2000 observations from the full graph. We compute the marginal covariance matrix as before (see Appendix B.4) and again solve (6.8) with the score matching loss to compute our estimator. Figure 6.8.2 shows the low rank component of the ground truth covariance and the low rank component obtained by our method. We clearly recover the latent structure of the graph, i.e., the four groups of 35 variables.

6.9 Conclusion

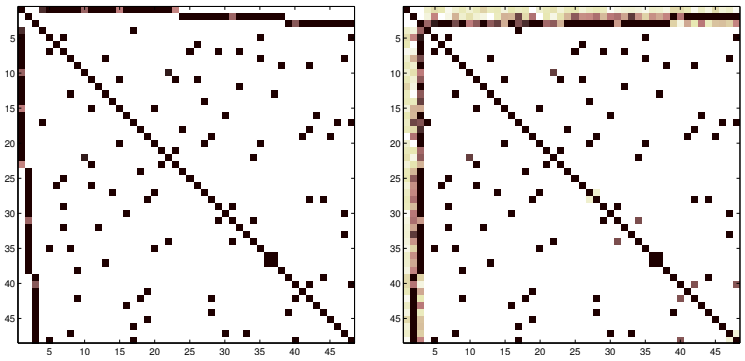
We considered a family of latent variable Gaussian graphical models whose marginal concentration matrix over the observed variables decomposes as a sparse matrix plus a low-rank matrix *with sparse factors*. We introduced a convex regularization to specifically induce this structure



(a) *model 1*, ours (d) *model 1*, $\ell_1 + \mathrm{tr}$



(b) *model 2*, ours (e) *model 2*, $\ell_1 + \mathrm{tr}$



(c) *model 3*, ours (f) *model 3*, $\ell_1 + \mathrm{tr}$

Figure 6.3: Estimated $|K_{ij}|$, for K the complete concentration matrices where the three (resp. four) first rows and columns correspond to the latent variables of *model 1* and *model 3* (resp. *model 2*) : for *model 1* in (a) ours and (d) $\ell_1 + \mathrm{tr}$ regularization; for *model 2* in (b) ours and (e) $\ell_1 + \mathrm{tr}$ regularization; for *model 3* in (c) ours and (f) $\ell_1 + \mathrm{tr}$ regularization

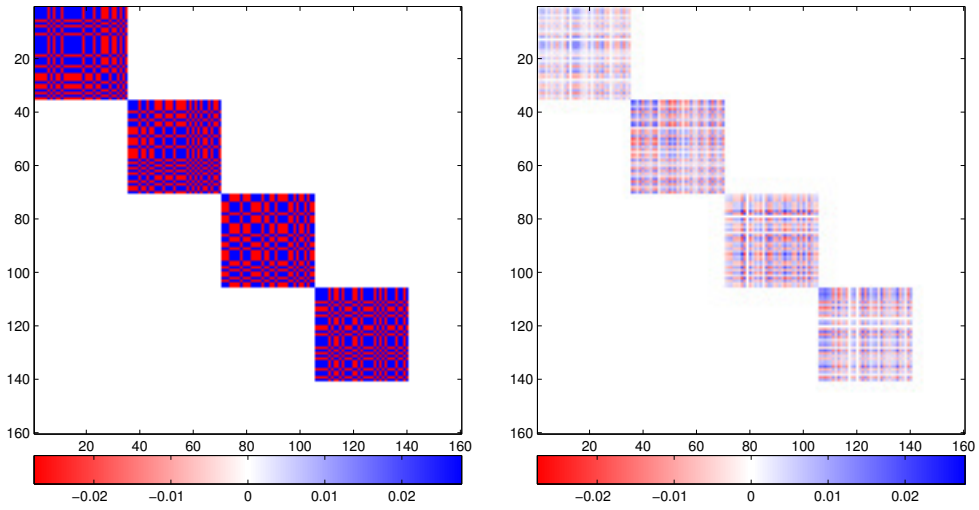


Figure 6.4: Experiment on on model with $p = 160$ observed variables and 4 unobserved. $n = 2000$ and $k = 35$. (left) low rank component of the ground truth covariance (right) low rank component obtained by our method.

on the low rank component, proposed a convex formulation to estimate both components, based on a regularized score matching loss, and proposed an efficient algorithm to solve it. We provided as well an identifiability result, that guarantees that, in the limit of an infinite amount of data, and when the blocks associated with each latent variable are disjoint, the graph structure of the whole graph, including connectivity between latent and observed variables is recovered by the proposed formulation.

Our experiments show promising results in terms of recovery of the structure of the whole graph, including when there is overlap or when cliques associated with latent variables have different sizes. Future work could study more precisely the formulations that allows for different clique sizes, and extend identifiability/recovery results in different directions.

In this chapter we considered the problem of learning the structure of Gaussian Graphical models with unobserved variables. It is worth noting that the same kind of formulation used in this work could be applied to k-sparse representative problem(Elhamifar et al., 2012). One can reformulate the k-sparse representative problem as a graphical model and define the latent structure as being the k representative, and all other data points are connected only to the representatives. In future work it would be interesting to consider this application. This would need improvements of our method Vinyes and Obozinski (2017) to scale, as discussed at the end of Chapter 4.

Chapter 7

Convex demixing by gauge minimization

In this work, we consider the problem of signal demixing into two or more components via the minimization of a sum of norms or gauges, encoding each a structural prior on the corresponding components to recover. In particular, we provide general exact recovery guarantees in the noiseless setting based on *local cumulative coherence* measures that are related to the *cumulative coherence* measures introduced in Tropp (2004), for combinations of norms (possibly coupled with a subspace constraint), that satisfy a decomposition property of the subgradient. In the case of demixing of two components, we provide finer recovery result applicable to general coercive gauges. Our general results subsume specific results from the literature for Basis Pursuit, Morphological Component Analysis, sparse+ low rank matrix decomposition and others.

7.1 Introduction

A common problem in the signal processing literature which is also highly relevant in the machine learning and statistics communities is the *decomposition* or *demixing* problem: given a signal y obtained as a linear combination of signals $x_i^\star$, that is $y = \sum_{i=1}^m x_i^\star$, with all $x_i^\star$ belonging to a finite[1] dimensional vector space that we identify with $\mathbb{R}^d$ and some prior information on the characteristics or structure of the $x_i^\star$, under what conditions are they identifiable unambiguously?

7.1.1 Formulation

We are in particular interested by the case where each $x_i^\star$ is "simple" for a given measure of complexity which is a norm or gauge ν_i, and which encourages properties such as sparsity, low

[1] All general results could be extended to infinite dimensional vector spaces without difficulties other than notational.

rankness and, more generally, the fact that x_i^* belongs to a union of subspaces or submanifolds associated with the gauge.

The motivation for considering norms (or gauges) is that it is common to consider underlying structures that are combinatorial, like sparsity, and that would a priori lead to untractable formulations for the characterization of x_i^*, but for which it is possible to construct convex relaxations that typically take the form of a gauge. More precisely, as we will illustrate by several examples in Section 7.1.2, in a number of settings, it is natural to assume that x_i^* is a combination of a small number of elements of the ambient vector space, called *atoms*, and picked from a collection $\mathcal{A}_i$, and a gauge naturally associated with this type of structure is the atomic gauge associated with $\mathcal{A}_i$ (Chandrasekaran et al., 2012).

We formulate the problem of recovering the components x_i^* as a that of finding a minimal complexity decomposition, as follows

$$\min_{x_1,\dots,x_m \in \mathbb{R}^p} \quad \sum_{i=1}^m \nu_i(x_i) \qquad \text{s.t.} \quad y = \sum_{i=1}^m x_i, \tag{7.1}$$

where for all i, ν_i is either a norm of more generally a (closed) gauge[2]. In particular, our main theorem applies to symmetric coercive gauges[3]. Considering gauges that are not norms is motivated by the fact that a symmetric coercive gauge γ can always be written under the form $\gamma(x) = \omega(x) + \iota_{\{x \in \mathcal{E}_\gamma\}}$, where ω is a norm and $\iota_{\{x \in \mathcal{E}_\gamma\}} = 0$ if $x \in \mathcal{E}_\gamma$ and $\iota_{\{x \in \mathcal{E}_\gamma\}} = \infty$ else, with $\mathcal{E}_\gamma$ a subspace associated with γ. Proving the result for these gauges thus allows us to cover a fairly natural setting in which the x_i are explicitly constrained to live in a subspace $\mathcal{E}_i$.

Notations

For any set $C \subset \mathbb{R}^d$, we denote by $\operatorname{span}(C)$ the subspace spanned linearly by elements in C, and we will denote by $\operatorname{ri}(C)$ the relative interior of C, that is the interior of C for the topology of the affine hull of C.

7.1.2 Examples

A number of problems that can either be formulated as (7.1) or as variants accounting for the presence of noise or of an additional linear map combining the elements x_i^* to form y have been considered in the literature. We provide hereafter a certain number of examples, illustrated in Figures 7.1 to 7.4, focussing for the most on the noiseless and design-less case, which correspond to the setting we will study.

Sparse + low rank. The most emblematic example of this type of decomposition in the recent literature is probably the sparse + low rank matrix decomposition problem studied in

[2] A closed gauge is a positively homogeneous proper lower semi-continuous convex function

[3] A gauge γ is said to be coercive if $(\gamma(x) = 0) \Rightarrow (x = 0)$. It is symmetric if $\gamma(x) = \gamma(-x)$. Note that although the main theorem requires symmetry, Theorem 7, which is specialized version of the main theorem for just two subspaces, has an immediate generalization to the case of non-symmetric coercive gauges.

Chandrasekaran et al. (2011). The decomposition is illustrated in Figure 7.1. To decompose a matrix $M = S + L$ into its sparse part S and low rank part L, Chandrasekaran et al. (2011) solve (7.1) with $\nu_1(S) = \gamma\|S\|_1$, where γ is a scalar, and $\nu_1(L) = \|L\|_{tr}$, where $\|\cdot\|_{tr}$ is the trace norm.

Robust PCA. Low rank + sparse matrix decomposition has extensively been studied by Candès et al. (2011) for the problem of robust PCA. Xu et al. (2010) focus on the PCA problem where some data points are outliers: given a matrix of n observations $X \in \mathbb{R}^{n\times p}$, some rows of X are outliers. The authors propose a decomposition of $X \approx L + C$, ilustrated in Figure 7.2, where L is low rank and C is row sparse. The penalty used is a combination of the trace norm and the ℓ_1/ℓ_2 norm, that writes

$$\lambda\|L\|_{tr} + \mu\|C\|_{1,2},$$

and corresponds to $\nu_1 := \lambda\|.\|_{tr}$ and $\nu_2 := \mu\|.\|_{1,2}$ with our notations.

Morphological Component Analysis. MCA (Elad et al., 2005) was introduced as the problem of decomposition of a signal onto the union of two (or more) orthogonal bases capturing each different morphological components of the signal. The original example is the *sine and spike model* applied to the separation of wispy galaxies from a starry sky background, using a discrete Fourier or DCT basis and the canonical basis in a discretized image. If the matrix of the cosines basis is denoted U then MCA can be formulated as

$$\min_{x_1,x_2'} \|x_1\|_1 + \|x_2'\|_1 \quad \text{s.t.} \quad y = x_1 + Ux_2',$$

given that U is orthonormal

$$\min_{x_1,x_2} \|x_1\|_1 + \|U^\top x_2\|_1 \quad \text{s.t.} \quad y = x_1 + x_2,$$

where we recognize (7.1) with $\nu_1 = \|\cdot\|_1$ and $\nu_1 = \|U^\top \cdot\|_1$.

Low rank tensors. Two notions of rank are classically associated to tensors: the *canonical rank* and the n-rank, the latter being the list of the mode k rank for all $k \in [\![K]\!]$, i.e., the list of the ranks of the mode-k matricizations of the tensor (Kolda and Bader, 2009). The mode k matricization of a tensor $\mathcal{W} \in \mathbb{R}^{n_1\times\ldots\times n_K}$ is the n_k by $\prod_{i\neq k} n_i$ matrix whose columns are all the mode-k fibers[4] of $\mathcal{W}$. For large tensors, one way to control the complexity is to penalize the sum (or the product) of the ranks of the mode k matricization, and some convex relaxation have been considered (Tomioka and Suzuki, 2013). Wimalawarne et al. (2014, 2016) take a different point of view and propose to learn a tensor $\mathcal{W}$ that decomposes as sum of K tensors $\mathcal{W} = \sum_{k=1}^K \mathcal{W}^{(k)}$ such that $\mathcal{W}^{(k)}$ has low mode-k rank. Since the trace-norm provides a convex relaxation of the rank constraint, the authors consider a formulation with a regularization by

[4]The mode-k fibers are the n_k dimensional vectors obtained by fixing all the indices but the kth index.

the sum of the trace norms of the matrices $W^{(k)}_{(k)}$, which are the mode k matricization of the tensor $\mathcal{W}^{(k)}$. This induces a regularization on $\mathcal{W}$ which they call the *latent trace norm*. The corresponding demixing problem matching our framework would be

$$\min_{(W^{(k)})_{k=1..K}} \sum_{k=1}^{K} \left\| W^{(k)}_{(k)} \right\|_{tr} \quad \text{s.t.} \quad \mathcal{W} = \sum_{k=1}^{K} \mathcal{W}^{(k)}.$$

where $W^{(k)}_{(k)}$ is the mode-k matricization of the tensor $\mathcal{W}^{(k)}$. Figure 7.4 illustrates the decomposition of a tensor in $\mathbb{R}^3$ in three tensors $\mathcal{W}^{(k)}$ of mode-k rank one. Haeffele and Vidal (2015) use gauge functions for tensor factorization to show, in a certain framework, the existence of multiple equivalent local minima in deep networks.

Shared and individual variable selection in multitask linear regression. Jalali et al. (2010) consider the problem of simultaneous variable selection for a group of linear regressions sharing the same variable space, in which one assumes that some variable are relevant to all regressions and some variables are only relevant to one or a few of the regressions. In this setting, each regression indexed by $k \in [\![K]\!]$ is based on n_k observations of the form $(x_i^{(k)}, y_i^{(k)})$. To couple variable selection between the different regressions, if $X^{(k)} \in \mathbb{R}^{n_k \times p}$ and $y^{(k)} \in \mathbb{R}^{n_k}$ are respectively the design matrix and the vector of labels for the kth linear regression, Jalali et al. (2010) propose to solve a problem of the form

$$\min_{M,S,R \in \mathbb{R}^{p \times K}} \sum_{k=1}^{K} \|y^{(k)} - X^{(k)} M_{:,k}\|_2^2 + \nu_1(S) + \nu_2(R) \quad \text{s.t.} \quad M = S + R,$$

where $\nu_1(S) := \mu \|S\|_1$ and $\nu_2(R) = \lambda \sum_{j=1}^{p} \|R_{j,:}\|_\infty$, so that M is decomposed as the sum of a row-sparse and of a sparse matrix (see Figure 7.3). Jalali et al. (2010) call this type of model "dirty statistical model".

The question that we investigate in this paper is under what conditions on $x_i^\star$ and ν_i the components $x_i^\star$ are the unique solution of problem (7.1).

7.2 Related work

McCoy and Tropp (2014) consider optimization problems of the form

$$\min_{x_1,x_2} \nu_1(x_1) \quad \text{s.t.} \quad \nu_2(x_2) \le \alpha, \quad y = x_1 + Q x_2,$$

with $y = x_1^\star + Q x_2^\star$, and for Q an orthonormal basis matrix drawn uniformly at random from the Haar measure one the Stiefel manifold, and show results for successful demixing when $\alpha = \nu_2(x_2^\star)$. More precisely, their work is based on a characterization for successful demixing that generalizes the *null space property* (Cohen et al., 2009), and which essentially specifies that the Minkowski sum of the tangent cones should be a direct sum. Considering *intrinsic volumes*

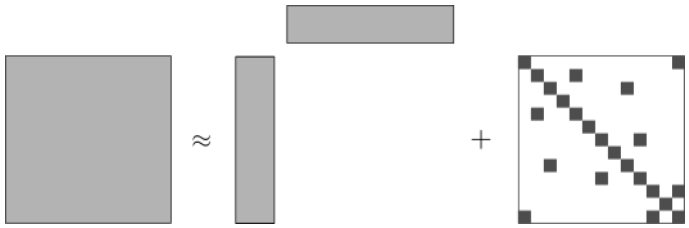


Figure 7.1: low rank + sparse matrix decomposition for a symetric matrix

Figure 7.2: low rank + row sparse matrix decomposition for PCA with noisy outliers

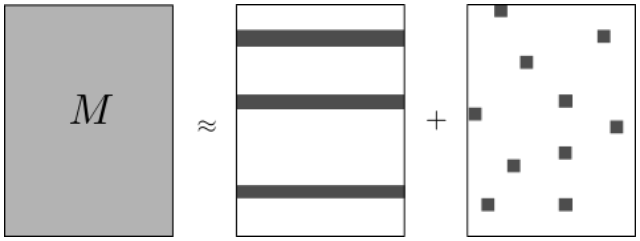


Figure 7.3: row sparse + sparse for dirty statistical models

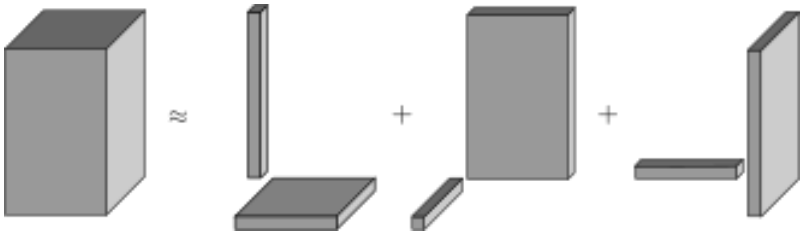


Figure 7.4: low rank tensors

of these cones, they characterize regimes in which these do or do not shrink exponentially quickly as a function of the ambient dimension, which allows them to characterize the regimes in which, asymptotically, in large dimension, with high probability, the cones are in direct sum or not, so that demixing with the convex formulation is successful or not.

Amelunxen et al. (2014), in a paper that makes several other contributions, sharpen these results by providing guarantees for fixed ambient dimension d, based on the value of the sum of the statistical dimensions (a.k.a. Gaussian complexities) of each of the cones. McCoy et al. (2014) focusses specifically on demixing and provides an accessible account of similar results. Foygel and Mackey (2014) consider again a similar formulation, in which however Q is now a rectangular sensing matrix, and provide also non-asymptotic demixing guarantees, including in the noisy setting, based on Gaussian complexity measures of the tangent cones and scaled subdifferentials.

Few papers study the case of more than two components: McCoy and Tropp (2013), who generalize the results of McCoy and Tropp (2014) to more than two components, consider a compressed sensing setting in which $y = A\sum_{i=1}^m U_i x_i^\star + \xi$, where $A \in \mathbb{R}^{N\times d}$ is a fixed sensing matrix, $U_i \in \mathbb{R}^{d\times d}$ is a random orthogonal matrix drawn at random from the Haar measure on the orthogonal group, and $\xi \in \mathbb{R}^N$ is a noise vector, one particular case being the case where A is the identity. The $x_i^\star$ are estimated by solving

$$\min_{(x_i)_{i=1..m}} \left\| A^\dagger\big(y - A\sum_{i=1}^m U_i x_i^\star\big)\right\|_2^2 \quad \text{s.t.} \quad f_i(x_i) \le f_i(x_i^\star), \tag{7.2}$$

where $(f_i)_{i=1..m}$ are convex functions. Note that solving the optimization problem requires to know the values $f_i(x_i^\star)$. The authors argue that the solution of this problem is the same as the minimizer of a Lagrangian in which the smooth part of the above objective is penalized by a term of the form $\sum_{i=1}^m \lambda_i f_i(x_i)$, for some unknown multipliers λ_i.

The constrained formulation above has the advantage that its optimality conditions are easier to analyze. In particular, when $A = I$ and $\xi = 0$, the problem reduces to a feasibility problem, which has a unique solution under a condition on the tangent cones[5] $\mathcal{C}_i$ associated with each of the f_i at $x_i^\star$.

More precisely a necessary and sufficient condition for the uniqueness of the decomposition $y = \sum_{i=1}^m U_i x_i^\star$ is that the rotated cones $\mathcal{C}_i' := U_i\mathcal{C}_i$ are such that there is no non null element $(h_1, \dots, h_m) \in \mathcal{C}_1' \times \dots \times \mathcal{C}_m'$ such that $\sum_{i=1}^m h_i = 0$, or equivalently that $-\mathcal{C}_i' \cap \sum_{j\neq i}\mathcal{C}_j' = \{0\}$, $\forall i = 1..m$.

Based on this characterization, the authors establish a phase transition in the probability of exact or stable demixing, as a function of N, d and the Gaussian complexity measures (statistical dimensions) of the tangent cones $\mathcal{C}_i$ associated with each of the f_i at $x_i^\star$. Their approach relies crucially on the use of the *random orientation model* based on the U_i, which places the cones $\mathcal{C}_i'$ in general position, and makes it possible to develop elegant arguments to bound the probabilities of cones intersection just as a function of their volumes, the latter being measured by Gaussian complexities. However, this *random orientation model* does not match

[5] The tangent cone of ν_i at $x_i^\star$ is the closure of the descent cone defined by $\big\{h_i \in \mathbb{R}^d \mid \exists \varepsilon_0 > 0, \forall \varepsilon \le \varepsilon_0,\ f_i(x_i^\star + \epsilon h_i) \le f_i(x_i^\star)\big\}$

the setting of a number of applications; and the computation of good bounds for Gaussian complexities is potentially difficult. Gu and Banerjee (2016) consider a similar setting, without the *random orientation model* and with a subgaussian sensing and noise, but they assume a *structural coherence* condition on the cones $\mathcal{C}_i$, which guarantees that the decomposition is unique when $A = I$ and $\xi = 0$, while our work is precisely concerned with means to establish this type of result from properties of the $x_i^\star$ and the ν_i.

The work that is closest to ours is Ong and Lustig (2016), which studies a particular matrix decomposition as a sum of several low rank matrices supported by sets of rows indexed by dyadic intervals. In particular, the analysis proposed in that work in the noiseless case is close to ours and introduces similar coherence measures between norms and tangent spaces; it is also similar to an analysis in Wright et al. (2013), which however introduced coherence measures between subspaces as measured by the classical operators norms, while in our work and in Ong and Lustig (2016), norms or gauge on operators induced by the ν_i are used.

Our work generalizes the exact demixing guarantees based on *cumulative coherence* measures in the noiseless setting, first by showing that it is applicable to a broader set of gauges, then by distinguishing several incoherence measures (between pairs of tangent spaces equipped with gauges, between tangent spaces and norms, and between subgradients and tangents spaces) that yield sharper conditions, and finally by proving an improved condition in the case of two components (see the discussion at the end of Section 7.4).

7.3 Subspace associated with a gauge at a point

If the gauges ν_i were simply indicators of linear subspaces $\mathcal{T}_i$ of $\mathbb{R}^d$, it would be sufficient for problem (7.1) to have a unique solution that the collection of subspaces $\mathcal{T}_i$ are in *direct sum*, which can be stated as the assumption that for any i, $\mathcal{T}_i \cap \operatorname{span}(\mathcal{T}_j)_{j \neq i} = \{0\}$.

In fact, we will show that, under some appropriate hypotheses, each of the gauges ν_i naturally associates to the corresponding element $x_i^\star \in \mathbb{R}^d$ a subspace $\mathcal{T}_i$ containing $x_i^\star$, and that these subspaces plays a key role in the analysis of problem (7.1). In particular, it is necessary that these associated subspaces $\mathcal{T}_i$ are in direct sum for (7.1) to have a unique solution, and we will sufficient conditions for uniqueness based on incoherence assumptions on the subspace.

Examples of such spaces $\mathcal{T}_i$ are the tangent spaces used in Chandrasekaran et al. (2011) and one of the subspaces introduced in Negahban et al. (2012) to formalize the concept of *separable* (a.k.a. *decomposable*) norms. This association of a tangent space $\mathcal{T}_x$ to a point x and for a norm ν, was also discussed for *separable* norms in Candès and Recht (2013); Wright et al. (2013) and Foygel and Mackey (2014) and extended to gauges in Vaiter et al. (2015a) (see also Fadili et al., 2013; Vaiter et al., 2015b).

This is best formalized based on a notion of decomposability of the subgradient (Foygel and Mackey, 2014; Vaiter et al., 2015a), as follows:

Definition 14. *We say that the gauge ν has an* orthogonally decomposable subdifferential *(o.d.s.) at $x \in \mathbb{R}^d$ if the projection of the origin 0 on the affine span of $\partial\nu(x)$ is in* $\operatorname{ri}(\partial\nu(x))$.

Indeed, Vaiter et al. (2015a) show the following characterization:

Proposition 8. *If ν has an o.d.s. at x then, denoting the projection of the origin 0 on the affine span of $\partial\nu(x)$ by q_x, letting $Q_x := \partial\nu(x) - q_x$, and $\mathcal{T}_x := span\big(Q_x\big)^\perp$ the subspace orthogonal[6] to Q_x then*

- $x, q_x \in \mathcal{T}_x$,
- *$\partial\nu(x)$ decomposes orthogonally on $(\mathcal{T}_x, \mathcal{T}_x^\perp)$ in the sense that we have*

$$\partial\nu(x) = \{q_x + q' \mid q' \in Q_x\}.$$

In other words, if $\nu_x^{\perp,\circ}$ denotes the gauge of Q_x, then $\nu_x^{\perp,\circ}$ is finite and we have

$$\partial\nu(x) = \big\{q \mid P_{\mathcal{T}_x} q = q_x,\ \nu_x^{\perp,\circ}(q - q_x) \le 1\big\}.$$

Proof. By definition we have $\langle q_x, q'\rangle = 0$ for all $q' \in Q_x$, and so we have $q_x \in (\mathcal{T}_x^\perp)^\perp$, and since $Q_x \subset \mathcal{T}_x^\perp$, then $\partial\nu(x)$ decomposes orthogonally on $\mathcal{T}_x$ and $\mathcal{T}_x^\perp$, with $\mathcal{P}_{\mathcal{T}_x}\partial\nu(x) = \{q_x\}$ and $\mathcal{P}_{\mathcal{T}_x^\perp}\partial\nu(x) = Q_x$. Then, we can write any $q \in \partial\nu(x)$ as $q = q_x + q'$ with $q' \in Q_x$ and, since $\langle x, q_x\rangle = \nu(x) = \langle x, q\rangle = \langle x, q_x\rangle + \langle x, q'\rangle$, this entails $\langle x, q'\rangle$ for any $q' \in Q_x$, so that we also have $x \in \mathcal{T}_x$. The fact that $q_x \in \mathrm{ri}(\partial\nu(x))$ (vs $q_x \in \partial\nu(x)$) or equivalently $0 \in \mathrm{ri}(Q_x)$ entails that the gauge $\nu_x^{\perp,\circ}$ is finite. □

Note that in general $\nu_x^{\perp,\circ}$ has no simple expression in terms of ν°. The reason why the o.d.s. definition assumes that $q_x \in \mathrm{ri}(\partial\nu(x))$ (vs $q_x \in \partial\nu(x)$) is that it entails that $\nu_x^\perp$ is coercive, which is a property that will be needed to be able to certify uniqueness of the solution of (7.1) with the characterization of Proposition 10. Note that if $\nu(x) = 0$, then $\partial\nu(x) = \{q \mid \langle q, x\rangle = 0,\ \nu^\circ(q) \le 1\} \ni 0$, so that $q_x = 0$. So in particular, if $x = 0$, then $\partial\nu(x) = \{q \mid \nu^\circ(q) \le 1\}$ and $\nu_x^{\perp,\circ} = \nu_x^\circ$.

Natural examples of gauges with o.d.s. are provided by *separable/decomposable* norms (Negahban et al., 2012).

Definition 15. *A gauge is said to be* separable *on a pair of subspaces $(\mathcal{M}, \mathcal{T}^\perp)$ if*

$$\nu(x + y) = \nu(x) + \nu(y) \quad \forall (x, y) \in \mathcal{M} \times \mathcal{T}^\perp.$$

Separable gauge are typical examples of gauge with an o.d.s. property.

Proposition 9. *If ν is a separable coercive gauge on $(\mathcal{M}, \mathcal{M}^\perp)$ and $x \in \mathcal{M}$ is a point at which the restriction of ν to $\mathcal{M}$ is differentiable, then ν has an o.d.s. at x with $q_x = \nabla(\nu|_{\mathcal{M}})(x)$ and $Q_x = \partial(\nu|_{\mathcal{M}^\perp})(0) = \{q \in \mathcal{M}^\perp \mid \nu^\circ(q) \le 1\}$. In general $\mathcal{M} \subset \mathcal{T}$, and if ν is coercive on $\mathcal{M}$, then $\mathcal{T} = \mathcal{M}$.*

[6] If the gauge ν has a domain (i.e. $\{x \mid \nu(x) < \infty\}$) which is included in a strict subspace $\mathcal{E}$ of $\mathbb{R}^d$ then $\mathcal{T}_x$ can also be defined as the orthogonal complement of $\mathrm{span}(Q_x)$ in $\mathcal{E}$ and not in $\mathbb{R}^d$.

Proof. It follows immediately from the rules of subdifferentiable calculus that $\partial\nu(x) = \{q_x\} \times Q_x$ with $q_x = \nabla(\nu|_{\mathcal{M}})(x)$ and $Q_x = \partial(\nu|_{\mathcal{M}^\perp})(0)$. By definition $q_x \in \mathcal{M}$ and $Q_x \subset \mathcal{M}^\perp$. So that in general, $\mathcal{T}_x^\perp \subset \mathcal{M}^\perp$ and $\mathcal{M} \subset \mathcal{T}$. We prove in Lemma 27 in the appendix, that if ν is separable on $(\mathcal{M}, \mathcal{M}^\perp)$, this entails that $\forall (p,q) \in (\mathcal{M}, \mathcal{M}^\perp)$, $\nu^\circ(p+q) = \max\big(\nu^\circ(p), \nu^\circ(q)\big)$. But this entails that $Q_x = \{q' \in \mathcal{M}^\perp \mid \nu^\circ(q') < 1\}$. Finally, if ν is coercive on $\mathcal{M}$, then ν° must be finite on $\mathcal{M}$, and Q_x has non-empty interior in $\mathcal{M}^\perp$.

□

There are many examples of norms separable on a couple $(\mathcal{M}, \mathcal{M}^\perp)$, including the Lasso and the group-Lasso. The trace norm is also separable but for a pair of spaces $(\mathcal{M}, \mathcal{T}^\perp)$ that are not the orthogonal complement of each other; however, given the form of the subdifferential of the trace norm, it is easy to see that it satisfies the o.d.s. property as well.

Note that separability at x alone is not sufficient to obtain an o.d.s. property, and vice-versa, the o.d.s. property does not imply separability. In particular, it is easy to see that a number of norms that are not separable have the o.d.s. property: we can cite among others some OWL norms like SLOPE and more generally the locally separable norm considered in Obozinski and Bach (2016).

This mismatch of concepts has led different authors to make assumptions that are hybrid between the o.d.s. property above and separability (Candès and Recht, 2013; Fadili et al., 2013; Vaiter et al., 2015a).

The o.d.s. property that we defined, while convenient, is in fact not necessary for the rest of our analysis and could be relaxed, to generalize our results to any (coercive) gauge. What is actually key is the fact that for any $q_x \in \partial\nu(x)$, the subspace $\mathcal{T}_x \coloneqq \operatorname{span}(q - q_x)^\perp$ is the same and that we have $x \in \mathcal{T}_x$, which is true for any gauge. The o.d.s. assumption could therefore be generalized by replacing the orthogonal projection in the definition by an oblique projection. This would be relevant for example in the case of the total variation on vectors associated with an undirected graph $G = (V, E)$ and defined by $\nu(x) = \sum_{\{i,j\} \in E} |x_i - x_j|$. Indeed, in that case the subdifferential at x is not orthogonally decomposable but since $\nu(x) = \|Dx\|_1$, where $D \in \mathbb{R}^{m \times d}$ is the matrix computing all pairwise differences on each edge, $\partial\nu(x) = D^\top \partial\|\cdot\|_1(Dx)$, and so the subdifferential of x is the image by a non-orthogonal transformation of a subdifferential of the ℓ_1-norm which itself is orthogonally decomposable. So changing the metric based on D would be relevant here.

7.3.1 Examples of gauges with an o.d.s.

Basic examples of norms with the o.d.s. property are the trace norm for inducing low rank structure and ℓ_1 norm for inducing sparsity

ℓ_1 norm in $\mathbb{R}^p$. For any $x \in \mathbb{R}^p$, we define the support of x as the set $\operatorname{Supp}(x) \coloneqq \{i \in [\![p]\!] \mid x_i \neq 0\}$. Let S be any particular subset of indices, and $S^c = [\![p]\!] \setminus S$, then $\|.\|_1$ is separable with

respect to $(\mathcal{T}(S), \mathcal{T}(S)^\perp)$ with $\mathcal{T}(S) := \{x \in \mathbb{R}^p \mid \mathrm{Supp}(x) \subset S\}$. In particular, for any x, we have $\mathcal{T}_x = \mathcal{T}(\mathrm{Supp}(x))$.

Trace norm for matrices in $\mathbb{R}^{p_1 \times p_2}$. If $X \in \mathbb{R}^{p_1 \times p_2}$ is a rank r matrix with $X = USV^\top$ its reduced singular decomposition (i.e. with $U \in \mathbb{R}^{p_1 \times r}$), $V \in \mathbb{R}^{p_2 \times r}$), $S \in \mathbb{R}^{r \times r}$, with U, V orthonormal and S diagonal, then the trace norm $\|.\|_{tr}$ is separable with respect to $\mathcal{M}(X) := \{UDV^\top \mid D \in \mathbb{R}^r \text{ diagonal}\}$ and $\mathcal{T}(X)^\perp := \{N \in \mathbb{R}^{p_1 \times p_2} \mid U^\top N = 0 \text{ and } NV = 0\}$ or equivalently

$$\mathcal{T}(X)^\perp := \left\{(I_{p_1} - UU^\top)M(I_{p_2} - VV^\top) \mid M \in \mathbb{R}^{p_1 \times p_2}\right\}.$$

As a consequence we have $\mathcal{T}_X = \mathcal{T}(X)$.

Trace norm with p.s.d. constraints. On the set of symmetric matrices (that can be viewed as $\mathbb{R}^{p(p+1)/2}$), the natural restriction of the trace norm on p.s.d. matrices is the gauge ν defined by $\nu(X) = \mathrm{tr}(X) + \iota_{\{X \succeq 0\}}$, where tr denotes the trace. The polar gauge is ν° with $\nu^\circ(B) = \lambda^+_{\max}(B)$ where $\lambda^+_{\max}$ is the largest nonnegative eigenvalue and 0 if all eigenvalues are negative. If X is a rank r p.s.d. matrix and $X = USU^\top$ its reduced eigenvalue decomposition, with $S \in \mathbb{R}^{r \times r}$, $S \succeq 0$, S diagonal and $U \in \mathbb{R}^{p^r}$ orthonormal, then one can check that $\partial\nu(X) = \{UU^\top + B \mid BU = 0,\ \lambda^+_{\max}(B) \leq 1\}$. We also have that ν is separable with respect to $\mathcal{M}(X) := \{UDU^\top \mid D \text{ diagonal}\}$ and $\mathcal{T}(X)^\perp := \{M \in \mathbb{R}^{p \times p} \mid M = M^\top,\ MU = 0\}$. Again, we have $\mathcal{T}_X = \mathcal{T}(X)$.

Atomic gauge. Consider the gauge $\Omega_{\mathcal{A}}$ defined by

$$\Omega_{\mathcal{A}}(x) = \inf\Big\{ \sum_{a \in \mathcal{A}} c_a \mid x = \sum_{a \in \mathcal{A}} c_a a,\ c_a \geq 0\Big\}.$$

It is a standard result that the polar gauge satisfies $\Omega^\circ_{\mathcal{A}}(s) = \max_{a \in \mathcal{A}} \langle a, s\rangle$ (see e.g., Chandrasekaran et al., 2012).

Let $\mathcal{A}_x$ be the set of elements of $\mathcal{A}$ that enter with a non-zero coefficient at least one of the optimal decompositions of x for the gauge $\Omega_{\mathcal{A}}$ and A_x the matrix whose columns are the elements of $\mathcal{A}_x$.

Then it is immediate to verify that $\partial\Omega_{\mathcal{A}}(x) = \{q \mid A_x q = 1,\ \Omega^\circ_{\mathcal{A}}(q) \leq 1\}$. But if q_x and Q_x are as in Proposition 8, then we must have $A_x q_x = 1$ and $Q_x \subset \{q' \mid A_x q' = 0\}$, which entails that $\mathrm{span}(\mathcal{A}_x) \subset \mathcal{T}_x$. We however do not have $\mathrm{span}(\mathcal{A}_x) = \mathcal{T}_x$ in general, as illustrated by the case of the group Lasso, or even more simply by the ℓ_2-norm. Indeed, in this last case $\mathcal{A}_x = \{q_x\} = \partial\nu(x)$ with $q_x = \frac{x}{\|x\|_2}$ and $\mathcal{T}_x$ is the entire space. Also, since any gauge can be written as an atomic gauge, atomic gauges do not necessarily satisfy the o.d.s. property.

7.4 Identifiability conditions

Without appropriate assumptions, Problem (7.1) is in general not well posed and even if the solution is unique there is no guarantee that the obtained decomposition yields the x_i^*.

However, under some incoherence conditions on tangent spaces we prove that the decomposition is exactly recovered. Proposition 10 states existence and uniqueness of the solution of (7.1) and Theorem 6 states simple conditions that guarantee exact decomposition for problem (7.1).

For all i, we consider the subdifferential of the gauge ν_i evaluated at x_i^*. We introduce again a number of the notations introduced in Proposition 8 but for each (ν_i, x_i^*) pair. In particular, assuming that ν_i has an o.d.s. at x_i^*, we will note q_i^* the projection of the origin on $\partial\nu_i(x_i^*)$, $Q_i = \partial\nu_i(x_i^*) - q_i^*$, $\mathcal{T}_i := \mathrm{span}(Q_i)^\perp$. Note that, as before, we have by construction $x_i^*, q_i^* \in \mathcal{T}_i$, $\langle x_i^*, q_i^* \rangle = \nu_i(x_i^*)$. For short, we will also denote $\tilde{\nu}_i^\circ := \nu_{x_i^*}^{\perp,\circ}$ the gauge of Q_i and P_i the projector on $\mathcal{T}_i$.

Proposition 10. *Let $y := \sum_{i=1}^m x_i^*$. Let $S = \{i \mid x_i^* \neq 0\}$ and $S^c = [\![m]\!] \backslash S$.*

If, for all i, ν_i is coercive[7] and has an o.d.s. at x_i^*, and if, with the above notations,

(i) the subspaces $(\mathcal{T}_i)_{i\in S}$ are in direct sum[8], i.e., $\forall i \in S$, $\mathcal{T}_i \cap span\big((\mathcal{T}_j)_{j\in S\backslash i}\big) = \{0\}$,

(ii) there exists a dual q such that

- $\forall i \in S, \quad P_i q = q_i^* \quad$ *and* $\quad \tilde{\nu}_i^\circ(P_i^\perp q) < 1$,
- $\forall j \in S^c, \quad \nu_j^\circ(q) < 1$.

then $\big(x_i^*\big)_{i=1..m}$ is the unique optimum of problem (7.1).

Proof. We start by showing that $\big(x_i^*\big)_{i=1..m}$ is an optimum before proving uniqueness. The objective can be rewritten $f(x) + g(x)$ with $f(x) = \sum_{i=1}^m \nu_i(x_i)$ and $g(x) = \iota_{\{y=\sum_i x_i\}}$. Since f is separable, its subdifferential is the Cartesian product $\partial f(x) = \partial\nu_1(x_1) \times \ldots \times \partial\nu_m(x_m)$, and since g is the indicator of a convex set (in fact a subspace), its subdifferential is the normal cone to that set at x (in fact the orthogonal subpsace), which for any x satisfying the inequality is the subspace $\partial g(x) = \{(q^\top, \ldots, q^\top)^\top \in \mathbb{R}^{m\times p} \mid q \in \mathbb{R}^d\}$. Since f and g are convex, $x^* = (x_i^*)_{i=1..m}$ is a minimum of $f + g$ if and only if $0 \in \partial(f+g)(x^*)$, but by the previous characterization, $0 \in \partial(f+g)(x^*)$ if and only if there exists $q \in \mathbb{R}^d$ such that $q \in \partial\nu_i(x_i^*)$, for all $1 \leq i \leq m$. Finally, by Proposition 8, for $i \in S$, $q \in \partial\nu_i(x_i^*)$ if and only if $P_i q = q_i^*$ and $\tilde{\nu}_i^\circ(P_i^\perp q) \leq 1$, and for $j \in S^c$, $q \in \partial\nu_j(0)$ if and only if $\nu_j^\circ(q) \leq 1$. It is thus clear that if there exists q satisfying the set of assumptions (ii), then $\big(x_i^*\big)_{i=1..m}$ is an optimum.

To prove uniqueness, suppose there is another solution $(x_i^* + n_i)_{i=1..m}$. Then $(x_i^* + n_i)_{i=1..m}$ is also a minimizer . Given the equality constraints we must have $\sum_{i=1}^m n_i = 0$ since $\sum_{i=1}^m x_i^* = y = \sum_{i=1}^m x_i^* + n_i$. Applying the subdifferential definition at $(\hat{x}_i)_{i=1..m}$, we have that for any $(q_1, ..., q_m)$ with $q_i \in \partial\nu_i(x_i^*)$,

$$\sum_{i=1}^m \nu_i(x_i^* + n_i) \geq \sum_{i=1}^m \big(\nu_i(x_i^*) + \langle q_i, n_i \rangle\big)$$

[7] We say that a gauge ν is *coercive* if $(\nu(x) = 0) \Rightarrow (x = 0)$.

[8] For a pair of subspaces, Chandrasekaran et al. (2011) use the expression "transverse subspaces".

Note that we can apply this decomposition for indices $j \in S^c$, for which $\mathcal{T}_j = \{0\}$ and we can let $q_j^* = 0$, which is consistent with it definition. Decomposing the subdifferentials on $\mathcal{T}_i$ and $\mathcal{T}_i^\perp$, yields $q_i = q_i^* + P_i^\perp(q_i)$ and $q = q_i^* + P_i^\perp(q)$. Using this decompositions we have,

$$\langle q_i, n_i \rangle = \left\langle q_i^* + P_i^\perp(q_i), n_i \right\rangle = \left\langle q - P_i^\perp(q) + P_i^\perp(q_i), n_i \right\rangle = \left\langle P_i^\perp(q_i - q), n_i \right\rangle + \langle q, n_i \rangle .$$

Thus,

$$\sum_{i=1}^m \langle q_i, n_i \rangle = \sum_{i=1}^m \left\langle P_i^\perp(q_i - q), n_i \right\rangle = \sum_{i=1}^m \left\langle P_i^\perp(q_i) - P_i^\perp(q), P_i^\perp n_i \right\rangle,$$

where in the first equality we use the fact that $\sum_{i=1}^m n_i = 0$. We can select any subgradient $(q_1, ..., q_m)$. Let $q_i^c \in \operatorname{Arg\,max}_{q' \in Q_{x_i^*}} \langle P_i^\perp n_i, q' \rangle$ where $Q_{x_i^*} = P_i^\perp \partial \nu_i(x_i^*)$ as in Proposition 8, and let $q_i = q_i^* + q_i^c$. Note that, by definition, $q_i^c \in \mathcal{T}_i^\perp$, so that $P_i^\perp q_i = q_i^c$, which entails that $\langle P_i^\perp n_i, P_i^\perp q_i \rangle = \tilde{\nu}_i(P_i^\perp n_i)$ and $\tilde{\nu}_i^\circ(P_i^\perp q_i) = 1$. Finally, by the Fenchel-Young inequality, we have $\langle P_i^\perp n_i, P_i^\perp q \rangle \leq \tilde{\nu}_i^\circ(P_i^\perp q)\, \tilde{\nu}_i(P_i^\perp n_i)$, and so we have

$$\sum_{i=1}^m \langle q_i, n_i \rangle \geq \sum_{i=1}^m \left(1 - \tilde{\nu}_i^\circ(P_i^\perp q)\right) \tilde{\nu}_i(P_i^\perp n_i). \tag{7.3}$$

But by assumption, $1 - \tilde{\nu}_i^\circ(P_i^\perp q) > 0$ for all $i = 1..m$ (in particular, this is true for $j \in S^c$, because $\tilde{\nu}_j^\circ = \nu_j^\circ$). So, $\sum_{i=1}^m \langle q_i, n_i \rangle$ is strictly positive unless $\tilde{\nu}_i(P_i^\perp n_i) = 0$ for all $i = 1..m$. For all $j \in S^c$, $P_j^\perp$ is the identity and $(\tilde{\nu}_j(n_j) = \nu_j(n_j) = 0) \Rightarrow (n_j = 0)$. For $i \in S$, the o.d.s. property implies (by Proposition 8) that $\tilde{nu}_i^\circ$ is finite which equivalently means that $\tilde{\nu}$ is coercive and so $(\tilde{\nu}_i(P_i^\perp n_i) = 0) \Rightarrow P_i^\perp n_i = 0$, which entails that $n_i \in \mathcal{T}_i$ for all i. Finally since $\sum_{i \in S} n_i = 0$, we have that $n_i = -\sum_{j \in S \setminus i} n_j$ so that $n_i \in \mathcal{T}_i \cap \operatorname{span}\left((\mathcal{T}_j)_{j \neq i}\right) = \{0\}$ by assumption (i). So $n_i = 0$ for all $i \in S$ as well and this completes the proof of uniqueness of the decomposition. □

Our main theorem essentially states that if the subspaces $\mathcal{T}_i$ equipped with the gauge ν_i are sufficiently *incoherent* (and in particular if each the vectors q_i^* dually associated with x_i^* are themselves sufficiently *incoherent* with the other subspaces $(\mathcal{T}_j)_{j \neq i}$, then Problem (7.1) has a unique solution, which is $(x_i^*)_{i=1..m}$.

This requires to introduce measures of coherence: we introduce measure of coherence similar to the ones defined in Chandrasekaran et al. (2011), and *cumulated coherence* that generalize the notion essentially introduced in Tropp (2004) (see the discussion after Theorem 6).

Definition 16. *(Generalized simple and cumulative coherences) Let $\nu_i^\circ, \mathcal{T}_i, \tilde{\nu}_i^\circ, P_i$ and $P_i^\top$ be defined as in Proposition 10. We define the coherences $\zeta_{ij}, \zeta_{ij}^\perp$ between pairs of subspaces, the coherence between a subgradient and a subspace $\zeta_{ij}^*, \zeta_{ij}^{*,\perp}$ and the corresponding cumulative coherences $\alpha, \alpha^*, \alpha_i^\perp, \alpha_i^{*,\perp}$ as follows:*

- $\zeta_{ij} := \max\left\{\nu_i^\circ(P_i u_j) \mid u_j \in \mathcal{T}_j,\ \nu_j^\circ(u_j) \leq 1\right\}$ *and* $\alpha = \max_i \sum_{j \neq i} \zeta_{ij}$,
- $\zeta_{ij}^\perp := \max\left\{\tilde{\nu}_i^\circ(P_i^\perp u_j) \mid u_j \in \mathcal{T}_j,\ \nu_j^\circ(u_j) \leq 1\right\}$ *and* $\alpha_i^\perp = \sum_{j \neq i} \zeta_{ij}$,
- $\zeta_{ij}^* := \nu_i^\circ(P_i q_j^*)$ *and* $\alpha^* = \max_i \sum_{j \neq i} \zeta_{ij}^*$,

- $\zeta_{ij}^{*,\perp} := \tilde{\nu}_i^\circ(P_i^\perp q_j^*)$ *and* $\alpha_i^{*,\perp} = \sum_{j\neq i} \zeta_{ij}^{*,\perp}$.

Theorem 6. *Let $y = \sum_{i=1}^m x_i^*$ and $(\nu_i)_{i=1..m}$ be a collection of symmetric gauges such that ν_i is coercive and has an o.d.s. at x_i^*. Let $\mathcal{T}_i$ be defined as before. Let P_i be the projector on $\mathcal{T}_i$. Let q_i^* be the unique element in $\partial\nu_i(x_i^*) \cap \mathcal{T}_i$. With the* cumulative coherences *from Definition 16: If $\alpha < 1$ and $\max_i \frac{\alpha^*}{1-\alpha}\alpha_i^\perp + \alpha_i^{*,\perp} < 1$, then $(x_i^*)_{i=1..m}$ is the unique solution to* (7.1).

As before, let $S := \{i \mid x_i^* \neq 0\}$. Note that $\zeta_{ij} = \zeta_{ij}^* = 0$ unless $i, j \in S$, and that $\zeta_{ij}^\perp = \zeta_{ij}^{*,\perp} = 0$ for $j \in S^c$, but that $\alpha_j^\perp$ and $\alpha_j^{*,\perp}$ are non zero in general for $j \in S^c$.

Proof. The proof consists in showing that the assumptions of Proposition 10 hold. First, by applying Lemma 24 in Appendix C.1 with $\Omega_i = \nu_i^\circ$ for $i = 1..m$, we get that $\alpha < 1$ implies that the spaces $\mathcal{T}_i$ are in direct sum.

Second we show that there exists $q \in \operatorname{span}(\mathcal{T}_i)_{i=1..m}$ such that for all $i = 1..m$, $P_i\, q = q_i^*$. In particular, we show that we can write $q = \sum_{i\in S}(q_i + \varepsilon_i)$ with $\varepsilon_i \in \mathcal{T}_i$. With this parameterization, and assuming, without loss of generality that $S = [\![s]\!]$, the previous equality constraints yield a linear system of equations that can be written in matrix form as

$$\begin{bmatrix} q_1^* \\ q_2^* \\ \vdots \\ q_s^* \end{bmatrix} = \begin{bmatrix} P_1 & P_1P_2 & \dots & P_1P_s \\ P_2P_1 & P_2 & \dots & P_2P_s \\ & \vdots & & \\ P_sP_1 & P_sP_2 & \dots & P_s \end{bmatrix} \begin{bmatrix} q_1^* + \varepsilon_1 \\ q_2^* + \varepsilon_2 \\ \vdots \\ q_s^* + \varepsilon_s \end{bmatrix},$$

with the constraint that $\varepsilon_i \in \mathcal{T}_i$ for all $i = 1..m$. To express subspace constraints, and assuming that $\mathcal{T}_i$ is a subspace of dimension d_i, let $T_i \in \mathbb{R}^{d\times r_i}$ be the matrix whose columns forms an orthonormal basis of $\mathcal{T}_i$. Then there exists unique $b_i, c_i \in \mathbb{R}^{r_i}$ such that $q_i^* = T_i b_i$ and $\varepsilon_i = T_i c_i$. Moreover, we must have $P_i = T_i T_i^\top$. The previous linear with subspace constraints is thus equivalent to the problem without constraints:

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_s \end{bmatrix} = \underbrace{\begin{bmatrix} I_{r_1} & T_1^\top T_2 & \dots & T_1^\top T_s \\ T_2^\top T_1 & I_{r_2} & \dots & T_2^\top T_s \\ & \vdots & & \\ T_s^\top T_1 & T_s^\top T_2 & \dots & I_{r_s} \end{bmatrix}}_{A} \begin{bmatrix} b_1 + c_1 \\ b_2 + c_2 \\ \vdots \\ b_s + c_s \end{bmatrix}. \tag{7.4}$$

Now, the matrix A above is naturally partitioned into $A = (A_{ij})_{1\leq i,j\leq s}$ with for all i, $A_{ii} = I_{r_i} \in \mathbb{R}^{r_i\times r_i}$ the identity matrix and in general $A_{ij} = T_i^\top T_j$.

By associating to $v_i \in \mathbb{R}^{r_i}$ the norm $\omega_i(v_i) := \nu_i^\circ(T_i v_i)$, we precisely have that $\omega_i(A_{ij}v_j) = \nu_i^\circ(T_i T_i^\top T_j\, v_j)$, so that

$$\max_{v_j\in\mathbb{R}^{r_j},\, \omega_j(v_j)\leq 1} \omega_i(A_{ij}v_j) = \max_{u_j\in\mathcal{T}_j,\, \nu_j^\circ(u_j)\leq 1} \nu_i^\circ(P_i u_j) \leq \zeta_{ij}.$$

Let $b := ({b_1}^\top, \dots, {b_s}^\top)^\top$ and similarly $c := (c_1^\top, \dots, c_s^\top)^\top$.

Given that $\alpha < 1$, by Lemma 25, A is invertible and if we let $B := A^{-1}$ and consider the partitioning of B into $(B_{ij})_{1\leq i,j\leq s}$, with same dimensions respectively as the blocks of $(A_{ij})_{1\leq i,j\leq s}$, then

$$\max\left\{\, \omega_i(B_{ij}v_j) \mid v_j \in \mathbb{R}^{r_j},\, \omega_j(v_j) \leq 1 \,\right\} \leq \frac{1}{1-\alpha}. \tag{7.5}$$

We can thus invert the system (7.4) to get the unique solution

$$b + c = A^{-1}b \quad \text{so that} \quad c = A^{-1}(I - A)b.$$

But if we let $\tilde{b} := (A - I)b$, then $\tilde{b}_i = \sum_{j\neq i} A_{ij}b_j$ so that $\omega_i(\tilde{b}_i) \leq \sum_{j\neq i} \zeta^*_{ij}\, \omega_j(b_j) = \nu^\circ_j(q^*_j) \leq \alpha^*$. Then, since by definition we have $\varepsilon_i = T_i\, c_i$, using inequality (7.5), we have

$$\max_i \nu^\circ_i(\varepsilon_i) = \max_i \omega_i(c_i) \leq \frac{1}{1-\alpha} \max_j \omega_j(-\tilde{b}_j) \leq \frac{\alpha^*}{1-\alpha}.$$

Finally, we have $\quad P^\perp_i q = \sum_{j\in S\setminus i} P^\perp_i \varepsilon_j + \sum_{j\in S\setminus i} P^\perp_i q^*_j, \quad$ so that, for all $i \in [\![m]\!]$,

$$\begin{aligned}
\tilde{\nu}^\circ_i(P^\perp_i q) &\leq \sum_{j\in S\setminus i} \tilde{\nu}^\circ_i(P^\perp_i \varepsilon_j) + \sum_{j\in S\setminus i} \tilde{\nu}^\circ_i(P^\perp_i q^*_j) \\
&\leq \sum_{j\in S\setminus i} \zeta^\perp_{ij}\, \nu^\circ_j(\varepsilon_j) + \sum_{j\in S\setminus i} \zeta^{*,\perp}_{ij}\, \nu^\circ_j(q^*_j) \\
&\leq \sum_{j\in S\setminus i} \zeta^\perp_{ij} \frac{\alpha^*}{1-\alpha} + \sum_{j\in S\setminus i} \zeta^{*,\perp}_{ij} \leq \alpha^\perp_i \frac{\alpha^*}{1-\alpha} + \alpha^{*,\perp}_i,
\end{aligned}$$

which concludes the proof. □

In this theorem, we chose to introduce four different quantities $\alpha, \alpha^\perp_i, \alpha^*$ and $\alpha^{\perp,*}_i$. These quantities can be viewed as generalizations of the *cumulative coherence* parameters discussed in Tropp (2004) and introduced in slightly earlier work by the same author. This line of work however sought to obtain uniform results over all possible signals expressed as any combination of a small number of atoms (each of the latter corresponding to one of our subspaces), while in our context the subspaces depend in general on the signal, which motivates several non-uniform definitions. For a more precise discussion of the connection with the results in Tropp (2004), see Section 7.5.1. The distinction between $\alpha^\perp_i$ and α is necessary since $\mathcal{T}_i$ and $\mathcal{T}^\perp_i$ are not equipped with the same gauge in general, and even if they do, taking into account the projection is important since the spaces are incoherent. Note that by definition $\alpha^* \leq \alpha$ and $\alpha^{*,\perp}_i \leq \alpha^\perp_i$. However, α and $\alpha^\perp_i$ are defined over entire subspaces $\mathcal{T}_i$, which are potentially large, and have unfavorable worst case elements, when q^*_i is an element that can have itself a very small projection over all other subspaces $\mathcal{T}_j$.

7.4.1 Special case of two blocks

In the special case where there are only two blocks, we can exploit the structure of A^{-1}, to obtain more precise bounds.

Theorem 7. *Consider the same setting and the same definitions for $\zeta_{ij}, \zeta^*_{ij}, \zeta^\perp_{ij}, \zeta^{*,\perp}_{ij}$ as in Theorem 6 in the particular case where $m = 2$, but without assuming that the gauges ν_i are symmetric. Let $\zeta^{*,-}_{ij} := \nu^\circ_i(-P_i q^*_j)$. A sufficient condition for (x^*_1, x^*_2) to be the unique solution of Problem (7.1) when $y = x^*_1 + x^*_2$, is that, for $(i,j) \in \{(1,2),(2,1)\}$,*

$$\zeta_{ji}\zeta_{ij} < 1 \qquad \text{and} \qquad \zeta^{*,\perp}_{ij} + \zeta^\perp_{ij}\frac{\zeta_{ji}\zeta^*_{ij} + \zeta^{*,-}_{ji}}{1-\zeta_{ji}\zeta_{ij}} < 1.$$

Note that, if ν_i is symmetric, we have $\zeta^{*,-}_{ij} = \zeta^*_{ij}$.

Proof. First, note that if $0 \le \zeta_{12}\,\zeta_{21} < 1$ and if $M \in \mathcal{T}_1 \cap \mathcal{T}_2$ then the inequalities

$$\nu^\circ_1(M) = \nu^\circ_1(P_1P_2M) \le \zeta_{12}\,\zeta_{21}\,\nu^\circ_1(M),$$

show that we must have $M = 0$, i.e. $\mathcal{T}_1 \cap \mathcal{T}_2 = \{0\}$. Second, $\zeta_{12}\,\zeta_{21} < 1$ entails that $I - P_1P_2$ is invertible. Indeed, for any $x_1 \in \mathcal{T}_1$,

$$\begin{aligned}(1-\zeta_{12}\zeta_{21})\nu^\circ_1(x_1) &\le \nu^\circ_1(x_1) - \zeta_{12}\nu^\circ_2(P_2x_1)\\ &\le \nu^\circ_1(x_1) - \nu^\circ_1(P_1P_2x_1) \le \nu^\circ_1\big((I-P_1P_2)x_1\big),\end{aligned}$$

so if $x_1 \in \mathcal{T}_1\backslash\{0\}$ then by Lemma 26, with $\nu^\circ_1(x_1) > 0$ or $\nu^\circ_1(-x_1) > 0$ which entails that $(I - P_1P_2)x_1 \neq 0$. This shows that $I - P_1P_2$ is invertible on $\mathcal{T}_1$ and that for all $x_1 \in \mathcal{T}_1$, $\nu^\circ_1\big((I - P_1P_2)x_1\big) \le (1-\zeta_{12}\zeta_{21})^{-1}\nu^\circ_1(x_1)$.

Finally, if we project the equality $q = q^*_1 + \varepsilon_1 + q^*_2 + \varepsilon_2$ on $\mathcal{T}_1$ and $\mathcal{T}_2$, we obtain respectively

$$\varepsilon_1 = -P_1q^*_2 - P_1\varepsilon_2 \qquad \text{and} \qquad \varepsilon_2 = -P_2q^*_1 - P_2\varepsilon_1.$$

But substituting one equation in the other and solving for the remaining ε_i yields

$$\varepsilon_1 = (I-P_1P_2)^{-1}(P_1P_2q^*_1 - P_1q^*_2) \quad \text{and} \quad \varepsilon_2 = (I-P_2P_1)^{-1}(P_2P_1q^*_2 - P_2q^*_1),$$

and we have, using the symmetry of the gauge, that

$$\nu^\circ_1(\varepsilon_1) \le \frac{\nu^\circ_1(P_1P_2q^*_1 - P_1q^*_2)}{1-\zeta_{12}\zeta_{21}} \le \frac{\zeta_{12}\zeta^*_{21} + \zeta^{*,-}_{12}}{1-\zeta_{12}\zeta_{21}} \quad \text{and symmetrically} \quad \nu^\circ_2(\varepsilon_2) \le \frac{\zeta_{21}\zeta^*_{12} + \zeta^{*,-}_{21}}{1-\zeta_{21}\zeta_{12}}.$$

Finally, for $(i,j) \in \{(1,2),(2,1)\}$,

$$\tilde\nu^\circ_i(P^\perp_i q) = \tilde\nu^\circ_i(P^\perp_i q^*_j + P^\perp_i\varepsilon^*_j) \le \zeta^{*,\perp}_{ij} + \zeta^\perp_{ij}\frac{\zeta_{ji}\zeta^*_{ij} + \zeta^{*,-}_{ji}}{1-\zeta_{ji}\zeta_{ij}},$$

hence the result. □

Note that this result could easily be extended to the case of asymmetric gauges by introducing $\zeta^{*,-}_{ij} := \nu^\circ_i(-P_iq^*_j)$, in which case, the condition simply becomes

$$\zeta^{*,\perp}_{ij} + \zeta^\perp_{ij}\frac{\zeta_{ji}\zeta^*_{ij} + \zeta^{*,-}_{ji}}{1-\zeta_{ji}\zeta_{ij}} < 1.$$

7.5 Illustrative examples

7.5.1 Basis Pursuit

As a first illustration we consider the case of Basis Pursuit. Consider the classical ℓ_1 minimization problem

$$\min_{\beta\in\mathbb{R}^m} \sum_{i=1}^m |\beta_i| \quad \text{s.t.} \quad y = \sum_{i=1}^m \beta_i u_i, \tag{7.6}$$

with $u_i \in \mathbb{R}^m$ and $\|u_i\| = 1$, and where $y = \sum_{i=1}^m \beta_i^* u_i$. If we let $x_i = \beta_i u_i$ and $\nu_i(x_i) = \|x_i\| + \iota_{\{|\langle x_i, u_i\rangle| = \|x_i\|\}}$ then (7.1) is equivalent to (7.6).

We clearly have $x_i^* = \beta_i^* u_i$ or equivalently $\beta_i^* = \langle u_i, x_i^*\rangle$. We also have $\nu_i^\circ(q) = |\langle u_i, q\rangle|$, and, as a consequence,

$$\partial\nu_i(x_i^*) = \{q_i^* + q_i^c \mid q_i^* = \operatorname{sign}(\beta_i^*)\, u_i,\ \langle u_i, q_i^c\rangle = 0\} \quad \text{and} \quad \mathcal{T}_i = \operatorname{span}(\beta_i^* u_i).$$

Computation of ζ_{ij} and ζ_{ij}^*. Clearly, for all $(i,j) \in S \times S$, we have

$$\zeta_{ij} = \max\{|\langle u_i, x_j\rangle| \mid \|x_j\| \le 1, x_j \in \mathcal{T}_j\} = |\langle u_i, u_j\rangle| = \zeta_{ij}^*.$$

Computation of $\zeta_{ij}^\perp$ and $\zeta_{ij}^{*,\perp}$. We should distinguish the case $i \in S$ and $i \in S^c$.

- If $i \in S$, then $Q_{x_i^*} = \{u \mid \langle u_i, u\rangle = 0\}$ so that $\tilde\nu_i^\circ(q) = \iota_{\{\langle u_i, q\rangle = 0\}}$. As a consequence $\tilde\nu_i^\circ(P_i^\perp q) = 0$ for any q, which entails that $\zeta_{ij}^{*,\perp} = \zeta_{ij}^\perp = 0$ and $\alpha_i^{*,\perp} = \alpha_i^\perp = 0$, for all $i \in S$.
- But if $i \in S^c$, then $\tilde\nu_i^\circ(q) = \nu_i^\circ(q) = |\langle u_i, q\rangle|$ and $P_i^\perp = I$ so that, for all $j \in S$, $\zeta_{ij}^{*,\perp} = \zeta_{ij}^\perp = |\langle u_i, u_j\rangle|$.

Finally, we get $\alpha^* = \alpha = \max_{i\in S} \sum_{j\in S\setminus i} |\langle u_i, u_j\rangle|$, and, $\forall j \in S^c$, $\alpha_j^{*,\perp} = \alpha_j^\perp = \sum_{i\in S} |\langle u_i, u_j\rangle|$. With these notations, Theorem 6 states that $(\beta_i^*)_{i=1..m}$ is the unique solution of (7.6) if $\alpha < 1$ and $\forall j \in S^c$, $\alpha_j^\perp \frac{\alpha}{1-\alpha} + \alpha_j^\perp < 1$, which is equivalent to $\alpha + \alpha_j^\perp < 1$. This leads to the sufficient condition:

Proposition 11. *A sufficient condition for exact recovery of $(\beta_i^*)_{i=1..m}$ in Problem (7.6) is*

$$\max_{j\in S} \sum_{i\in S\setminus j} |\langle u_i, u_j\rangle| + \max_{j\in S^c} \sum_{i\in S} |\langle u_i, u_j\rangle| < 1. \tag{7.7}$$

This condition is stronger than a classical condition based on the *coherence* or *mutual coherence*. The concept of *coherence* of a matrix U was original introduced by Donoho and Elad (2003) as the quantity $\mu := \|U^\top U\|_\infty$, where $\|\cdot\|_\infty$ is the entrywise ℓ_∞ norm and U is the matrix whose columns are dictionary elements u_i that match the ones appearing in our problem. An immediate consequence of Proposition 11 is that a sufficient condition for exact recovery in (7.6) is that $\mu(2s-1) < 1$, where $s = |S|$, which is a classical result first appearing as Theorem 7 in Donoho and Elad (2003).

In fact an even closer result is stated in Tropp (2004): Indeed, Tropp introduced *cumulative coherence* which is defined as $\mu_1(k) := \max_{|S'|\le k, S'\subset[\![m]\!]} \max_{j\in S'^c} \sum_{i\in S'} |\langle u_i, u_j\rangle|$. Since $\mu_1(k)$ is a

uniform upper bound over all choices of support of size k, we clearly have $\alpha \leq \mu_1(s-1)$ and $\alpha_j^\perp \leq \mu_1(s)$, which leads to a less stringent sufficient condition then the one based on *mutual coherence* and which is

$$\mu_1(s-1) + \mu_1(s) < 1,$$

a condition which is discussed in Proposition 3.7 in Tropp (2004). The quantities $\alpha, \alpha^*, \alpha_i^\perp$ and $\alpha_i^{*,\perp}$ that we introduced can thus be viewed as non-uniform generalizations of cumulative coherence, that are appropriate in our general setting.

7.5.2 MCA

We can treat the case or Morphological Component Analysis in two ways. First, it can be viewed as a particular case of Basis Pursuit. In that case, the support is composed of S_1 the non-zero coefficients of x_1 and of S_2 the non-zero coefficients of x_2'. Let I_1 and I_2 be respectively the index sets associated to the coefficients on Φ^1 and Φ^2. So that $S_1 = S \cap I_1$ and $S_2 = S \cap I_2$.

With $\mu^1(S_1) \coloneqq \sum_{i \in S_1} |\langle \phi_i^1, \phi_j^2 \rangle|$ and $\mu^2(S_2) \coloneqq \sum_{j \in S_2} |\langle \phi_i^1, \phi_j^2 \rangle|$,

$$\forall j \in S_1, \quad \sum_{i \in S \setminus j} |\langle u_i, u_j \rangle| = \mu^2(S_2) \quad \text{and} \quad \forall j \in S_2, \quad \sum_{i \in S \setminus j} |\langle u_i, u_j \rangle| = \mu^1(S_1),$$

$$\forall j \in I_1 \setminus S_1, \quad \sum_{i \in S} |\langle u_i, u_j \rangle| = \mu^2(S_2) \quad \text{and} \quad \forall j \in I_2 \setminus S_2, \quad \sum_{i \in S} |\langle u_i, u_j \rangle| = \mu^1(S_1),$$

so that condition (7.7), can be rewritten

$$2 \max\big(\mu^1(S_1), \mu^2(S_2)\big) < 1. \tag{7.8}$$

A uniform version of this condition in which, for $i \in \{1,2\}$, $\mu^i(S_i)$ is replaced by its uniform upper bound over all subsets S_i of size k_i of I_i has appeared as the *cluster coherence* condition.

But we can obtain a slightly improved condition using Theorem 7: Indeed, let $\nu_i(x_i) \coloneqq \|{\Phi^i}^\top x_i\|_1$, $i \in \{1,2\}$; since $\zeta_{ij}^* \leq \zeta_{ij}$ and $\zeta_{ij}^{*,\perp} \leq \zeta_{ij}^\perp$, the condition

$$\forall (i,j) \in \{(1,2),(2,1)\}, \qquad \zeta_{ij}^\perp \frac{1 + \zeta_{ji}}{1 - \zeta_{ji}\zeta_{ij}} < 1 \tag{7.9}$$

is sufficient for the condition of Theorem 7 to hold. But here, it is immediate to check that $\zeta_{12} = \zeta_{12}^\perp = \mu^2(S_2)$ and $\zeta_{21} = \zeta_{21}^\perp = \mu^1(S_1)$. So that (7.9) holds if and only if

$$\max\big(\mu^1(S_1), \mu^2(S_2)\big) + 2\,\mu^1(S_1)\,\mu^2(S_2) < 1, \tag{7.10}$$

which is strictly weaker than (7.8) since if $y < x$ then $x + 2xy < x + 2x^2$ and, for $x > 0$, $x + 2x^2 < 1$ is equivalent to $2x < 1$, which shows the result for $x \coloneqq \max\big(\mu^1(S_1), \mu^2(S_2)\big)$ and $y \coloneqq \min\big(\mu^1(S_1), \mu^2(S_2)\big)$.

7.6 Conclusion

Spaces defined as direct sum of subspaces immediately associate to signals a canonical decomposition into components, and thus demixing is well-posed. In general demixing problems are ill-posed unless appropriate complexity measures are introduced to favor simple decomposition.

Natural scale invariant complexity measures that lead to convex formulations of the demixing problems are norms or more generally gauges. In particular, sparsity inducing norms yield decompositions with a small number of elementary components. In has been shown in the literature that many very natural sparsity inducing norms are naturally associated with subspaces, and that if these subspaces are not too far from being pairwise orthogonal, or, stated slightly differently, if these spaces are *incoherent* for an appropriate measure of coherence associated with the norms, then it can be shown that the spaces supporting the true decomposition are in direct sum and that a convex optimization problem both identifies those subspaces containing the solution and yields the unique decomposition.

In this work, we show that this type of results generalizes to combinations of two or more symmetric coercive gauges (which includes the case of two or more norms). In particular, we propose a simple condition under which such a norm can be associated a pair of complementary orthogonal subspaces, and provide simple condition under the form of an inequality on a combination of *cumulated coherences* associated with individual subspaces to guarantee that a optimal decomposition can be recovered uniquely as a solution of a convex demixing optimization problem.

Our results recover known results for several particular cases and also produce novel results. In particular, we recover exactly recovery results for Basis Pursuit based on *cumulative coherence*, we obtain results based on *cluster coherence* that are slightly better than those reported in the literature.

Chapter 8

Conclusion and perspectives

In this thesis we considered machine learning problems in which the parameter is a structured matrix and formulations in which the structure is induced via convex regularization. In particular, we considered structured problems where the parameter is an additive combination of different components. In terms of the convex regularizers, we consider the general family of atomic norms. They have the advantage that they cover a broad family of norms, they have been studied and have associated theoretical guarantees, and they are relevant in the context of additive combination, because when considering two signals that are added, each one with a structure that is encoded by an atomic norm, their structure is naturally encoded by the atomic assiciated with the union of the atoms form the two original norms.

The first part of the work contributed by this thesis, considers the problem of solving efficiently empirical risk minimization problem when an atomic norm is used as a regularizer or as a constraint. There has been a significant amount of research on structured convex regularizations. From an algorithmic point of view, a number of algorithms have been proposed, in particular based on proximal methods. However there are many norms for which either the computation of a proximal operator is too costly or there is no efficient algorithm known to compute it. Another family of algorithms, the conditional gradient algorithms or Frank-Wolfe methods are well matched to the structure of atomic norms: indeed the LMO amounts to solve a maximization over all atoms of the dot product of an atom and the current gradient of the loss function, which correspond to the computation of a dual norm. Therefore, as soon as the dual norm can be computed efficiently Frank-Wolfe methods are well suited.

When there was a regain of interest for Frank-Wolfe methods in machine learning around 2012 (Bach et al., 2012c; Jaggi, 2013; Lacoste-Julien et al., 2012), one of the motivations was that these algorithms produced a sequence of sparse approximations to a solution and therefore seemed well matched to some optimization problems in which precisely a sparse solution is sought. However, a big gap in terms of convergence speed and sparsity of the solution remained between problems solved Frank-Wolfe methods and problems where proximal methods could be applied with Frank-Wolfe methods often showing sublinear convergence rate and not

necessary yielding sparse solution because of the large number of steps needed to converge. In recent work (Lacoste-Julien and Jaggi, 2015), with the use of away steps, the variant method pairwise Frank-Wolfe was proven to have a linear convergence when the set of atoms is bounded.

In this thesis (Chapter 4) we proposed to use a fully corrective Frank-Wolfe algorithm that can perform well in practice and keeps a sparse solution at every step but which needs to solve a complete optimization problem at each step. When considering sparse solutions, these intermediary problems remain relatively small and can be efficiently solved. The advantage is that the solution remains sparse along the optimization and the use of active-set for quadratic programming allows us to take advantage of warm start, which would not be possible with other techniques such as interior point methods. Our proposed algorithm in Chapter 4 performs well in practice as long as the final solution is sparse enough. We have considered a linear regression problem but our algorithm can be generalized to smooth loss functions. A recent paper of Locatello et al. (2017a) derives an extension of fully corrective Frank-Wolfe algorithm for smooth loss functions and derives convergence rates in the case of a bounded set of atoms. Later this year Locatello et al. (2017b) extended the convergence analysis to convex cones and derive sublinear ($\mathcal{O}(1/t)$) convergence on general smooth and convex objectives, and linear convergence ($\mathcal{O}(e^{-\alpha t})$) on strongly convex objectives.

With the algorithm developed in the first part of this thesis(Chapter 4), we can efficientlty solve problems regularized by complex matricx norms, for example the norms proposed in (Richard et al., 2014) which induce a decomposition with factors that are simultaneously low rank and sparse. This work on martix norms was inspired by a line of search on sparse + low rank models where several variants of these models were considered, among others, robust sparse PCA(Candès et al., 2011; Xu et al., 2010; Liu et al., 2010), and sparse subspace clustering with outliers (Elhamifar and Vidal, 2013). An important application the problem of recovering the structure of probabilistic graphical models, whith applications in genetics. In particular, a challenging problem is the one of confounding factors, corresponding to unobserved variables that have influence on observed variables. The work of Chandrasekaran et al. (2010) allowed to identify the sparse part of the model and the subspace spanned by unobserved variables but it did not allow to identify which observed variables were affected by each unobserved variable. From an algorithmic perspective, the proposed approach is an alternated algorithm that uses the algorithm proposed in Chapter 4 to update the component with complex structure and uses proximal steps to update the plain sparse component. From a theoretical perspective, we show that by using the norm of Richard et al. (2014) it is possible to identify this connectivity under specific assumptions and in the asymptotic setting. Our theoretical result suggests that in order to retrieve the full connectivity, the unobserved variables need to be connected to a certain number of observed variables. Indeed, a sufficient condition is the incoherence between the different matrix components that suggests that the effect of latent variables needs to be well spread over observed variables, so that it is possible to demix the sparse and low rank components. In the experiments we considered a variant of the regularization for graphical model where latent variables are associated with overlapping blocks of variables or blocks of

different sizes. For such graphs, the preliminary results show that we are able to recover the structure which would motivate further analysis.

In Chapter 7 we try to show how the previous recovery result can be extended to more than components and for general atomic norms. We consider the problem of signal demixing into two or more components via the minimization of a sum of norms or gauges, encoding each a structural prior on the corresponding components to recover. This work is motivated by the fact that a number of problems can be formulated this way: sparse + low rank decomposition Chandrasekaran et al. (2011), robust PCA (Candès et al., 2011; Xu et al., 2010), low rank tensors (Wimalawarne et al., 2014, 2016) among others. The conditions proposed are fairly simple, based on a notion of cumulative incoherence that generalizes the concepts introduced in Chapter Tropp (2004). Our results recover known results for several particular cases, for example Basis Pursuit, and also produce slightly better results than those reported in the literature for the case of two components.

In this thesis, we considered convex matrix sparsity problems by using the framework provided by atomic norms, which naturally allow to consider underlying structures that are combinatorial and are a relevant tool for demixing problems.

Appendix A

Column generation for atomic regularization

A.1 Proof of Proposition 1

Proposition 12. *If f is assumed lower bounded by 0 and if $\rho > f(0)$, or more generally if the level sets of $x \mapsto f(x) + \gamma_{\mathcal{A}}(x)$ are bounded and ρ is sufficiently large, then the sequence $(\bar{x}^t)_t$ produced by the FCFW algorithm applied to the truncated cone constrained problem (4) and initialized at $(\bar{x}^0; \tau^0) = (0; 0)$ is the same as the sequence $(x^t)_t$ produced by Algorithm (1) initialized with $x^0 = 0$, with equivalent sequences of subproblems, active sets and decomposition coefficients.*

Proof. :

Notations: 1_k (rep. 0_k) denotes the vector in $\mathbb{R}^k$ with all entries equal to 1 (resp. 0).

We being by showing that the Frank-Wolfe directions computed for the regularized and the constrained problems are related via a simple relation, already discussed in Yu et al. (2014); Harchaoui et al. (2015).

First note that, the set of extreme points of the truncated cone $\{(x, \tau) \mid \gamma_{\mathcal{A}}(x) \leq \tau \leq \rho\}$ is

$$\bar{\mathcal{A}} = \{(0; 0)\} \cup \big\{(\rho a; \rho) \mid a \in \mathcal{A}\big\}.$$

so that all its non zero extreme points are in bijection with those of $\mathcal{A}$. Then, for a given point x, the Frank-Wolfe directions computed respectively by FCFW in problems (5) and (4) are

$$\begin{cases} a^* & := \arg\max_{a \in \mathcal{A}} \langle \nabla f(x), a \rangle \\ \bar{a}^* & := \arg\max_{(\rho a; \rho u) \in \bar{\mathcal{A}}} \langle \nabla f(x), a \rangle + u, \end{cases}$$

and we have

$$\bar{a}^* = \begin{cases} (0; 0) & \text{if} \quad \gamma^\circ_{\mathcal{A}}(\nabla f(x)) \leq 1 \\ (\rho a^*; \rho) & \text{otherwise,} \end{cases}$$

which shows that unless the atom $(0;0)$ is selected in $\bar{\mathcal{A}}$, it is the image of the regular FW direction mapped via $a \mapsto (\rho a; \rho)$. Note also that the atom $(0;0)$ is special in $\bar{\mathcal{A}}$ in that it is the only one for which the second component is different than ρ.

We now prove, by induction on t, the following statement:
$\mathcal{P}^t$: *Letting $(\bar{x}^t, \bar{A}^t, \bar{c}^t)$ denote the triple of values of x, the matrix of active atoms of $\bar{\mathcal{A}}$ and the vector of coefficients $\bar{c}$ of decomposition of x on these atoms, all generated by the FCFW algorithm, then (a) the first column of $\bar{A}^t$ is a column of zeroes corresponding to the atom $(0;0)$, so that we can write*

$$\bar{A}^t = \begin{pmatrix} 0 & \rho A^t \\ 0 & \rho u^t \end{pmatrix} \in \mathbb{R}^{(d+1)\times(1+k_t)} \quad \text{and} \quad \bar{c}^t = \begin{pmatrix} d_0^t \\ d^t \end{pmatrix} \in \mathbb{R}^{1+k_t},$$

(b) setting $x^t := \bar{x}^t$, and $c^t = \rho d^t$ we have that (x^t, A^t, c^t) is the t-th corresponding triple produced by Algorithm (1) and (c) $\tau^t = \rho(1 - d_0^t) < \rho$ so that the truncation constraint $\{\tau \le \rho\}$ is inactive.

To prove $\mathcal{P}^0$, note that if $(x^0; \tau^0) = (0;0)$, then we trivially have $\bar{A}^0 = (0;0)$ and $\bar{\mathcal{A}}^0$ has the desired form, we have $\bar{x}^0 = c_0^0 \cdot 0_d = x^0$ with $\bar{c}^0 = d_0^0 = 1$ so that $\bar{c}^0$ satisfies the simplex constraints; finally $\tau^0 < \rho$.

We now assume $\mathcal{P}^{t-1}$ is true and prove that so is $\mathcal{P}^t$. In the FCFW algorithm, the new direction chosen cannot be $(0;0)$ since $d_0^{t-1} > 0$, which entails this atoms is already in the active set and because the algorithm is fully corrective (which prevents the forward direction to be an atom already in the active set), so that it must be of the form $(\rho a^t, \rho)$ which, given that by induction $\bar{x}^{t-1} = x^{t-1}$, entails that a^t is indeed the same direction as the one chosen by Algorithm (1) .

Letting $\bar{A}^t$ is the matrix whose columns are the atoms used in the expansion of x^t, then $\bar{x}^t = A^t c^t$ and letting $x^t = \bar{x}^t$, then the triple (x^t, A^t, c^t) is the one generated by Algorithm (1) . This entails that $\bar{A}^t$ is indeed of the announced form and that the sub-matrix A^t is indeed the one used by Algorithm (1) .

Now the optimization problem solved in the corrective step of FCFW is thus

$$\begin{aligned} \min_{x,\tau,d} \quad & f(x) + \tau \quad \text{s.t.} \\ & x = \rho A^t d, \quad \tau = \rho u^t d, \quad \bar{c} = (c_0; d) \in \Delta^{k_t+1}, \end{aligned}$$

with $u^t = 1_{k_t}^\top$ and k_t the number of currently active atoms.

Eliminating x and τ we obtain

$$\min_{d \ge 0} f(\rho A^t d) + \rho\, 1_{k_t}^\top d \quad \text{s.t.} \quad 1_{k_t}^\top d \le 1,$$

and with the change of variable $c = \rho d$, we get

$$\min_{c \ge 0} f(A^t c) + \|c\|_1 \quad \text{s.t.} \quad \|c\|_1 \le \rho$$

But, since $\gamma_{\mathcal{A}^t}(x) = \inf\left\{\|c\|_1 \mid c \in \mathbb{R}_+^{k_t},\ x = A^t c\right\}$, we can rewrite the previous problem equivalently as

$$\min_x f(x) + \gamma_{\mathcal{A}^t}(x) \quad \text{s.t.} \quad \gamma_{\mathcal{A}^t}(x) \le \rho.$$

We first conclude the argument assuming $f \geq 0$ and $\rho > f(0)$. In that case, we have

$$\gamma_{\mathcal{A}^t}(x^t) \leq f(x^t) + \gamma_{\mathcal{A}^t}(x^t) \leq f(0) + \gamma_{\mathcal{A}^t}(0) = f(0) < \rho,$$

so that the inequality constraint is inactive for all t at the optimum in the two last problems above and can be removed. We thus showed that the optimization problem of the corrective step of the FCFW algorithm on problem (4) is equivalent to the problem solved at step 6 of Algorithm (1) , and that $\|c^t\|_1 < \rho$ which entails that $d_0^t = 1 - \|d^t\|_1 = 1 - \frac{1}{\rho}\|c\|_1 > 0$ and so that the atom $(0;0)$ remains in $\bar{\mathcal{A}}^{t+1}$. The induction step is completed which thus proves the result.

Now, if we do not assume that f is lower bounded, but we assume instead that the level sets of $f + \gamma_{\mathcal{A}}$ are bounded, then Algorithm (1) generates a sequence x^t which is bounded since the sequence $\left(f(x^t) + \gamma_{\mathcal{A}^t}(x^t)\right)_t$ is a monotonically decreasing sequence. But since for all x, $f(x) + \gamma_{\mathcal{A}^t}(x) \geq f(x) + \gamma_{\mathcal{A}}(x)$, the monotonicity also implies that the sequence $(x^t)_t$ remains in the bounded set $\{x \mid f(x) + \gamma_{\mathcal{A}}(x) \leq f(0)\}$. Since f is assumed continuous this entails that $\left(f(x^t)\right)_t$ is bounded which entails that so is $\left(\gamma_{\mathcal{A}^t}(x^t)\right)_t$ so if ρ is chosen such that $\rho > \sup_t \gamma_{\mathcal{A}^t}(x^t)$ then the FCFW algorithm applied on problem (4) will generate the same sequence as Algorithm (1) . This value of ρ is not known a priori, but is required by neither algorithms. □

A.2 Rank one updates of the Hessian and its inverse in active-set

Let H^t be the Hessian of the quadratic problem in active-set algorithm and B^t its inverse. Let Q be the Hessian of the quadratic function f. We have $H^t = A^{t\top} Q A^t$. We use the Sherman–Morrison–Woodbury matrix inversion formula in the following equations.

When we add an atom a_{t+1}, we have updates

$$H^{t+1} = \begin{bmatrix} H^t & v \\ v^\top & a_{t+1}^\top Q a_{t+1} \end{bmatrix}$$

and

$$B^{t+1} = \begin{bmatrix} B^t + \alpha B^t v v^\top B^t & -\alpha B^t v \\ -\alpha (B^t v)^\top & \alpha \end{bmatrix}$$

where $v = A^{t\top} Q a_{t+1}$ and $\alpha = (a_{t+1}^\top Q a_{t+1} - v^\top B v)^{-1}$.

When removing an atom, H^{t+1} is obtained removing the corresponding column and row. For clarity, let us assume that we want to remove the last atom. We have

$$H^t = \begin{bmatrix} \tilde{H}^t & v \\ v^\top & \nu \end{bmatrix}$$

and

$$B^{t+1} = \begin{bmatrix} \tilde{B}^t & w \\ w^\top & \beta \end{bmatrix}.$$

Then,

$$H^{t+1} = \tilde{H}^t,$$

$$B^{t+1} = \tilde{B}^t + \frac{\beta \tilde{B}^t v v^\top \tilde{B}^t - (w^\top v - 1)(w v^\top \tilde{B}^t + \tilde{B}^t v w^\top) + v^\top \tilde{B} v w w^\top}{(w^\top v - 1)^2 - \beta v^\top \tilde{B}^t v}.$$

Appendix B

Learning the effect of latent variables in Gaussian Graphical models with unobserved variables

Proof of Lemma 9

Claim 1. *Let $Y \in \mathbb{R}^{p\times p}$ be a symmetric matrix. The polar gauge of Ω writes*

$$\Omega^\circ(Y) = \max_{I\in\mathcal{G}_k^p} \lambda^+_{\max}(Y_{II}). \tag{B.1}$$

Proof. $\Omega^\circ(Y) = \max_{\Omega(X)\leq 1} \operatorname{tr}(Y^\top X) = \max_{\substack{\|u\|_0=k\\ \|u\|_2=1}} u^\top Y u = \max_{I\in\mathcal{G}_k^p} \lambda^+_{\max}(Y_{II}).$ □

Lemmas charactering the subgradients

In the following lemmas we express the subgradients of the ℓ_1 norm and Ω as decomposed on the tangent subspaces. The result for the ℓ_1-norm is well known.

Lemma 17. *(Characterization of ℓ_1 subgradient) $Q \in \gamma\partial\|.\|_1(S^*)$ if and only if*

(A.1) $\mathcal{P}_{\mathcal{T}_0}(Q) = \gamma\, sign(S^*)$

(A.2) $\|\mathcal{P}_{\mathcal{T}_0^c}(Q)\|_\infty \leq \gamma$

We then characterize the subgradient of the gauge we have introduced.

Lemma 18. *(Characterization of the subgradient of Ω)*
If L^* is of the form $L^* = \sum_{i=1}^r s_i u^i u^{i\top}$, with $\operatorname{Supp}(u^i) \subset I_i$ and $I_i \cap I_j = \varnothing$ for all $i \neq j$, we have that $Q \in \partial\Omega(L^*)$ if and only if

(B.1) $\forall i \in [\![r]\!]$, $\mathcal{P}_{\mathcal{T}_i}(Q) = u^i u^{i\top}$

(B.2) $\forall i \in [\![r]\!],\ \lambda^+_{\max}\big(\mathcal{P}_{\mathcal{T}_i^c}(Q)\big) \leq 1$

(B.3) $\forall J \in \mathcal{G}^p_k \backslash \{I_1, \dots, I_r\},\ \lambda^+_{\max}(Q_{JJ}) \leq 1$

Proof. By the characterization of the subgradient of a gauge we have $Q \in \partial\Omega(L^*)$ if and only if

$$\max_{I \in \mathcal{G}^p_k} \lambda^+_{\max}(Q_{II}) \leq 1 \quad \text{and} \quad \langle Q, L^* \rangle = \Omega(L^*). \tag{B.2}$$

The inequality implies immediately **(L.3)** and that $u^\top Q u \leq 1$ for any unit vector u such that $\|u\|_0 \leq k$. By definition of L^*, the equality becomes $\sum_{I_i \in \mathcal{I}} s_i(u^{i\top} Q u^i - 1) = 0$. Since all terms of the sum are non negative we must have $u^{i\top} Q u^i = 1$. Since $1 = u^{i\top} Q u^i = u^{i\top} Q_{I_i I_i} u^i$ and we have $\lambda^+_{\max}(Q_{I_i I_i}) \leq 1$, u^i must be an eigenvector of $Q_{I_i I_i}$ with eigenvalue 1. Given that $Q_{I_i I_i}$ as a real symmetric matrix, admits an orthonormal basis of eigenvectors, we can thus write $Q_{I_i I_i} = u^i u^{i\top} + W_i$ with $W_i \in \mathcal{T}_i^c$ and $\lambda^+_{\max}(W_i) \leq 1$. Since the previous decomposition shows that $W_i = \mathcal{P}_{\mathcal{T}_i^c}(Q)$ and $\mathcal{P}_{\mathcal{T}_i}(Q) = u^i u^{i\top}$ we have shown **(L.1)** and **(L.2)**. □

Proof of Proposition 6

Claim 2. *The pair (S^*, L^*) is the unique optimum of* (6.11) *if*

(T) $\forall i \in [\![r]\!], \quad \mathcal{T}_0 \cap \mathcal{T}_i = \{0\},$

and there exists a dual matrix $Q \in \mathbb{R}^{p \times p}$ such that:

(S.1) $\mathcal{P}_{\mathcal{T}_0}(Q) = \gamma\ sign(S^*)$

(S.2) $\|\mathcal{P}_{\mathcal{T}_0^c}(Q)\|_\infty < \gamma$

(L.1) $\forall i \in [\![r]\!], \quad \mathcal{P}_{\mathcal{T}_i}(Q) = u^i u^{i\top}$

(L.2) $\forall i \in [\![r]\!], \quad \lambda^+_{\max}\big(\mathcal{P}_{\mathcal{T}_i^c}(Q)\big) < 1$

(L.3) $\forall J \in \mathcal{G}^p_k \backslash \{I_1, \dots, I_r\}, \quad \lambda^+_{\max}(Q_{JJ}) < 1.$

Proof. The **(S.1)**, **(S.2)**, **(L.1)**, **(L.2)** and **(L.3)** clearly imply that there exist a dual matrix Q such that $Q \in \big(\gamma \partial \|\cdot\|_1(S^*)\big) \cap \partial\Omega(L^*)$, which is the first order subgradient condition that characterizes the optima of (6.11).

To show that the solution is *unique* we show that (S^*, L^*) must be obtained as the unique solution of an equivalent minimization problem. Indeed, consider the gauge $\gamma_I(M) = \mathrm{tr}(M) + \iota_{\{M \succeq 0\}} + \iota_{\{\mathrm{Supp}(M) \subset I \times I\}}$. It is immediate to verify that the polar gauge is γ_I° such that $\gamma_I^\circ(Q) = \lambda^+_{\max}(Q_{II})$. Thus $\Omega^\circ(Q) = \max_{I \in \mathcal{G}^p_k} \gamma_I^\circ(Q)$ and, taking polars, we get that

$$\Omega(M) = \inf \Big\{ \sum_{I \in \mathcal{G}^p_k} \gamma_I(M^{(I)}) \mid M = \sum_{I \in \mathcal{G}^p_k} M^{(I)} \Big\}. \tag{B.3}$$

As a consequence, problem (6.11) is equivalent to

$$\min_{S, (L^{(I)})_{I \in \mathcal{G}^p_k}} \gamma \|S\|_1 + \sum_{I \in \mathcal{G}^p_k} \gamma_I(L^{(I)}) \quad \text{s.t.} \quad M = S + \sum_{I \in \mathcal{G}^p_k} L^{(I)}. \tag{B.4}$$

In particular, if $\big(S^*, (L^{(I)*})_{I\in\mathcal{G}_k^p}\big)$ is an optimal solution of (B.4), and if $L^* = \sum_{I\in\mathcal{G}_k^p} L^{(I)*}$, then (S^*, L^*) is an optimal solution of (6.11). Conversely, (S^*, L^*) is an optimal solution of (6.11), then any optimal decomposition of L^* obtained from (B.3) yields an optimal solution of (B.4).

So clearly, if the solution to (B.4) is unique, then so must be that of (6.11).

Let's then assume that $\big(S^* + N_0, (L^{(I)*} + N^{(I)})_{I\in\mathcal{G}_k^p}\big)$ is another optimal solution to (B.4). Since matrices in both solutions sum to M, we must necessarily have

$$N_0 + \sum_{I\in\mathcal{G}_k^p} N^{(I)} = 0. \tag{B.5}$$

Let $Q^{(I)} \in \partial\gamma_I(L^{(I)*})$ and $Q_0 \in \partial\|\cdot\|_1(S^*)$. Then, by convexity, we have

$$\begin{aligned}\gamma\|S^*\|_1 + \sum_{I\in\mathcal{G}_k^p} \gamma_I(L^{(I)*}) &= \gamma\|S^* + N_0\|_1 + \sum_{I\in\mathcal{G}_k^p} \gamma_I(L^{(I)*} + N^{(I)}) \\ &\geq \gamma\|S^*\|_1 + \sum_{I\in\mathcal{G}_k^p} \gamma_I(L^{(I)*}) + \langle Q_0, N_0\rangle + \sum_{I\in\mathcal{G}_k^p} \langle Q^{(I)}, N^{(I)}\rangle.\end{aligned} \tag{B.6}$$

Consistently with previous notations, we denote by $\mathcal{I} = \{I_i, \dots, I_r\}$ the set of blocks such that $L^{(I)*} \neq 0$, and $Q_i := Q_{I_i}$, $N_i := N_{I_i}$.

Now, γ_I is a decomposable gauge in the sense of Negahban et al. (2012): in particular if $L^{(I_i)*} = L_i^* := U^i D^i U^{i\top}$, with U^i an orthonormal matrix and D^i a diagonal matrix, then $\partial\gamma_{I_i}(L_i^*) = \big\{Q_i^* + Q_i^c \mid Q_i^c \in \mathcal{T}_i^c,\ \gamma_{I_i}^\circ(Q_i^c) \leq 1\big\}$, with $Q_i^* = U^i U^{i\top}$. Note that, since $\mathcal{T}_i$ and $\mathcal{T}_i^c$ are orthogonal, for all $i \in [\![r]\!]$, any $Q_i \in \gamma_{I_i}(L_i^*)$ is such that $\mathcal{P}_{\mathcal{T}_i}(Q_i) = Q_i^*$. In the rest, of the proof, we choose $Q_i = Q_i^* + Q_i^c$ with $Q_i^c \in \mathcal{T}_i^c$ such that

$$\gamma_{I_i}\big(\mathcal{P}_{\mathcal{T}_i^c}(N_i)\big) = \langle \mathcal{P}_{\mathcal{T}_i^c}(N_i), Q_i^c\rangle = \langle \mathcal{P}_{\mathcal{T}_i^c}(N_i), Q_i\rangle \tag{B.7}$$

(this is clearly possible because for $M \in \mathcal{T}_i^c$, we have precisely that $\gamma_{I_i}(M) = \max\{\langle M, Z\rangle \mid Z \in \mathcal{T}_i^c,\ \gamma_{I_i}^\circ(Z) \leq 1\}$).

Given that there exists, by assumption of the theorem, Q such that conditions **(S.1)**,**(S.2)**,**(L.1)**,**(L.2)**,**(L.3)** are satisfied, we have in particular that $\mathcal{P}_{\mathcal{T}_i}(Q) = Q_i^*$, $\forall i \in \{0\} \cup [\![r]\!]$, with $Q_0^* = \gamma\,\mathrm{sign}(S^*)$.

So, we have

$$
\begin{aligned}
0 &\overset{\text{(B.6)}}{\geq} \langle Q_0, N_0\rangle + \sum_{I\in\mathcal{G}_k^p} \langle Q^{(I)}, N^{(I)}\rangle \\
&= \sum_{i=0}^{r} \big(\langle Q_i^*, N_i\rangle + \langle \mathcal{P}_{\mathcal{T}_i^c}(Q_i), N_i\rangle\big) + \sum_{I\in\mathcal{G}_k^p\setminus\mathcal{I}} \langle Q^{(I)}, N^{(I)}\rangle \\
&= \sum_{i=0}^{r} \big(\langle Q, N_i\rangle + \langle \mathcal{P}_{\mathcal{T}_i^c}(Q_i - Q), N_i\rangle\big) + \sum_{I\in\mathcal{G}_k^p\setminus\mathcal{I}} \langle Q^{(I)}, N^{(I)}\rangle \\
&\overset{\text{(B.5)}}{=} \sum_{i=0}^{r} \langle Q_i - Q, \mathcal{P}_{\mathcal{T}_i^c}(N_i)\rangle + \sum_{I\in\mathcal{G}_k^p\setminus\mathcal{I}} \langle Q^{(I)} - Q, N^{(I)}\rangle \\
&\overset{\text{(B.7)}}{\geq} \gamma\|\mathcal{P}_{\mathcal{T}_0^c}(N_0)\|_1\big(1 - \tfrac{1}{\gamma}\|\mathcal{P}_{\mathcal{T}_0^c}(Q)\|_\infty\big) + \sum_{i=1}^{r} \gamma_{I_i}\big(\mathcal{P}_{\mathcal{T}_i^c}(N_i)\big)\big(1 - \gamma_{I_i}^\circ\big(\mathcal{P}_{\mathcal{T}_i^c}(Q)\big)\big) \\
&\qquad\qquad + \sum_{I\in\mathcal{G}_k^p\setminus\mathcal{I}} \gamma_I(N^{(I)})\big(1 - \gamma_I^\circ(Q)\big),
\end{aligned}
$$

where the last inequality is an instance of the Fenchel-Young inequality. But this last expression is non negative and, as a consequence of conditions **(S.2)**,**(L.2)** and **(L.3)**, can only be equal to zero if,

$$
\begin{cases}
 & \|\mathcal{P}_{\mathcal{T}_0^c}(N_0)\|_1 = 0, \\
\forall i \in [\![r]\!], & \gamma_{I_i}\big(\mathcal{P}_{\mathcal{T}_i^c}(N_i)\big) = 0, \\
\forall I \in \mathcal{G}_k^p\setminus\mathcal{I}, & \gamma_I(N^{(I)}) = 0.
\end{cases}
$$

So $\forall I \notin \mathcal{I}$, $N^{(I)} = 0$, and for all $0 \leq i \leq r$, $N_i \in \mathcal{T}_i$. Finally by (B.5), we have $\sum_{i=0}^r N_i = 0$, and by projecting this equality on $\bar{\mathcal{T}}_i$ we get $N_{0,i} + N_i = 0$ with $N_{0,i} \coloneqq \mathcal{P}_{\bar{\mathcal{T}}_i}(N_0) \in \mathcal{T}_0$ and $N_i \in \mathcal{T}_i$. But, by **(T)**, $\mathcal{T}_0 \cap \mathcal{T}_i = \{0\}$, i.e. the two spaces are in direct sum, in which case the fact that $N_{0,i} + N_i = 0$ implies $N_{0,i} = 0$ and $N_i = 0$. We clearly have $N_0 = \mathcal{P}_{\bar{\mathcal{T}}_{00}}(N_0) + \sum_{i=1}^r N_{i,0} = 0$, since $\mathcal{P}_{\bar{\mathcal{T}}_{00}}(N_0) = 0$ by projection of $\sum_{i=0}^r N_i = 0$ on $\bar{\mathcal{T}}_{00}$. And so finally, for all $0 \leq i \leq r$, $N_i = 0$, which shows that the solution is necessarily unique. □

Proof of Lemma 10

Claim 3. *(Bounds on ζ) Let us consider the elements of Definition 13. Given the definitions of k_0 and $\overline{\tau}$, we have*

(1) $\zeta_{i\to 0} \leq \sqrt{\frac{2\bar{\tau}}{k}}$

(2) $\zeta_{0\to i} \leq k_0$

(3) $\zeta'_{i\to 0} \leq \sqrt{\frac{2\bar{\tau}}{k}}$

(4) $\zeta'_{0\to i} \leq 2k_0\sqrt{\frac{k_0\bar{\tau}}{k}}$

Proof. (1) Let M be any matrix in $\mathcal{T}_i$ such that $\|M\|_{\text{op}} \leq 1$. We know that $\exists v$ with $\text{Supp}(v) \subset I_i$ such that $M = u^i v^\top + v u^{i^\top}$. The condition $\|M\|_{\text{op}} \leq 1$ imposes in particular $|u^{i^\top} M v / \|v\|| \leq 1$ which becomes $\|u^i\|^2 \|v\| \leq 1 - (u^{i^\top} v)^2 / \|v\|$. Hence $\|v\| \leq 1$, and

$$\begin{aligned}
\|M\|_\infty &= \|u^i v^\top + v u^{i^\top}\|_\infty \\
&\leq \max_{k,l} \left[|u^i_k||v_l| + |u^i_l||v_k| \right] \\
&\leq \|u^i\|_\infty \max_{k,l} \left[|v_l| + |v_k| \right] \leq \|u^i\|_\infty \sqrt{2} \sqrt{v_l^2 + v_k^2} \leq \sqrt{\frac{2\bar{\tau}}{k}},
\end{aligned}$$

since $\|u^i\|_\infty^2 \leq \frac{\bar{\tau}}{k}$.

(3) Since $\|\mathcal{P}_{\mathcal{T}_0}(M)\|_\infty \leq \|M\|_\infty$, we have $\zeta'_{i\to 0} \leq \zeta_{i\to 0}$.
For the other two inequalities, let Z be any matrix in $\mathcal{T}_0$ such that $\|Z\|_\infty \leq 1$. Then we know that $\text{Supp}(Z) \subset \text{Supp}(S^*)$. Let us introduce variables δ such that $\delta_{ij} = 1$ if $S^*_{ij} \neq 0$ and $\delta_{ij} = 0$ otherwise. We notice that, for any $v \in \mathbb{R}^p$,

$$\begin{aligned}
\|Zv\|_2 &= \max_{w : \|w\|_2 \leq 1} |w^\top Z v| \\
&= \max_{w : \|w\|_2 \leq 1} \sum_{i,j} |v_i||w_j||Z_{ij}| \\
&\leq \|Z\|_\infty \max_{w : \|w\|_2 \leq 1} \sum_{i,j} |v_i||w_j| \delta_{ij} \\
&\leq \|Z\|_\infty \max_{w : \|w\|_2 \leq 1} \sqrt{\sum_{i,j} \delta_{ij}} \sqrt{\sum_{i,j} v_i^2 w_j^2 \delta_{ij}} \leq \|Z\|_\infty \sqrt{\|Z\|_0} \|v\|_2, \qquad \text{(B.8)}
\end{aligned}$$

where the second inequality uses Cauchy-Schwarz and the last inequality uses the fact that $\sum_{i,j} \delta_{ij} = \|Z\|_0 \leq k_0^2$ and the fact that $|\delta_{ij}| \leq 1$.
It follows immediately from (B.8) that

$$\|Z\|_{\text{op}} \leq \|Z\|_\infty \sqrt{\|Z\|_0}. \qquad \text{(B.9)}$$

Inequality (2) follows from (B.9) and the fact that $\|Z\|_0^2$.

To prove (4), note that since $\mathcal{P}_{\mathcal{T}_i}(Z) = u^i u^{i^\top} Z - u^i u^{i^\top} Z u^i u^{i^\top} + Z u^i u^{i^\top}$,

$$\|\mathcal{P}_{\mathcal{T}_i}(Z)\|_{\text{op}} = \|u^i u^{i^\top} Z (I - u^i u^{i^\top})\|_{\text{op}} + \|Z u^i u^{i^\top}\|_{\text{op}} \leq 2 \|Z u^i\|_2.$$

But then using the same derivation as the one leading to (B.8), we have

$$\begin{aligned}
\|Z u^i\|_2 &\leq \|Z\|_\infty \max_{w : \|w\|_2 \leq 1} \sqrt{\sum_{j,j'} \delta_{jj'}} \sqrt{\sum_{j,j'} {u^i_j}^2 w^2_{j'} \delta_{jj'}} \\
&\leq \|Z\|_\infty k_0 \|u^i\|_\infty \sqrt{\sum_{j'} w^2_{j'} \big(\sum_j \delta_{jj'} \big)} \leq 2 \|Z\|_\infty k_0 \sqrt{\frac{k_0 \bar{\tau}}{k}}.
\end{aligned}$$

□

Proof of Lemma 11

Claim 4 (Transversality condition)**.** *Let* $\alpha := k_0\sqrt{\frac{2\bar{\tau}}{k}}$*. If* $\alpha < 1$*, then, for all* $i \in [\![r]\!]$*,* $\mathcal{T}_0 \cap \mathcal{T}_i = \{0\}$*.*

Proof. Let $M \in \mathcal{T}_0 \cap \mathcal{T}_i$, then by definition of $\zeta_{0\to i}$ and $\zeta_{i\to 0}$ we have

$$\|M\|_\infty = \|\mathcal{P}_{\mathcal{T}_0} \circ \mathcal{P}_{\mathcal{T}_i}(M)\|_\infty \le \zeta_{i\to 0}\,\zeta_{0\to i}\|M\|_\infty.$$

Hence, if $\zeta_{i\to 0}\,\zeta_{0\to i} < 1$ the only possible solution is $M = 0$. But given the upper bounds on $\zeta_{i\to 0}$ and $\zeta_{0\to i}$ established in Lemma 10 we get the result as soon as $\sqrt{\frac{2\bar{\tau}}{k}}k_0 < 1$. □

B.1 Technical lemmas from the proof of Theorem 5

Proof of Lemma 12

Claim 5. *Let* $A := \begin{bmatrix} I & P_{\mathcal{T}_0} \\ P_{\mathcal{T}_i} & I \end{bmatrix}$*. Then, with Definition 13, if* $(1 - \zeta_{0\to i}\,\zeta_{i\to 0}) > 0$*, then* A *is invertible and its inverse is*

$$B := \begin{bmatrix} I & -P_{\mathcal{T}_0} \\ -P_{\mathcal{T}_i} & I \end{bmatrix}\begin{bmatrix} (I - P_{\mathcal{T}_0}P_{\mathcal{T}_i})^{-1} & 0 \\ 0 & (I - P_{\mathcal{T}_i}P_{\mathcal{T}_0})^{-1} \end{bmatrix}.$$

Proof. Clearly, $AB = I$. We need to show that $(I - P_{\mathcal{T}_0}P_{\mathcal{T}_i})$ and $(I - P_{\mathcal{T}_i}P_{\mathcal{T}_0})$ are invertible. Let x be any matrix in $\mathbb{R}^{p\times p}$. From Definition 13, we have

$$\begin{aligned}\|(I - P_{\mathcal{T}_0}P_{\mathcal{T}_i})x\|_\infty &\ge \|x\|_\infty - \|P_{\mathcal{T}_0}P_{\mathcal{T}_i}x\|_\infty \\ &\ge \|x\|_\infty - \zeta_{i\to 0}\|P_{\mathcal{T}_i}x\|_{\mathrm{op}} \ge \|x\|_\infty - \zeta_{i\to 0}\,\zeta_{0\to i}\|x\|_\infty.\end{aligned}$$

Hence, if $x \neq 0$, $\|(I - P_{\mathcal{T}_0}P_{\mathcal{T}_i})x\|_\infty \ge (1-\alpha)\|x\|_\infty > 0$ which shows that $(I - P_{\mathcal{T}_0}P_{\mathcal{T}_i})$ is invertible. Moreover if we let $x = (I - P_{\mathcal{T}_0}P_{\mathcal{T}_i})^{-1}v$ in this inequality, we get the last inequality at the end of the theorem. The case of $I - P_{\mathcal{T}_i}P_{\mathcal{T}_0}$ is exactly symmetric. □

Proof of Lemma 13

Claim 6. *(Bounds on* $\|P_{\mathcal{T}_0^c}q\|_\infty$ *and* $\|P_{\mathcal{T}_i^c}q\|_{\mathrm{op}}$*)*

$$\begin{aligned}\|P_{\mathcal{T}_0^c}q\|_\infty &\le \max_{i\in[\![r]\!]}\|q_i^*\|_\infty + \zeta_{i\to 0}\|\varepsilon_i\|_{\mathrm{op}}, \\ \|P_{\mathcal{T}_i^c}q\|_{\mathrm{op}} &\le \|q_0^*\|_{\mathrm{op}} + \zeta_{0\to i}\|\varepsilon_0\|_\infty\end{aligned}$$

Proof. By Equation (6.15),

$$\begin{aligned}\|P_{\mathcal{T}_0^c}q\|_\infty &= \Big\|\sum_{i=1}^r P_{\mathcal{T}_0^c}q_i^* + P_{\mathcal{T}_0^c}\varepsilon_i\Big\|_\infty \\ &\le \max_{i\in[\![r]\!]}\|P_{\mathcal{T}_0^c}q_i^* + P_{\mathcal{T}_0^c}\varepsilon_i\|_\infty \\ &\le \max_{i\in[\![r]\!]}\|q_i^* + \varepsilon_i\|_\infty \le \max_{i\in[\![r]\!]}\big(\|q_i^*\|_\infty + \|\varepsilon_i\|_\infty\big) \le \max_{i\in[\![r]\!]}\big(\|q_i^*\|_\infty + \zeta_{i\to 0}\|\varepsilon_i\|_{\mathrm{op}}\big),\end{aligned}$$

where the first inequality is due to the fact that for each $i \in \{1, \dots, r\}$, $P_{\mathcal{T}_0^c} q_i^* + P_{\mathcal{T}_0^c} \varepsilon_i$ has its support in $I_i \times I_i$ and I_i are disjoint. The second inequality comes from the fact that for any matrix A, $\|P_{\mathcal{T}_0^c} A\|_\infty = \max_{i,j \notin \mathrm{supp}(S^*)} |A_{ij}| \le \|A\|_\infty$.
For $\|P_{\mathcal{T}_i^c} q\|_{\mathrm{op}}$, we have

$$\|P_{\mathcal{T}_i^c} q\|_{\mathrm{op}} \le \|q\|_{\mathrm{op}} \le \|q_0^* + \varepsilon_0\|_{\mathrm{op}} \le \|q_0^*\|_{\mathrm{op}} + \|\varepsilon_0\|_{\mathrm{op}} \le \|q_0^*\|_{\mathrm{op}} + \zeta_{0\to i} \|\varepsilon_0\|_\infty,$$

where the first inequality is due to the fact that for any matrix Z,

$$\|P_{\mathcal{T}_i^c} Z\|_{\mathrm{op}} = \|(I - u^i u^{i\top}) Z (I - u^i u^{i\top})\|_{\mathrm{op}} \le \|Z\|_{\mathrm{op}}.$$

□

Proof of Lemma 14

Claim 7. *(Bounds on ε_i) If $\zeta_{0\to i}\, \zeta_{i\to 0} \le \alpha < 1$, and $(\varepsilon_i)_{i \in [\![r]\!]}$ be defined as in the previous lemma, then*

$$\|\varepsilon_0\|_\infty \le \tfrac{1}{1-\alpha}\Big(\frac{\bar\tau}{k} + \zeta'_{i\to 0} 2\gamma k_0\Big) \qquad \text{and} \qquad \|\varepsilon_i\|_{\mathrm{op}} \le \tfrac{1}{1-\alpha}\Big(2\gamma k_0 + \zeta'_{0\to i}\frac{\bar\tau}{k}\Big).$$

Proof. By Lemma 12, we have

$$\begin{bmatrix} \varepsilon_{0,i} \\ \varepsilon_i \end{bmatrix} = \begin{bmatrix} I & -P_{\mathcal{T}_0} \\ -P_{\mathcal{T}_i} & I \end{bmatrix} \begin{bmatrix} (I - P_{\mathcal{T}_0} P_{\mathcal{T}_i})^{-1} & 0 \\ 0 & (I - P_{\mathcal{T}_i} P_{\mathcal{T}_0})^{-1} \end{bmatrix} \begin{bmatrix} \eta_0 \\ \eta_i \end{bmatrix} \tag{B.10}$$

So, if, for $i \in [\![r]\!]$, we let $\tilde\eta_{0,i} := (I - P_{\mathcal{T}_0} P_{\mathcal{T}_i})^{-1} \eta_0$ and $\tilde\eta_i := (I - P_{\mathcal{T}_i} P_{\mathcal{T}_0})^{-1} \eta_i$, then $\varepsilon_0, \dots, \varepsilon_r$ are uniquely defined by

$$\begin{cases} \varepsilon_0 = \sum_{i=1}^r \varepsilon_{0,i} \quad \text{where } \varepsilon_{0,i} = \tilde\eta_{0,i} - P_{\mathcal{T}_i} \tilde\eta_i, \\ \varepsilon_i = \tilde\eta_i - P_{\mathcal{T}_0} \tilde\eta_{0,i} \quad \text{for } i \in [\![r]\!]. \end{cases}$$

In the rest of the proof, we use the fact that $\zeta_{0\to i}\, \zeta_{i\to 0} \le \alpha$. First, using the inequalities proved in Lemma 12, we have, for $i \ge 1$,

$$\|\tilde\eta_{0,i}\|_\infty \le \tfrac{1}{1-\alpha}\|\eta_0\|_\infty \quad \text{and} \quad \|\tilde\eta_{0,i}\|_{\mathrm{op}} \le \tfrac{1}{1-\alpha}\|\eta_i\|_{\mathrm{op}}.$$

Then, we can bound $\|\varepsilon_{0,i}\|_\infty$ as follows

$$\begin{aligned} \|\varepsilon_{0,i}\|_\infty &= \|\tilde\eta_{0,i} - P_{\mathcal{T}_0}\tilde\eta_i\|_\infty \\ &\le \|\tilde\eta_{0,i}\|_\infty + \|P_{\mathcal{T}_0}\tilde\eta_i\|_\infty \le \|\tilde\eta_{0,i}\|_\infty + \zeta'_{i\to 0}\|\tilde\eta_i\|_{\mathrm{op}} \le \tfrac{1}{1-\alpha}\big(\|\eta_0\|_\infty + \zeta'_{i\to 0}\|\eta_i\|_{\mathrm{op}}\big), \end{aligned}$$

and since all $\varepsilon_{0,i}$ have disjoint supports, $\|\varepsilon_0\|_\infty \le \frac{1}{1-\alpha} \max_{i \in [\![r]\!]} \big(\|\eta_0\|_\infty + \zeta'_{i\to 0}\|\eta_i\|_{\mathrm{op}}\big)$.

On the other hand,

$$\begin{aligned} \|\varepsilon_i\|_{\mathrm{op}} &= \|\tilde\eta_i - P_{\mathcal{T}_i}\tilde\eta_{0,i}\|_{\mathrm{op}} \\ &\le \|\tilde\eta_i\|_{\mathrm{op}} + \|P_{\mathcal{T}_i}\tilde\eta_{0,i}\|_{\mathrm{op}} \le \|\tilde\eta_i\|_{\mathrm{op}} + \zeta'_{0\to i}\|\tilde\eta_0\|_\infty \le \tfrac{1}{1-\alpha}\big(\|\eta_i\|_{\mathrm{op}} + \zeta'_{0\to i}\|\eta_0\|_\infty\big). \end{aligned}$$

Finally,

$$\|\eta_0\|_\infty = \|\mathcal{P}_{\mathcal{T}_0}(u^i u^{i\top})\|_\infty \le \|u^i u^{i\top}\|_\infty \le \|u^i\|_\infty^2 \le \frac{\bar\tau}{k},$$
$$\|\eta_i\|_{\mathrm{op}} = \gamma\|\mathcal{P}_{\mathcal{T}_i}\big(\operatorname{sign}(S^*)\big)\|_{\mathrm{op}} \le 2\gamma\|\operatorname{sign}(S^*)u^i\|_2 \le 2\gamma k_0,$$

where we used the fact that $\|\mathcal{P}_{\mathcal{T}_i}(M)\|_{\mathrm{op}} \le \|M\|_{\mathrm{op}} + \|\mathcal{P}_{\mathcal{T}_i^c}(M)\|_{\mathrm{op}} \le 2\|M\|_{\mathrm{op}}$ (see the end of the proof of Lemma 13). This concludes the proof. □

Proof of Lemma 8

Claim 8 (Simplified bounds on $\|P_{\mathcal{T}_0^c}q\|_\infty$ and $\|P_{\mathcal{T}_i^c}q\|_{\mathrm{op}}$)**.** *Let $\alpha := k_0\sqrt{\frac{2\bar\tau}{k}}$. If $\alpha < 1$, for q as in Lemma 13, we have*

$$\|P_{\mathcal{T}_0^c}q\|_\infty \le \frac{\bar\tau}{k}\frac{1-\alpha+\alpha^2\sqrt{2/k_0}}{1-\alpha} + \gamma\frac{2\alpha}{1-\alpha}, \quad \|P_{\mathcal{T}_i^c}q\|_{\mathrm{op}} \le \gamma k_0\frac{1+\alpha}{1-\alpha} + \frac{\bar\tau}{k}\frac{k_0}{1-\alpha}.$$

Proof. First note that, by Lemma 10, we have $\zeta_{0\to i}\,\zeta_{i\to 0} \le \alpha$, so that the results of previous lemmas apply.

We thus start from results of Lemma 13. From definitions, we have

$$\|q_i^*\|_\infty = \|u^i u^{i\top}\|_\infty \le \max_{k,l}|u_k^i||u_l^i| \le \frac{\bar\tau}{k}.$$

and from Lemma 10, we have $\|q_0^*\|_{\mathrm{op}} = \|\gamma\operatorname{sign}(S^*)\|_{\mathrm{op}} \le \gamma\zeta_{0\to i} \le \gamma k_0$.

Then, applying results from Lemma 14, we get

$$\begin{aligned}
\|P_{\mathcal{T}_0^c}q\|_\infty &\le \frac{\bar\tau}{k} + \frac{\zeta_{i\to 0}}{1-\alpha}\big(2\gamma k_0 + \zeta'_{0\to i}\frac{\bar\tau}{k}\big)\\
&= \frac{\bar\tau}{k}\left(1 + \frac{\zeta'_{0\to i}\zeta_{i\to 0}}{1-\alpha}\right) + \gamma\frac{2k_0\zeta_{i\to 0}}{1-\alpha},\\
\|P_{\mathcal{T}_i^c}q\|_{\mathrm{op}} &\le \gamma k_0 + \frac{\zeta_{0\to i}}{1-\alpha}\big(\frac{\bar\tau}{k} + \zeta'_{i\to 0}2\gamma k_0\big)\\
&= \gamma k_0\left(1 + \frac{2\zeta_{0\to i}\,\zeta'_{i\to 0}}{1-\alpha}\right) + \frac{\bar\tau}{k}\frac{\zeta_{0\to i}}{1-\alpha},
\end{aligned}$$

and then, using agin bounds on ζ from Lemma 10,

$$\|P_{\mathcal{T}_0^c}q\|_\infty \le \frac{\bar\tau}{k}\left(1 + \frac{\alpha^2\sqrt{2/k_0}}{1-\alpha}\right) + \gamma\frac{2\alpha}{1-\alpha},$$
$$\|P_{\mathcal{T}_i^c}q\|_{\mathrm{op}} \le \gamma k_0\left(1 + \frac{2\alpha}{1-\alpha}\right) + \frac{\bar\tau}{k}\frac{k_0}{1-\alpha}.$$

□

Proof of Lemma 16

Claim 9. *Let $\alpha := k_0\sqrt{\frac{2\bar\tau}{k}}$. If $\alpha + \frac{\alpha^2}{2k_0} < \frac{1}{3}$, then the interval $\Gamma := \big[\frac{\bar\tau}{k}\frac{1}{1-3\alpha}, \frac{1}{k_0}\frac{1-k_0\bar\tau/k}{1+\alpha}\big)$ is not empty, and for any $\gamma \in \Gamma$, the dual matrix q defined in Lemma 13 satisfies conditions **(S.2)** and **(L.2)**.*

Proof. Given the inequalities of the previous lemma, a sufficient condition for the inequality $\|P_{\mathcal{T}_0^c} q\|_\infty < \gamma$ to hold is if

$$\frac{\bar\tau}{k}\frac{1-\alpha+\alpha^2\sqrt{2/k_0}}{1-\alpha} < \gamma\Big(1-\frac{2\alpha}{1-\alpha}\Big).$$

Note that $\alpha\sqrt{\frac{2}{k_0}} \le \frac{\sqrt{2}}{3} < 1$. As a consequence the previous inequality is implied by the simpler

$$\frac{\bar\tau}{k}\frac{1}{1-\alpha} < \gamma\Big(1-\frac{2\alpha}{1-\alpha}\Big).$$

Clearly, we have $1-3\alpha > 0$, so that multiplying the last inequality by $\frac{1-\alpha}{1-3\alpha}$, the last inequality is equivalent to

$$\gamma > \frac{\bar\tau}{k}\frac{1}{1-3\alpha}. \tag{B.11}$$

Similarly, the condition $\big\|P_{\mathcal{T}_i^c} q\big\|_{\mathrm{op}} < 1$ is satisfied if

$$\gamma k_0\frac{1+\alpha}{1-\alpha} < 1 - \frac{\bar\tau}{k}\frac{k_0}{1-\alpha}, \quad \text{or equivalently} \quad \gamma < \frac{1}{k_0}\frac{1-k_0\bar\tau/k}{1+\alpha}. \tag{B.12}$$

Finally combining (B.12) and (B.11), we obtain the sufficient condition

$$\frac{\bar\tau}{k}\frac{1}{1-3\alpha} \le \gamma < \frac{1}{k_0}\frac{1-k_0\bar\tau/k}{1+\alpha}.$$

For $k_0 \ge 1$, this interval is non empty if and only if $2\bar\tau x_0(1-\alpha)/(1-3\alpha) < 1$, with $x_0 := k_0/k$. But $2\bar\tau x_0 = \frac{\alpha^2}{k_0}$, and $3\alpha + \frac{3\alpha^2}{2k_0} < 1$ implies that $(1-3\alpha)^{-1} < \frac{2k_0}{3\alpha^2}$. So that $2\bar\tau x_0(1-\alpha)/(1-3\alpha) \le \frac{2}{3}(1-\alpha) < 1$, which shows the desired result. □

The final step of the proof of Theorem 5 is to prove Proposition 7, which is more involved. The next appendix is devoted to its proof.

B.2 Proof of Proposition 7

Let $m := |\{i \mid I_i \cap J \neq \varnothing\}|$ denote the number of blocks of the support that are intersecting J. Let $k_i := |I_i \cap J|$. We assume here w.l.o.g. that, for the set J we consider, $\{i \mid I_i \cap J \neq \varnothing\} = [\![m]\!]$ and that $k_1 \ge k_2 \ge \ldots \ge k_m$. In the rest of the proof we will let $x_0 := \frac{k_0}{k}$ and $x_i := \frac{k_i}{k}$. We will also write $\tilde I_i := (I_1 \cup \ldots \cup I_{i-1})^c$.

A A recursive decomposition of each submatrix Q_{JJ}

We consider a recursive decomposition of this matrix in four blocks

$$Q_{JJ} = \begin{bmatrix} Q_{J\cap I_1, J\cap I_1} & Q_{J\cap I_1, J\cap I_1^c} \\ Q_{J\cap I_1^c, J\cap I_1} & Q_{J\cap I_1^c, J\cap I_1^c} \end{bmatrix},$$

then, we redecompose the lower right block as follows

$$Q_{J\cap I_1^c, J\cap I_1^c} = \begin{bmatrix} Q_{J\cap I_2, J\cap I_2} & Q_{J\cap I_2, J\cap (I_1\cup I_2)^c} \\ Q_{J\cap (I_1\cup I_2)^c, J\cap I_2} & Q_{J\cap (I_1\cup I_2)^c, J\cap (I_1\cup I_2)^c} \end{bmatrix}.$$

etc, see Figure B.1.

In particular, we will construct upper bounds $\lambda^{(i)}$ and $\tilde{\lambda}^{(i)}$ such that

$$\lambda_{\max}(Q_{J\cap I_i, J\cap I_i}) \le \lambda^{(i)} \quad \text{and} \quad \lambda_{\max}(Q_{J\cap \tilde{I}_i, J\cap \tilde{I}_i}) \le \tilde{\lambda}^{(i)}.$$

To construct an upper bound of $\lambda_{\max}(Q_{J\cap I_i, J\cap I_i})$ it is necessary to take into account the structure of $Q_{J\cap I_i, J\cap I_i}$ and in particular the fact that, for the operator norm, the component of Q_{I_i,I_i} on $\mathcal{T}_i$ will contribute most strongly to the largest eigenvalue of $Q_{J\cap I_i, J\cap I_i}$, especially when the overlap $J \cap I_i$ is large.

Let $P^{\perp}_{u^i} := I - u^i (u^i)^\top$ for short. Note that since $\mathcal{P}_{\mathcal{T}_i^c}(Q) = P^{\perp}_{u^i} Q_{I_i I_i} P^{\perp}_{u^i}$ and $P^{\perp}_{u^i}$ is idempotent, we have $\mathcal{P}_{\mathcal{T}_i^c}(Q) = P^{\perp}_{u^i} \mathcal{P}_{\mathcal{T}_i^c}(Q) P^{\perp}_{u^i}$.

With these notations and remarks, we have

$$Q_{I_i\cap J, I_i\cap J} = u^i_J {u^i_J}^\top + [P^{\perp}_{u^i} \mathcal{P}_{\mathcal{T}_i^c}(Q) P^{\perp}_{u^i}]_{JJ}. \tag{B.13}$$

Let $\check{u}^i_J = \frac{u^i_J}{\|u^i_J\|}$ and $[\check{u}^i_J, U^i_J]$ be an orthormal basis matrix, obtained from $\check{u}^i_J$ by Gram-Schmidt orthonormalization. Since the matrix $[\check{u}^i_J, U^i_J]^\top Q_{I_i\cap J, I_i\cap J} [\check{u}^i_J, U^i_J]^\top$, has the same largest eigenvalue as $Q_{I_i\cap J, I_i\cap J}$, we consider the four blocks of the former matrix, bound separately the operator norms of each of the blocks and then construct the upper bound $\lambda^{(i)}$ from these.

Indeed, using (B.13) and the fact that $\mathrm{Supp}(\check{u}^i_J) \subset J$, we have

$$|(\check{u}^i_J)^\top Q_{I_i\cap J, I_i\cap J} \check{u}^i_J| \le \|u^i_J\|_2^2 + \|P^{\perp}_{u^i} \check{u}^i_J\|_2 \|\mathcal{P}_{\mathcal{T}_i^c}(Q)\|_{\mathrm{op}} \|P^{\perp}_{u^i} \check{u}^i_J\|_2, \tag{B.14}$$

$$\|(U^i_J)^\top Q_{I_i\cap J, I_i\cap J} \check{u}^i_J\|_2 \le \|\mathcal{P}_{\mathcal{T}_i^c}(Q)\|_{\mathrm{op}} \|P^{\perp}_{u^i} \check{u}^i_J\|_2, \tag{B.15}$$

$$\|(U^i_J)^\top Q_{I_i\cap J, I_i\cap J} U^i_J\|_{\mathrm{op}} \le \|\mathcal{P}_{\mathcal{T}_i^c}(Q)\|_{\mathrm{op}}. \tag{B.16}$$

We will discuss in the next section how we can leverage these bounds to obtain a bound $\lambda^{(i)}$. We first discuss how the various terms appearing in the right hand sides can be bounded based on the assumption and previous results.

As a consequence of the assumed inequalities (6.14) on u^i, we have $x_i \underline{\tau} \le \|u^i_J\|_2^2 \le x_i \overline{\tau}$, and, using the same formula for $u^i_{J\setminus I_i}$ and combining,

$$\|u^i_J\|^2 \le \min(x_i \overline{\tau}, 1 - \underline{\tau} + x_i \underline{\tau}). \tag{B.17}$$

Note that we have $x_i\overline{\tau} < 1 - \underline{\tau} + x_i\underline{\tau}$ if and only if $2k_i < k$.

We have $\|P^{\perp}_{u^i}\check{u}^i_J\|_2^2 = \|\check{u}^i_J - u^i u^{i\top}\check{u}^i_J\|_2^2 = 1 - (u^{i\top}\check{u}^i_J)^2 = 1 - \|u^i_J\|_2^2$, so that

$$\|P^{\perp}_{u^i}\check{u}^i_J\|_2^2 \le \min\big(1 - \underline{\tau}x_i, \overline{\tau}(1 - x_i)\big). \tag{B.18}$$

Again, which of the two elements in the upper bound is smaller depends on whether $x_i \le \frac{1}{2}$.

As in the statement of the theorem, we set $\gamma := \frac{\mu\overline{\tau}}{k}$ with $\mu := (1 - 3\alpha)^{-1}$ and, as before, $\alpha = k_0\sqrt{\frac{2\overline{\tau}}{k}}$.

Using this value of γ in the upper bound obtained in Lemma 8, we have

$$\|\mathcal{P}_{\mathcal{T}_i^c}(Q)\|_{\mathrm{op}} \le r := \frac{\mu\overline{\tau}}{k}k_0\frac{1+\alpha}{1-\alpha} + \frac{\bar{\tau}}{k}\frac{k_0}{1-\alpha} = 2\mu\overline{\tau}x_0. \tag{B.19}$$

We need to upper bound also the off-diagonal blocks. For this, note that all off-diagonal blocks are in $\mathcal{T}_0$ and that, given that$\|S^*_{i\cdot}\|_0 \le k_0$, for any sets J', J'' with $|J'| = k'$ and $|J''| = k''$, it follows from (B.9) that

$$\forall Z \in \mathcal{T}_0, \quad \|Z_{J'J''}\|_{\mathrm{op}} \le \|Z\|_\infty\sqrt{\|Z\|_0} \le \|Z\|_\infty\sqrt{\min(k', k_0)\min(k'', k_0)}. \tag{B.20}$$

In particular, we have

$$\|Q_{J\cap I_i, J\cap\tilde{I}_{i+1}}\|_{\mathrm{op}} \le \gamma k\sqrt{\min(x_i, x_0)\min(\tilde{x}_i, x_0)}. \tag{B.21}$$

with $\tilde{x}_i = 1 - \sum_{j=1}^{i-1} x_j$. Letting $\check{z} := \min(x_0, 1 - x_1)$, this entails

$$\|Q_{J\cap I_i, J\cap\tilde{I}_{i+1}}\|_{\mathrm{op}} \le \gamma k_0 \qquad \text{and} \qquad \|Q_{J\cap I_1, J\cap I_1^c}\|_{\mathrm{op}} \le \gamma\sqrt{k_0 k\check{z}}. \tag{B.22}$$

B Bounding eigenvalues of different blocks within Q_{JJ}

To write concisely various bounds we introduce several notations. First, given a two-by-two matrix M of the form

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \text{with} \quad a, b, c, d \ge 0,$$

we denote its largest eigenvalue $\lambda_{\max}(a, b, c, d)$.

For $0 \le x \le 1/2$, with $\overline{\eta} := \overline{\tau} - r\underline{\tau}$, and using r defined in (B.19), we denote

$$\begin{cases} a_l(x) = \overline{\tau}x + (1 - \underline{\tau}x)r = \overline{\eta}x + r \\ b_l(x) = r = c_l(x) \\ d_l(x) = r \end{cases}$$

and we write $\lambda_l(x) = \lambda_{\max}(a_l(x), b_l(x), c_l(x), d_l(x))$. Note that $x \mapsto \lambda_l(x)$ is clearly an increasing function.

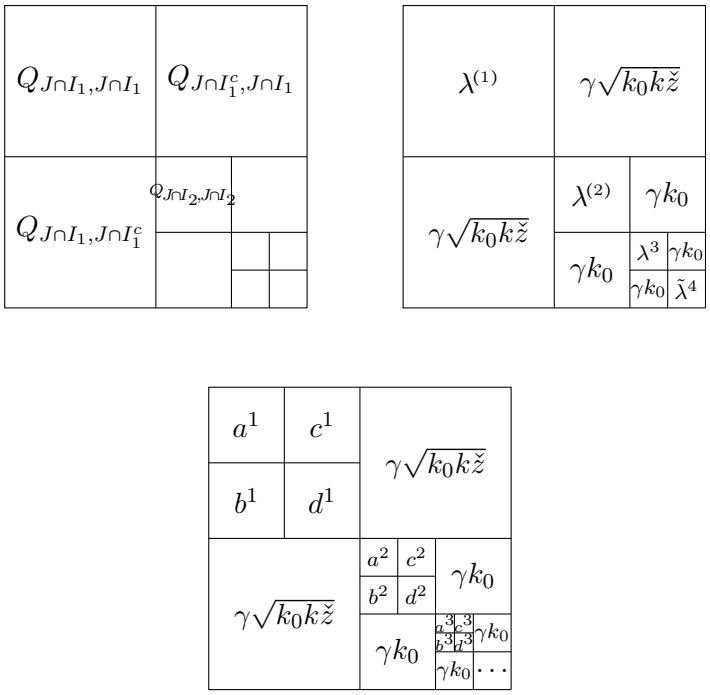


Figure B.1: Matrix blocks and corresponding upper bounds on largest singular values: (top left) Recursive partitioning of blocks of Q_{JJ} introduced in Section A (top right and bottom) Upper bounds on the operator norms of (sub)blocks introduced in inequalities (B.22), (B.23) and in Proposition 13.

Combining inequalities (B.14) to (B.19) we get that, for all $i \in [\![m]\!]$, if $x_i \in (0, \frac{1}{2}]$, we let $a^i := a_l(x_i), b^i := b_l(x_i), c^i := c_l(x_i), d^i := d_l(x_i)$, we have

$$\begin{cases} |(\check{u}^i_J)^\top Q_{I_i\cap J, I_i\cap J}\, \check{u}^i_J| & \le a^i, \\ \|(U^i_J)^\top Q_{I_i\cap J, I_i\cap J}\, \check{u}^i_J\|_2 & \le b^i = c^i, \\ \|(U^i_J)^\top Q_{I_i\cap J, I_i\cap J}\, U^i_J\|_{\mathrm{op}} & \le d^i. \end{cases} \tag{B.23}$$

(We could get a smaller value for $b^i = c^i$ based on (B.18), but this us not useful for our proof)

Symmetrically, for $1 > x > 1/2$, then if $z = 1 - x$, and with $\underline{\eta} := \underline{\tau} - r\overline{\tau}$, we define

$$\begin{cases} a_u(z) = 1 - \underline{\tau} z + \overline{\tau} z r = 1 - \underline{\eta} z \\ b_u(z) = r\sqrt{\overline{\tau} z} = c_u(z) \\ d_u(z) = r. \end{cases}$$

We will denote again $\lambda_u(z) = \lambda_{\max}(a_u(z), b_u(z), c_u(z), d_u(z))$.

Combining again inequalities (B.14) to (B.19), we get that, for $x_1 \in [\frac{1}{2}, 1)$, if $a^1 := a_l(1 - x_1), b^1 := b_l(1 - x_1), c^1 := c_l(1 - x_1), d^1 := d_l(1 - x_1)$, then the set of inequalities (B.23) holds again. Note that, since by definition $\sum_{i=1}^m x_i \le 1$, only x_1 can possibly be larger than $\frac{1}{2}$.

Proposition 13. *If for all $i \in [\![m]\!]$, we let $\lambda^{(i)} := \lambda^+_{\max}(a^i, b^i, c^i, d^i)$, then*

$$\|Q_{I_i\cap J, I_i\cap J}\|_{\mathrm{op}} \le \lambda^{(i)}.$$

Proof. The result follows from Lemma 23 and the fact that, for $a, b, c, d \ge 0$, if we have $a' \ge a, b' \ge b, c' \ge c, d' \ge d$, then $\lambda^+_{\max}(a, b, c, d) \le \lambda^+_{\max}(a', b', c', d')$. □

Proposition 14. *For $i \ge 2$,* $\|Q_{\tilde{I}_i\cap J, \tilde{I}_i\cap J}\|_{\mathrm{op}} \le \tilde{\lambda}^{(i)} := \lambda^{(i)} + \gamma k_0$.

Proof. To keep notations as simple as possible we prove the result for $\tilde{\lambda}^{(2)}$. The proof is the same for larger values of i.

$$\begin{aligned} \|Q_{J\cap I_1^c, J\cap I_1^c}\|_{\mathrm{op}} &\le \max_{2\le i\le m} \|Q_{J\cap I_i, J\cap I_i}\|_{\mathrm{op}} + \|Q_{(J\cap I_1^c \times J\cap I_1^c)\setminus \cup_{2\le i\le m} I_i\times I_i}\|_{\mathrm{op}} \\ &\le \max_{2\le i\le m} \lambda^{(i)} + \gamma k_0 \le \lambda^{(2)} + \gamma k_0, \end{aligned}$$

where the second inequality is a variant of (B.22) due to (B.20), and because we have $\lambda^{(2)} \ge \ldots \ge \lambda^{(m)}$ given that $z \mapsto \lambda_l(z)$ is non-decreasing. □

Note that by Lemma 22, we have $\lambda_l(x) \le a_l(x) + \sqrt{b_l(x)c_l(x)} \le \overline{\tau} x + 2r$ and $\lambda_u(z) \le 1 - \underline{\eta} z + r$, since $r < 1 - \underline{\eta} z$ for $z \le \frac{1}{2}$.

C Some technical lemmas to quantify eigenvalue bounds

We first derive a bound applicable to $\lambda^{(1)}$ if $x_1 \le \frac{1}{2}$ and to all $(\lambda^{(i)})_{2\le i\le m}$ since, for $i \ge 2$, $0 < x_i \le \frac{1}{2}$. First note that, $k_0 \le \frac{1}{7}\sqrt{k}$ entails that $\mu \le 7$, for $C \ge 182$, we have $x_0 \le \frac{1}{182}$, and so $r \le 2\mu\overline{\tau}x_0 \le \frac{1}{4}$ and $\gamma k_0 \le \mu\overline{\tau}x_0 \le \frac{1}{8}$.

Lemma 19. *For $0 < x \le \frac{1}{2}$, we have $\lambda_l(x) < 1 - 2\gamma k_0$.*

Proof. We show that $(1-2\gamma k_0 - a_l(x))(1-2\gamma k_0 - d_l(x)) \ge b_l(x)c_l(x)$. In the calculation, we will write $\overline{\eta} = \overline{\tau} - \underline{\tau}r$ for short.

We have $r \le \frac{1}{4}$ and $\gamma k_0 \le \frac{1}{8}$, which, given that $\overline{\tau} \ge 1$, entails that $\overline{\tau} - 2r - 2\overline{\eta}\gamma k_0 + r^2\underline{\tau} \ge \overline{\tau}(1-2\gamma k_0) - 2r \ge \frac{1}{4} > 0$.

As a consequence,

$$\begin{aligned}
& (1-2\gamma k_0 - a_l(x))(1-2\gamma k_0 - d_l(x)) - b_l(x)c_l(x) \\
= \;& (1-2\gamma k_0 - \overline{\eta}x - r)(1-2\gamma k_0 - r) - r^2 \\
= \;& (1-r-2\gamma k_0)^2 - \overline{\eta}x(1-r-2\gamma k_0) - r^2 \\
= \;& (1-r-2\gamma k_0)^2 - (\overline{\tau} - \underline{\tau}r)x(1-r) - r^2 + 2\overline{\eta}\gamma k_0 x \\
= \;& (1-r-2\gamma k_0)^2 - (\overline{\tau} - \underline{\tau}r - \overline{\tau}r + \underline{\tau}r^2)x - r^2 + 2\overline{\eta}\gamma k_0 x \\
= \;& (1-2r) - 4(1-r)\gamma k_0 + 4\gamma^2 k_0^2 - (\overline{\tau} - 2r - 2\overline{\eta}\gamma k_0 + r^2\underline{\tau})x \\
\ge \;& (1-2r) - 4(1-r)\gamma k_0 - \tfrac{1}{2}\overline{\tau} + r + \overline{\eta}\gamma k_0 - \tfrac{1}{2}r^2\underline{\tau} \\
\ge \;& \tfrac{1}{2}\underline{\tau}(1 - 4\gamma k_0 - r^2) - r \ge \tfrac{1}{2}\overline{\tau}x_0(\kappa\overline{\tau}(1 - \tfrac{1}{2} - \tfrac{1}{16}) - 4\mu) > 0,
\end{aligned}$$

The last equality and the second inequality use $\overline{\tau} + \underline{\tau} = 2$, the first inequality uses that the expression is a decreasing function of x on $(0, \frac{1}{2}]$, the penultimate inequality uses that, by assumption, the inequalities (6.14) hold, and in particular $\underline{\tau} \ge \kappa\overline{\tau}^2 x_0$ and again that $r \le \frac{1}{4}$ and $\gamma k_0 \le \frac{1}{8}$, and, the final positivity stems from the assumption, made in the statement of the theorem, that $\kappa > 16\mu$. □

Corollary 8. *We have, for all $i \ge 2$, $\tilde{\lambda}^{(i)} \le 1 - \gamma k_0$.*

Proof. Immediate from the previous result since $\tilde{\lambda}^{(i)} \le \lambda^{(i)} + \gamma k_0$ □

We now upper bound $\lambda^{(1)}$ in the case where $x_1 > \frac{1}{2}$. Indeed, in that case we have $\lambda^{(1)} = \lambda_u(1 - x_1)$ and the bound is provided by the following result.

Lemma 20. *For $0 < z < \frac{1}{2}$,*

$$\lambda_u(z) < 1 - (\underline{\eta} - \xi)z \qquad \textit{with} \quad \underline{\eta} := \underline{\tau} - \overline{\tau}r \quad \textit{and} \quad \xi := 2\overline{\tau}r^2/(\overline{\tau} - 2r).$$

Proof. First note that

$$\begin{aligned}
\underline{\eta} - \xi = \underline{\tau} - \overline{\tau}r - \frac{2\overline{\tau}r^2}{\overline{\tau} - 2r} = \frac{\underline{\tau}\overline{\tau} - \overline{\tau}^2 r - 2\underline{\tau}r}{\overline{\tau} - 2r} \;&\ge\; \frac{\overline{\tau}}{\overline{\tau} - 2r}(\underline{\tau} - \overline{\tau}r - 2r) \\
&\ge\; (\kappa - 6\mu)\overline{\tau}x_0 > 0.
\end{aligned}$$

We clearly have $\big(1-(\underline{\eta}-\xi)z\big)-a_u(z)=\xi z>0$ and $\big(1-(\underline{\eta}-\xi)z\big)-d_u(z)\geq\frac{1}{2}-r>0$. Moreover, we have

$$\begin{aligned}&\Big((1-(\underline{\eta}-\xi)z\big)-a_u(z)\Big)\Big((1-(\underline{\eta}-\xi)z)-d_u(z)\Big)-b_u(z)c_u(z)\\ =\;&\xi z\big(1-(\underline{\eta}-\xi)z-r\big)-\overline{\tau}r^2z=\xi z(1-r)-\xi z^2(\underline{\eta}-\xi)-\overline{\tau}r^2z>0\end{aligned}$$

because

$$\xi(1-r)-\xi z(\underline{\eta}-\xi)-\overline{\tau}r^2>\xi(1-r)-\xi\tfrac{\tau}{2}-\overline{\tau}r^2\geq\xi(1-r)-\xi\tfrac{\tau}{2}-(\tfrac{\overline{\tau}}{2}-r)\xi=0,$$

where the first strict inequality is obtained using $\underline{\eta}-\xi>\underline{\tau}$ and the fact that, given that $\underline{\eta}-\xi>0$, $z=\frac{1}{2}$ must minimize the expression. This proves the result by application of Lemma 21. □

D Combining bounds on eigenvalues of sublocks of Q_{JJ}

We can finally prove the claim of Proposition 7:

Claim 10. *Under the assumptions of Theorem 5, and setting C in the statement of the theorem to $C=182$, then for all $J\in\mathcal{G}_k^p\backslash\{I_i\}_{1\leq i\leq m}$, we have $\lambda_{\max}^+(Q_{JJ})<1$.*

Proof. Note first that, as discussed at the beginning of the previous section, under these assumptions, we have $r\leq\frac{1}{4}$ and $\gamma k_0\leq\frac{1}{8}$.

To prove the result, we distinguish four cases:

1st case: $0\leq x_1\leq\frac{1}{2}$. If $0\leq x_1\leq\frac{1}{2}$, then by the same argument as in Corollary 8, we have

$$\lambda_{\max}^+(Q_{JJ})\leq\tilde{\lambda}^{(1)}\leq(1-\gamma k_0)<1.$$

2nd case: $\frac{1}{4}\leq z:=1-x_1\leq\frac{1}{2}$.

If $x_1>\frac{1}{2}$, then we let $z=1-x_1$, and we can upper bound the largest eigenvalue of the upper left block in Figure B.1 by $\lambda^{(1)}=\lambda_u(z)$ and the lower right block by $\tilde{\lambda}^{(2)}=\tilde{\lambda}_l(z):=\lambda_l(z)+\gamma k_0$, given Proposition 14.

First, we consider the case $z:=1-x_1$ with $\frac{1}{4}\leq z\leq\frac{1}{2}$. In that case, we have $\lambda_{\max}^+(Q_{JJ})\leq\lambda_{\max}^+(\lambda^{(1)},\gamma k_0,\gamma k_0,\tilde{\lambda}^{(2)})$. But by Lemma 20 and Corollary 8, we have

$$\big(1-\lambda_u(z)\big)\big(1-\tilde{\lambda}_l(z)\big)-\gamma^2k_0^2\quad\geq\quad(\underline{\eta}-\xi)\frac{1}{4}\gamma k_0-\gamma^2k_0^2,$$

and $\underline{\eta}-\xi-4\gamma k_0>(\kappa-6\mu-4\mu)\overline{\tau}x_0>0$, using the same lower bound for $\underline{\eta}-\xi$ as the one established in Lemma 20.

3rd case: $x_0\leq z=1-x_1\leq\frac{1}{4}$. We have

$$\lambda_u(z)\leq a_u(z)+\frac{b_u(z)c_u(z)}{a_u(z)-d_u(z)}=1-\underline{\eta}z+\frac{r^2\overline{\tau}z}{1-\underline{\eta}z-r}\quad\text{and}\quad\tilde{\lambda}_l(z)\leq\overline{\tau}z+2r+\gamma k_0.$$

As a consequence the function f defined by

$$f(z):=\Big(\underline{\eta}z-\frac{r^2\overline{\tau}z}{1-\underline{\eta}z-r}\Big)\big(1-\overline{\tau}z-2r-\gamma k_0\big)$$

provides the lower bound $f(z) \le \big(1-\lambda_u(z)\big)\big(1-\tilde\lambda_l(z)\big)$.

We first show that this function is increasing on the interval $[x_0, \frac14]$. Indeed, given that, for $z \le \frac14$, we have $\underline\eta z + r \le \frac12$, we have

$$\begin{aligned}
f'(z) &= \Big(\underline\eta - \frac{r^2\overline\tau}{1-\underline\eta z - r} - \frac{\underline\eta r^2\overline\tau z}{(1-\underline\eta z - r)^2}\Big)(1-\overline\tau z - 2r - \gamma k_0) - \underline\eta z\overline\tau + \frac{r^2\overline\tau^2 z}{1-\underline\eta z - r}\\
&= \Big(\underline\eta - \frac{r^2\overline\tau}{1-\underline\eta z - r} - \frac{\underline\eta r^2\overline\tau z}{(1-\underline\eta z - r)^2}\Big)(1-2\overline\tau z - 2r - \gamma k_0) - \frac{\underline\eta r^2\overline\tau^2 z^2}{(1-\underline\eta z - r)^2}\\
&\ge \big(\underline\eta - 2r^2\overline\tau - \underline\eta r^2\overline\tau\big)\big(\frac{\tau}{2} - 2r - \gamma k_0\big) - \frac14\underline\eta r^2\overline\tau^2\\
&\ge \big((\kappa-2\mu) - \mu - \frac12\mu\big)\frac12\big(\kappa - 8\mu - \mu\big)\overline\tau^3 x_0^2 - \mu^2\overline\tau^4 x_0^2\\
&\ge \frac{\overline\tau^3 x_0^2}{2}\big[(\kappa-4\mu)(\kappa-9\mu) - 4\mu^2\big] > 0.
\end{aligned}$$

Therefore the minimal value of f is attained for $z = x_0$. Note that

$$\underline\eta(1-\underline\eta x_0 - r) - r^2\overline\tau = \underline\tau - r\overline\tau - \underline\eta^2 x_0 - r\underline\tau = \underline\tau - 2r - \underline\eta^2 x_0 \ge \underline\tau(1-x_0) - 2\overline\tau r$$

which entails

$$\begin{aligned}
(1-\underline\eta x_0 - r)\big[f(x_0) - \gamma^2 k_0^2\big] &\ge (1-\underline\eta x_0 - r)f(x_0) - \gamma^2 k_0^2\\
&\ge x_0(\underline\tau(1-x_0) - 2\overline\tau r)(1-\overline\tau x_0 - 2r - \mu\overline\tau x_0) - \mu^2\overline\tau^2 x_0^2\\
&\ge \overline\tau^2 x_0^2\big[(\kappa(1-x_0) - 4\mu)\tfrac34 - \mu^2\big] \ge 4\,\overline\tau^2 x_0^2\,\mu > 0,
\end{aligned}$$

since $\mu \le 7$ and since, the assumption $x_0 \le \frac{1}{182}$ entails that $\overline\tau x_0 + 2r + \mu\overline\tau x_0 \le \frac14$.

But this shows that $f(x_0) - \gamma^2 k_0^2 > 0$ and since this is a lower bound on

$$\big(1-\lambda_u(z)\big)\big(1-\tilde\lambda_l(z)\big) - \gamma^2 k_0^2$$

on the interval $x_0 \le z \le \frac14$ we again have that $\lambda^+_{\max}(Q_{JJ}) < 1$ by Lemma 21.

4th case: $0 < z = 1 - x_1 \le x_0$. When z becomes very small, the off-diagonal block $Q_{J\cap I_1, J\cap I_1^c}$ becomes a very thin vertical block. As a consequence the bound $\|Q_{J\cap I_1, J\cap I_1^c}\|_{\rm op} \le \gamma k_0$ is no longer sufficient, but using Equation (B.20) we also have that $\|Q_{J\cap I_1, J\cap I_1^c}\|_{\rm op} \le \tilde b(z)$ with $\tilde b(z) = \gamma\sqrt{k_0 k z}$. As a consequence, we have

$$\lambda^+_{\max}(Q_{JJ}) \le \lambda^+_{\max}(\lambda^{(1)}, \tilde b(z), \tilde b(z), \tilde\lambda^{(2)}),$$

with $\lambda^{(1)} = \lambda_u(z)$, $\tilde\lambda^{(2)} = \tilde\lambda_l(z)$ and $z = 1 - x_1$. Reasoning like for the 3rd case, since $f(z) - \gamma^2 k_0 k z \le \big(1-\lambda_u(z)\big)\big(1-\tilde\lambda_l(z)\big) - \tilde b(z)^2$, it is sufficient to prove that $f(z) - \gamma^2 k_0 k z > 0$. But since $0 < z \le x_0$, we simply have

$$\begin{aligned}
\frac{x_0}{z}(f(z) - \gamma^2 k_0 k z) &= x_0\Big(\underline\eta - \frac{r^2\overline\tau}{1-\underline\eta z - r}\Big)\big(1-\overline\tau z - 2r - \gamma k_0\big) - \gamma^2 k_0^2\\
&\ge x_0\Big(\underline\eta - \frac{r^2\overline\tau}{1-\underline\eta x_0 - r}\Big)\big(1-\overline\tau x_0 - 2r - \gamma k_0\big) - \gamma^2 k_0^2\\
&= f(x_0) - \gamma^2 k_0^2 > 0,
\end{aligned}$$

where the last inequality was proven in the analysis of the 3rd case. This shows that for all $0 < z \le x_0$, we have

$$0 < f(z) - \gamma^2 k_0 k z \le \big(1 - \lambda_u(z)\big)\big(1 - \tilde{\lambda}_l(z)\big) - \tilde{b}(z)^2,$$

so that $\lambda^+_{\max}(Q_{JJ}) < 1$ by Lemma 21. □

B.3 Lemmas to control eigenvalues

In this section, we establish general bounds on eigenvalues of two-by-two matrices and of matrices that can be partitioned in two-by-two blocks.

Consider a two-by-two matrix M of the form

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \text{with} \quad a, b, c, d \ge 0.$$

We denote its largest eigenvalue $\lambda_{\max}$.

Since $\lambda_{\max} + \lambda_{\min} = a + d$ and $\lambda_{\max}\lambda_{\min} = ad - bc$, the eigenvalues are the roots of $x^2 - (a + d)x + ad - bc$, and by the quadratic formula, we have

$$2\lambda_{\max} = a + d + \sqrt{(a-d)^2 + 4bc}.$$

Given that $a, b, c, d \ge 0$, we must have $(a-d)^2 + 4bc > 0$ and the eigenvalues of M are real.

Lemma 21. $\lambda_{\max} < \nu \quad \Leftrightarrow \quad \max(a, d) < \nu \quad \textit{and} \quad bc < (\nu - a)(\nu - d).$

Proof. Indeed we clearly have $\lambda_{\max} < \nu \Rightarrow \max(a, d) < \nu$. And conversely, if $\max(a, d) < \nu$, using the quadratic formula, we have

$$\begin{aligned} \lambda_{\max} < \nu \quad &\Leftrightarrow \quad a + d + \sqrt{(a-d)^2 + 4bc} < 2\nu \\ &\Leftrightarrow \quad (a-d)^2 + 4bc < (2\nu - (a+d))^2 \\ &\Leftrightarrow \quad -2ad + 4bc < 4\nu^2 - 4\nu(a+d) + 2ad \\ &\Leftrightarrow \quad bc < \nu^2 - 2(a+d) + ad. \end{aligned}$$

where the second equivalence uses that $\max(a, d) < \nu \;\Rightarrow\; 2\nu - a - d > 0$. □

Lemma 22. *If $a > d$, we have $\lambda_{\max} \le a + \dfrac{bc}{a-d}$ and $\lambda_{\max} \le a + \sqrt{bc}$.*

Proof. Indeed, if $a > d$,

$$\sqrt{(a-d)^2 + 4bc} \le (a-d)\sqrt{1 + \frac{4bc}{(a-d)^2}} \le (a-d)\Big(1 + \frac{2bc}{(a-d)^2}\Big) \le a - d + \frac{2bc}{a-d}.$$

So that by the quadratic formula, we have

$$2\lambda_{\max} = a + d + \sqrt{(a-d)^2 + 4bc} \le a + d + a - d + \frac{2bc}{a-d} = 2a + \frac{2bc}{a-d}.$$

To prove the second inequality, note that $\sqrt{(a-d)^2 + 4bc} \le a - d + 2\sqrt{bc}$ which yields the result. □

Lemma 23.

$$\lambda_{\max}\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}\right) \leq \lambda_{\max}\left(\begin{bmatrix} \lambda_{\max}(A) & \|B\|_{\mathrm{op}} \\ \|C\|_{\mathrm{op}} & \lambda_{\max}(D) \end{bmatrix}\right)$$

Proof. Since, for $y_1 = \|x_1\|$ and $y_2 = \|x_2\|$, we have

$$x_1^\top A x_1 + x_1^\top B x_2 + x_2^\top C x_1 + x_2^\top D x_2 \leq \lambda_{\max}(A)\, y_1^2 + (\|B\|_{\mathrm{op}} + \|C\|_{\mathrm{op}})\, y_1 y_2 + \lambda_{\max}(D)\, y_2^2,$$

maximizing on both sizes of the inequality under the constraint $y_1^2 + y_2^2 = 1$ yields the result. □

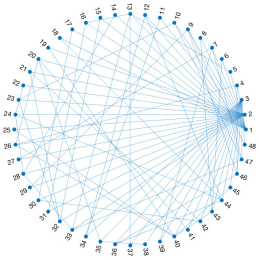
B.4 Construction of sparse precision matrices

In this appendix, we provide details on the construction of the precision matrices used in the experiments.

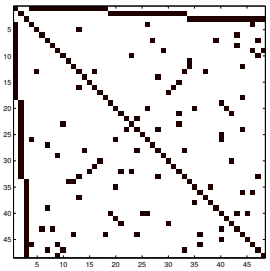
Constructing valid concentration matrices for a sparse Gaussian graphical model associated with a given graph is not completely immediate. In our synthetic experiment, we generate random concentration matrices from a model that yields sparse counterparts to Wishart matrices.

Given a graph $G = (V, E)$, where V and E are the set of vertices and edges respectively, we first build an incidence matrix $B \in \mathbb{R}^{n \times m}$ for G (where $n = |V|$ and $m = |E|$, and with $B_{i,j} = 1$ if the vertex v_i and edge e_j are incident and 0 otherwise). We then compute a sparse random matrix $\tilde{B}$ with sparsity pattern given by B, and with its nonzero coefficients drawn i.i.d. standard Gaussian. Finally, the matrix $K = \tilde{B}\tilde{B}^\top$ is a random concentration matrix with the imposed sparse structure: indeed, by construction, the non-zero pattern of K matches exactly the adjacency structure E of the graph G, and the obtained matrix K is clearly p.s.d. .

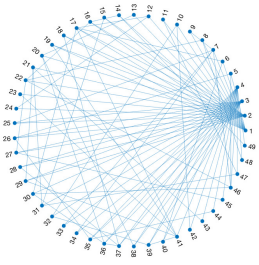
B.5 Experiments



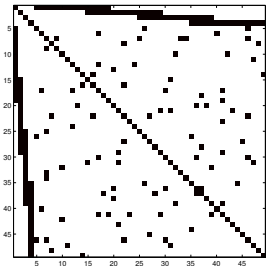
(a) *model 1*



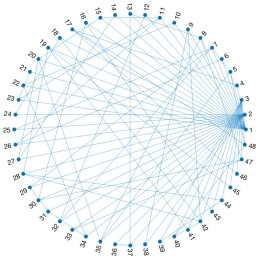
(b) structure of concentration matrix for *model 1*



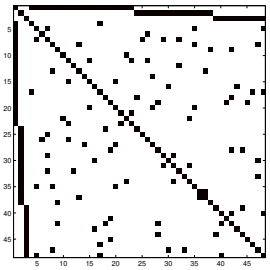
(c) *model 2*



(d) structure of concentration matrix for *model 2*



(e) *model 3*



(f) structure of concentration matrix for *model 3*

Figure B.2: Structures of graphical models for the synthetic experiments

Appendix C

Convex demixing by gauge minimization

C.1 Lemmas for the main theorem

We provide in this appendix the technical results needed to prove the main theorem.

First, given a collection of subspaces $\mathcal{T}_i$, each equipped with a norm (or symmetric coercive gauge) Ω_i, and given the projectors P_i each of the subspaces, we provide a sufficient condition on a collection of operator-norms of the P_i that are induced by the subspace norms, that guarantee that the subspaces are in direct sum.

Lemma 24. *Let $(\Omega_i)_{i=1..m}$ be a collection of norms (or symmetric coercive gauges). Let $\zeta_{ij} := \max_{u_j \in \mathcal{T}_j,\, \Omega_j(u_j)\leq 1} \Omega_i(P_i u_j)$, with P_i the projector on the subspace $\mathcal{T}_i$. Let $\alpha := \max_i \sum_{j\neq i} \zeta_{ij}$. If $\alpha < 1$, then $\forall i$, $\mathcal{T}_i \cap span\big((\mathcal{T}_j)_{j\neq i}\big) = \{0\}$.*

Proof. We reason by contradiction. Assume that there exist $(u_i)_{1\leq i\leq m}$, with $u_i \in \mathcal{T}_i$ and $\sum_{j=1}^m u_j = 0$. Then, if $\Omega_i(u_i) = \Omega(u) := \max_j \Omega_j(u_j)$, we have

$$\Omega(u) = \Omega_i(u_i) = \Omega_i(-P_i u_i) = \Omega_i\Big(\sum_{j\neq i} P_i u_j\Big) \leq \sum_{j\neq_i} \Omega_i(P_i u_j) \leq \sum_{j\neq_i} \zeta_{ij}\Omega_j(u_j) \leq \alpha\Omega(u).$$

Since $\alpha < 1$ this entails $\Omega(u) = 0$ and so $u_i = 0$ for all i. □

The following lemma generalizes a classical upper bound on $|||A^{-1}|||_\infty$ discussed in Varga (1976) for diagonally dominant matrices, where $|||\cdot|||_\infty$ is the operator ℓ_∞ norm (equal to the maximal ℓ_1-norm of all rows), not to be confused with the ℓ_∞ norm of the vectorized matrix, also known as the max-norm that we denote with $\|\cdot\|_\infty$ throughout the paper.

Lemma 25. *Let $A = (A_{ij})$ a matrix defining a linear operator from $\mathbb{R}^{r_1}\times\ldots\times\mathbb{R}^{r_m}$ to itself, with $A_{ii} = Id_{r_i}$. Consider a collection of norms ω_i each defined on $\mathbb{R}^{r_i}$ and define $\omega(x) = \max_i \omega_i(x_i)$. Define the matrix operator norm $|||A|||_{\omega,\omega} = \max_{x:\,\omega(x)\leq 1} \omega(Ax)$ and consider the quantities:*

$\zeta_{ij} = \max_{x_j:\omega_j(x_j)\leq 1} \omega_i(A_{ij}x_j)$. *Then if*

$$\alpha := \max_i \sum_{j\neq i} \zeta_{ij}$$

is such that $\alpha < 1$, *then* A *is invertible and* $|||A^{-1}|||_{\omega,\omega} < \frac{1}{1-\alpha}$.

Proof. Consider a vector x and assume that $i = \arg\max_k \omega_k(x_k)$ then

$$\begin{aligned}
(1-\alpha)\,\omega(x) = (1-\alpha)\,\omega_i(x_i) &\leq \omega_i(x_i) - \sum_{j\neq i} \zeta_{ij}\,\omega_i(x_i) \leq \omega_i(x_i) - \sum_{j\neq i} \zeta_{ij}\,\omega_j(x_j) \\
&\leq \omega_i(Id_{r_i}x_i) - \sum_{j\neq i} \omega_i(A_{ij}x_j) \leq \omega_i(A_{ii}x_i) - \omega_i\Big(\sum_{j\neq i} A_{ij}x_j\Big) \\
&\leq \omega_i\Big(\sum_{j=1}^{p} A_{ij}x_j\Big) \leq \max_i \omega_i(A_{i\cdot}x) = \omega(Ax)
\end{aligned}$$

Since this inequality is true for all x, it proves that for all x, $Ax \neq 0$ which entails that A is invertible. Furthermore,

$$(1-\alpha) \leq \inf_{x\neq 0} \frac{\omega(Ax)}{\omega(x)} = \inf_{y\neq 0} \frac{\omega(y)}{\omega(A^{-1}y)},$$

given that y is invertible, and so $\sup_{y\neq 0} \frac{\omega(A^{-1}y)}{\omega(y)} \leq \frac{1}{1-\alpha}$ which is the announced result. □

C.2 Technical results on gauges

Lemma 26. *(Partial coercivity of* ν° *on* $\mathcal{T}_x$*) Let* ν *be a gauge with an o.d.s. (cf Definition 14) and let* $\mathcal{T}_x$ *be defined as in Proposition 8. Then, for any* $x \in \mathbb{R}^d$, *the largest subspace of* $\mathcal{T}_x$ *on which* $\nu^\circ \equiv 0$ *is* $\{0\}$, *or equivalently, for any* $q \in \mathcal{T}_x\backslash\{0\}$, *either* $\nu^\circ(q) > 0$ *or* $\nu^\circ(-q) > 0$.

Proof. Before we prove the result, we discuss the equivalence of the two statements: clearly, if, for all $q \in \mathcal{T}_x$ with $q \neq 0$, either $\nu^\circ(q) > 0$ or $\nu^\circ(-q) > 0$, then ν° cannot be identically 0 on the span of q for any q which entails that there are no non trivial subspaces of $\mathcal{T}_x$ on which ν° is zero. Conversely, if there exists $q \in \mathcal{T}_x$ with $q \neq 0$, such that $\nu^\circ(q) = 0$ and $\nu^\circ(-q) = 0$, then by positive homogeneity of gauges, we must have $\nu^\circ(\lambda q)$ for all $\lambda \in \mathbb{R}$.

To then prove the result, first, note that there exists a unique maximal subspace $\mathcal{T}_0$ for the inclusion such that, $\forall q \in \mathcal{T}_0$, $\nu^\circ(q) = 0$. Indeed, assuming that $\mathcal{T}_0$ and $\mathcal{T}_0'$ are two distinct such maximal subspaces, then, by convexity, $\mathrm{span}(\mathcal{T}_0, \mathcal{T}_0')$ would be a strictly larger subspace with the same property. Now, note that, for all $q_0 \in \mathbb{R}^d$ and $q \in \mathcal{T}_0$, we must have $\nu^\circ(q_0 + q) = \nu^\circ(q_0)$. Indeed $\nu^\circ(q_0) \leq \nu^\circ(q_0 + q) + \nu^\circ(-q) = \nu^\circ(q_0 + q) \leq \nu^\circ(q_0) + \nu^\circ(q) = \nu^\circ(q_0)$.

Now we show that, for all x, $\mathcal{T}_x \cap \mathcal{T}_0 = \{0\}$. First, if x is not in the domain of ν, i.e., if $\nu(x) = \infty$, then $\partial\nu(x) = \mathbb{R}^d$ and, as a consequence, $\mathcal{T}_x = \{0\}$. We now consider x such that $\nu(x) < \infty$; note that the domain of ν must be included in $\mathcal{T}_0^\perp$: indeed, for any x, we have

$\nu(x) \geq \max_{q \in \mathcal{T}_0} \langle x, q \rangle$, which entails that, if $x \notin \mathcal{T}_0^\perp$, we must have $\nu(x) = \infty$. But then, if $x \in \mathcal{T}_0^\perp$, and if q_x is the orthogonal projection of the origin on $\partial\nu(x)$, then for all $q \in \mathcal{T}_0$, we have $\langle q, x \rangle = 0$ so that $\langle q_x + q, x \rangle = \nu(x)$ and $\nu^\circ(q_x + q) = \nu^\circ(q_x)$ by the previous point so that $q_x + q \in \partial\nu(x)$, which entails that $q \in Q_x \subset \mathcal{T}_x^\perp$, and we have $\mathcal{T}_0 \subset \mathcal{T}_x^\perp$ or equivalently $\mathcal{T}_x \subset \mathcal{T}_0^\perp$, which entails $\mathcal{T}_x \cap \mathcal{T}_0 = \{0\}$. This proves the result because, by maximality of $\mathcal{T}_0$, the largest subspace of $\mathcal{T}_x$ on which $\nu^\circ \equiv 0$ must be exactly $\mathcal{T}_x \cap \mathcal{T}_0$. □

Note that, in general, we do not have $\nu^\circ(q) > 0$ for any $q \in \mathcal{T}_x \backslash \{0\}$: clearly, this is only true if ν is symmetric.

Lemma 27 (Polar of a separable gauge)**.** *If ν is a* separable *gauge with respect to $\mathcal{M}$ and $\mathcal{M}^\perp$, then*

$$\nu^\circ(x + y) = \max(\nu^\circ(x), \nu^\circ(y)) \quad \forall (x, y) \in \mathcal{M} \times \mathcal{M}^\perp.$$

Proof. Let $(x, y) \in \mathcal{M} \times \mathcal{M}^\perp$.

$$\begin{aligned}
\nu^\circ(x + y) &= \max_{\nu(p)+\nu(q) \leq 1} \langle x, p \rangle + \langle x, q \rangle \\
&= \max_{\nu(p) \leq \eta, \nu(q) \leq \eta', \eta + \eta' \leq 1} \langle x, p \rangle + \langle x, q \rangle \\
&= \max_{\eta + \eta' \leq 1} \eta\, \nu^\circ(x) + \eta'\, \nu^\circ(y) \\
&= \max(\nu^\circ(x), \nu^\circ(y))
\end{aligned}$$

where in the first equality we use the decomposability of the gauge with respect to $\mathcal{M}$ and $\mathcal{M}^\perp$. □

Bibliography

Agarwal, A., Negahban, S., Wainwright, M. J., et al. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197.

Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294.

Argyriou, A., Evgeniou, T., and Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, 73(3):243–272.

Argyriou, A., Foygel, R., and Srebro, N. (2012). Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems 25*, pages 1466–1474.

Bach, F. (2013). Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2):145–373.

Bach, F. (2015). Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012a). Optimization with sparsity-inducing penalties. *Foundation and Trends in Machine Learning*, 1(4):1–106.

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al. (2012b). Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106.

Bach, F., Lacoste-Julien, S., and Obozinski, G. (2012c). On the equivalence between herding and conditional gradient algorithms. *arXiv preprint arXiv:1203.4523*.

Bach, F., Mairal, J., and Ponce, J. (2008). Convex sparse matrix factorizations. *arXiv preprint arXiv:0812.1869*.

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516.

Bertsekas, D. P. (1999). *Nonlinear programming*. Athena Scientific Belmont.

Bertsekas, D. P. (2015). *Convex optimization algorithms*. Athena Scientific Belmont.

Bien, J., Taylor, J., Tibshirani, R., et al. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141.

Borwein, J. M. and Lewis, A. S. (2006). *Convex analysis and nonlinear optimization*. Springer.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Bredies, K., Lorenz, D. A., and Maass, P. (2009). A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and Applications*, 42(2):173–193.

Candès, E. and Recht, B. (2013). Simple bounds for recovering low-complexity models. *Mathematical Programming*, 141(1-2):577–589.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717.

Champion, M., Picheny, V., and Vignes, M. (2018). Inferring large graphs using ℓ_1-penalized likelihood. *Statistics and Computing*, 28(4):905–921.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2010). Latent variable graphical model selection via convex optimization. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1610–1613. IEEE.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467.

Cohen, A., Dahmen, W., and DeVore, R. (2009). Compressed sensing and best k-term approximation. *Journal of the American mathematical society*, 22(1):211–231.

Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer.

Cook, S. A. (1985). A taxonomy of problems with fast parallel algorithms. *Information and control*, 64(1-3):2–22.

Dahinden, C., Parmigiani, G., Emerick, M. C., and Bühlmann, P. (2007). Penalized likelihood for sparse contingency tables with an application to full-length cdna libraries. *BMC bioinformatics*, 8(1):476.

d'Aspremont, A., Bach, F., and Ghaoui, L. E. (2008a). Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9(Jul):1269–1294.

d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008b). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66.

d'Aspremont, A., Ghaoui, L. E., Jordan, M. I., and Lanckriet, G. R. (2005). A direct formulation for sparse PCA using semidefinite programming. In *Advances in Neural Information Processing Systems*, pages 41–48.

Defazio, A. and Caetano, T. S. (2012). A convex formulation for learning scale-free networks via submodular relaxation. In *Advances in Neural Information Processing Systems*, pages 1250–1258.

Ding, C., Li, T., and Jordan, M. I. (2010). Convex and semi-nonnegative matrix factorizations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(1):45–55.

Donoho, D. L. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202.

Drton, M. and Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393.

Elad, M. and Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745.

Elad, M., Starck, J.-L., Querre, P., and Donoho, D. L. (2005). Simultaneous cartoon and texture image inpainting using morphological component analysis (mca). *Applied and Computational Harmonic Analysis*, 19(3):340–358.

Elhamifar, E., Sapiro, G., and Vidal, R. (2012). See all by looking at a few: Sparse modeling for finding representative objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1600–1607. IEEE.

Elhamifar, E. and Vidal, R. (2013). Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781.

Fadili, J. M., Peyré, G., Vaiter, S., Deledalle, C.-A., and Salmon, J. (2013). Stable recovery with analysis decomposable priors. In *Proc. SampTA'13*, pages 113–116.

Forsgren, A., Gill, P. E., and Wong, E. (2015). Primal and dual active-set methods for convex quadratic programming. *Mathematical Programming*, pages 1–40.

Forsgren, A., Gill, P. E., and Wong, E. (2016). Primal and dual active-set methods for convex quadratic programming. *Mathematical Programming*, 159(1):469–508.

Foygel, R. and Mackey, L. (2014). Corrupted sensing: Novel guarantees for separating structured signals. *IEEE Transactions on Information Theory*, 60(2):1223–1247.

Foygel, R., Srebro, N., and Salakhutdinov, R. R. (2012). Matrix reconstruction with the local max norm. In *Advances in Neural Information Processing Systems*, pages 935–943.

Franc, V., Sonnenburg, S., and Werner, T. (2011). Cutting plane methods in machine learning. In Sra, S., Nowozin, S., and Wright, S. J., editors, *Optimization for Machine Learning*. MIT Press, Cambridge, MA.

Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics (NRL)*, 3(1-2):95–110.

Friedlander, M. P., Macedo, I., and Pong, T. K. (2014). Gauge optimization and duality. *SIAM Journal on Optimization*, 24(4):1999–2022.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805.

Geiger, D., Verma, T., and Pearl, J. (1990). Identifying independence in bayesian networks. *Networks*, 20(5):507–534.

Gu, Q. and Banerjee, A. (2016). High dimensional structured superposition models. In *Advances In Neural Information Processing Systems*, pages 3691–3699.

Gu, Q., Wang, Z., and Liu, H. (2016). Low-rank and sparse structure pursuit via alternating minimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 600–609.

Haeffele, B. D. and Vidal, R. (2015). Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*.

Harchaoui, Z., Juditsky, A., and Nemirovski, A. (2015). Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1–2):75–112.

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243.

Hong, M., Wang, X., Razaviyayn, M., and Luo, Z. (2013). Iteration complexity analysis of block coordinate descent methods. *Preprint, available online*.

Hosseini, M. J. and Lee, S.-I. (2016). Learning sparse gaussian graphical models with overlapping blocks. In *Advances in Neural Information Processing Systems*, pages 3808–3816.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *ICML*.

Jaggi, M. (2013). Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435.

Jalali, A., Johnson, C. C., and Ravikumar, P. K. (2011). On learning discrete graphical models using greedy methods. In *Advances in Neural Information Processing Systems*, pages 1935–1943.

Jalali, A., Sanghavi, S., Ruan, C., and Ravikumar, P. K. (2010). A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems*, pages 964–972.

Jenatton, R., Audibert, J., and Bach, F. (2011a). Structured variable selection with sparsity-inducing norms. *JMLR*, 12:2777–2824.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. (2011b). Proximal methods for hierarchical sparse coding. *JMLR*, 12:2297–2334.

Kahn, A. B. (1962). Topological sorting of large networks. *Communications of the ACM*, 5(11):558–562.

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3):455–500.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kowalski, M., Weiss, P., Gramfort, A., and Anthoine, S. (2011). Accelerating ista with an active set strategy. In *OPT 2011: 4th International Workshop on Optimization for Machine Learning*, page 7.

Krishnamurthy, V., Ahipasaoglu, S. D., and d'Aspremont, A. (2011). A pathwise algorithm for covariance selection. In *Optimization for Machine Learning*, pages 479–494. MIT Press.

Lacoste-Julien, S. and Jaggi, M. (2015). On the global linear convergence of Frank-Wolfe optimization variants. *Advances in Neural Information Processing Systems 28*, pages 496–504.

Lacoste-Julien, S., Jaggi, M., Schmidt, M., and Pletscher, P. (2012). Block-coordinate frank-wolfe optimization for structural svms. *arXiv preprint arXiv:1207.4747*.

Lam, W. and Bacchus, F. (1993). Using causal information and local measures to learn bayesian networks. In *Uncertainty in Artificial Intelligence, 1993*, pages 243–250. Elsevier.

Larsson, T., Migdalas, A., and Patriksson, M. (2015). A generic column generation principle: derivation and convergence analysis. *Operational Research*, 15(2):163–198.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788.

Lee, S.-I., Ganapathi, V., and Koller, D. (2007). Efficient structure learning of markov networks using l_1-regularization. In *Advances in neural Information processing systems*, pages 817–824.

Levina, E., Rothman, A., and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, pages 245–263.

Li, F. and Yang, Y. (2005). Using modified lasso regression to learn large undirected graphs in a probabilistic framework. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 801.

Lin, L., Drton, M., Shojaie, A., et al. (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics*, 10(1):806–854.

Liu, G., Lin, Z., and Yu, Y. (2010). Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 663–670.

Liu, J., Musialski, P., Wonka, P., and Ye, J. (2013). Tensor completion for estimating missing values in visual data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):208–220.

Locatello, F., Khanna, R., Tschannen, M., and Jaggi, M. (2017a). A Unified Optimization View on Generalized Matching Pursuit and Frank-Wolfe. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 860–868, Fort Lauderdale, FL, USA. PMLR.

Locatello, F., Tschannen, M., Rätsch, G., and Jaggi, M. (2017b). Greedy algorithms for cone constrained optimization with convergence guarantees. In *Advances in Neural Information Processing Systems*, pages 773–784.

Mairal, J., Bach, F., Ponce, J., et al. (2014). Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3):85–283.

Maurer, A. and Pontil, M. (2012). Structured sparsity and generalization. *The Journal of Machine Learning Research*, 13(1):671–690.

McCoy, M. B., Cevher, V., Dinh, Q. T., Asaei, A., and Baldassarre, L. (2014). Convexity in source separation: Models, geometry, and algorithms. *IEEE Signal Processing Magazine*, 31(3):87–95.

McCoy, M. B. and Tropp, J. A. (2013). The achievable performance of convex demixing. *arXiv preprint arXiv:1309.7478*.

McCoy, M. B. and Tropp, J. A. (2014). Sharp recovery bounds for convex demixing, with applications. *Foundations of Computational Mathematics*, 14(3):503–567.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462.

Meng, Z., Eriksson, B., and Hero, A. (2014). Learning latent variable gaussian graphical models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1269–1277.

Moghaddam, B., Weiss, Y., and Avidan, S. (2006). Spectral bounds for sparse pca: Exact and greedy algorithms. In *Advances in Neural Information Processing Systems*, pages 915–922.

Moreau, J.-J. (1962). Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Ser. A Math.*, 255:2897–2899.

Moreau, J.-J. (1965). Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299.

Ndiaye, E., Fercoq, O., Gramfort, A., and Salmon, J. (2017). Gap safe screening rules for sparsity enforcing penalties. *J. Mach. Learn. Res*, 18(128):1–33.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Nesterov, Y. (2012). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362.

Nesterov, Y. et al. (2015). Complexity bounds for primal-dual methods minimizing the model of objective function. *Center for Operations Research and Econometrics, CORE Discussion Paper*.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Obozinski, G. and Bach, F. (2012). Convex relaxation for combinatorial penalties. *arXiv preprint arXiv:1205.1240*.

Obozinski, G. and Bach, F. (2016). A unified perspective on convex structured sparsity: Hierarchical, symmetric, submodular norms and beyond. *Hal preprint hal-01412385*.

Obozinski, G., Jacob, L., and Vert, J.-P. (2011). Group Lasso with overlaps: the Latent Group Lasso approach. *preprint HAL - inria-00628498*.

Obozinski, G., Taskar, B., and Jordan, M. (2006). Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, 2.

Obozinski, G., Taskar, B., and Jordan, M. I. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252.

Ong, F. and Lustig, M. (2016). Beyond low rank+ sparse: Multiscale low rank matrix decomposition. *IEEE journal of selected topics in signal processing*, 10(4):672–687.

Pace, R. K. and Barry, R. (1997). Sparse spatial autoregressions. *Statistics and Probability Letters*, 33(3):291 – 297.

Qin, Z., Scheinberg, K., and Goldfarb, D. (2013). Efficient block-coordinate descent algorithms for the group lasso. *Mathematical Programming Computation*, 5(2):143–169.

Rao, N., Shah, P., and Wright, S. (2015). Forward -Backward Greedy Algorithms for Atomic Norm Regularization. *IEEE Transactions on Signal Processing*, 63(21):5798–5811.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. (2009). High-dimensional graphical model selection using l1-regularized logistic regression. *Annals of Statistics*.

Ray, A., Sanghavi, S., and Shakkottai, S. (2015). Improved greedy algorithms for learning graphical models. *IEEE Transactions on Information Theory*, 61(6):3457–3468.

Richard, E., Bach, F. R., Vert, J.-P., et al. (2013). Intersecting singularities for multi-structured estimation. In *ICML (3)*, pages 1157–1165.

Richard, E., Obozinski, G. R., and Vert, J.-P. (2014). Tight convex relaxations for sparse matrix factorization. In *Advances in Neural Information Processing Systems*, pages 3284–3292.

Rockafellar, R. (1970). *Convex Analysis*. Princeton Univ. Press.

Roth, V. and Fischer, B. (2008). The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th international conference on Machine learning*, pages 848–855. ACM.

Rothman, A. J., Bickel, P. J., Levina, E., Zhu, J., et al. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Schmidt, M., Niculescu-Mizil, A., Murphy, K., et al. (2007). Learning graphical model structure using l1-regularization paths. In *AAAI*, volume 7, pages 1278–1283.

Shalev-Shwartz, S. and Tewari, A. (2011). Stochastic methods for l1-regularized loss minimization. *Journal of Machine Learning Research*, 12(Jun):1865–1892.

She, Y. and Jiang, H. (2014). Group regularized estimation under structural hierarchy. *arXiv preprint arXiv:1411.4691*.

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72.

Tan, K. M., London, P., Mohan, K., Lee, S.-I., Fazel, M., and Witten, D. M. (2014). Learning graphical models with hubs. *Journal of Machine Learning Research*, 15(1):3297–3331.

Tao, S., Sun, Y., and Boley, D. (2017). Inverse covariance estimation with structured groups. In *26th International Joint Conference on Artificial Intelligence, IJCAI 2017*.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc. B*, 58(1).

Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences.

Tomioka, R. and Suzuki, T. (2013). Convex tensor decomposition via structured Schatten norm regularization. In *Advances in Neural information Processing Systems*, pages 1331–1339.

Tropp, J. A. (2004). Just relax: Convex programming methods for subset selection and sparse approximation. *ICES report*, 404.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494.

Vaiter, S., Golbabaee, M., Fadili, J., and Peyré, G. (2015a). Model selection with low complexity priors. *Information and Inference: A Journal of the IMA*, 4(3):230–287.

Vaiter, S., Peyré, G., and Fadili, J. (2015b). Low complexity regularization of linear inverse problems. In *Sampling Theory, a Renaissance*, pages 103–153. Springer.

Varga, R. S. (1976). On diagonal dominance arguments for bounding $\|a^{-1}\|_\infty$. *Linear Algebra and its applications*, 14(3):211–217.

Vidal, R. (2011). Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68.

Vinyes, M. and Obozinski, G. (2017). Fast column generation for atomic norm regularization. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 547–556. PMLR.

Vinyes, M. and Obozinski, G. (2018). Learning the effect of latent variables in gaussian graphical models with unobserved variables. *arXiv preprint arXiv:1807.07754*.

Wimalawarne, K., Sugiyama, M., and Tomioka, R. (2014). Multitask learning meets tensor factorization: task imputation via convex optimization. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2825–2833. Curran Associates, Inc.

Wimalawarne, K., Tomioka, R., and Sugiyama, M. (2016). Theoretical and experimental analyses of tensor-based regression and classification. *Neural Computation*, 4(28):686–715.

Wolfe, P. (1970). Convergence theory in nonlinear programming. *Integer and nonlinear programming*, pages 1–36.

Wolfe, P. (1976a). Finding the nearest point in a polytope. *Mathematical Programming*, 11(1):128–149.

Wolfe, P. (1976b). Finding the nearest point in a polytope. *Mathematical Programming*, 11(1):128–149.

Wright, J., Ganesh, A., Min, K., and Ma, Y. (2013). Compressive principal component pursuit. *Information and Inference: A Journal of the IMA*, 2(1):32–68.

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34.

Xu, H., Caramanis, C., and Sanghavi, S. (2010). Robust pca via outlier pursuit. In *Advances in Neural Information Processing Systems*, pages 2496–2504.

Xu, P., Ma, J., and Gu, Q. (2017). Speeding up latent variable gaussian graphical model estimation via nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 1930–1941.

Yan, X. and Bien, J. (2015). Hierarchical sparse modeling: A choice of two regularizers. *arXiv preprint arXiv:1512.01631*.

You, C., Li, C.-G., Robinson, D. P., and Vidal, R. (2016). Oracle based active set algorithm for scalable elastic net subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3928–3937.

Yu, Y., Zhang, X., and Schuurmans, D. (2014). Generalized conditional gradient for sparse estimation. *arXiv preprint arXiv:1410.4828*.

Yuan, M. and Lin, Y. (2006a). Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67.

Yuan, M. and Lin, Y. (2006b). Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc. B*, 68:49–67.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, pages 19–35.

Yuan, X.-T. and Zhang, T. (2013). Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14(Apr):899–925.

Zhang, Y., d'Aspremont, A., and El Ghaoui, L. (2012). Sparse pca: Convex relaxations, algorithms and applications. In *Handbook on Semidefinite, Conic and Polynomial Optimization*, pages 915–940. Springer.

Zhou, Y. (2011). Structure learning of probabilistic graphical models: a comprehensive survey. *arXiv preprint arXiv:1111.6925*.

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

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286.