Sparse Modeling for Image and Vision Processing

Julien Mairal
Inria
julien.mairal@inria.fr

Francis Bach
Inria
francis.bach@inria.fr

Jean Ponce
Ecole Normale Supérieure
jean.ponce@ens.fr

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Computer Graphics and Vision

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

J. Mairal, F. Bach and J. Ponce. *Sparse Modeling for Image and Vision Processing*. Foundations and Trends® in Computer Graphics and Vision, vol. 8, no. 2-3, pp. 85–283, 2012.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-009-5

Foundations and Trends® in Computer Graphics and Vision
Volume 8, Issue 2-3, 2012

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Computer Graphics and Vision publishes survey and tutorial articles in the following topics:

- Rendering
- Shape
- Mesh simplification
- Animation
- Sensors and sensing
- Image restoration and enhancement
- Segmentation and grouping
- Feature detection and selection
- Color processing
- Texture analysis and synthesis
- Illumination and reflectance modeling
- Shape representation
- Tracking
- Calibration
- Structure from motion
- Motion estimation and registration
- Stereo matching and reconstruction
- 3D reconstruction and image-based modeling
- Learning and statistical methods
- Appearance-based matching
- Object and scene recognition
- Face detection and recognition
- Activity and gesture recognition
- Image and video retrieval
- Video analysis and event recognition
- Medical image analysis
- Robot localization and navigation

Information for Librarians

Foundations and Trends® in Computer Graphics and Vision, 2012, Volume 8, 4 issues. ISSN paper version 1572-2740. ISSN online version 1572-2759. Also available as a combined paper and online subscription.

Foundations and Trends® in
Computer Graphics and Vision
Vol. 8, No. 2-3 (2012) 85–283

DOI: 10.1561/0600000058

now
the essence of knowledge

Sparse Modeling for Image and Vision Processing

Julien Mairal
Inria [1]
julien.mairal@inria.fr

Francis Bach
Inria[2]
francis.bach@inria.fr

Jean Ponce
Ecole Normale Supérieure[3]
jean.ponce@ens.fr

[1]LEAR team, laboratoire Jean Kuntzmann, CNRS, Univ. Grenoble Alpes, France.
[2]SIERRA team, département d'informatique de l'Ecole Normale Supérieure, ENS/CNRS/Inria UMR 8548, France.
[3]WILLOW team, département d'informatique de l'Ecole Normale Supérieure, ENS/CNRS/Inria UMR 8548, France.

Contents

Abstract

In recent years, a large amount of multi-disciplinary research has been conducted on sparse models and their applications. In statistics and machine learning, the sparsity principle is used to perform model selection—that is, automatically selecting a simple model among a large collection of them. In signal processing, sparse coding consists of representing data with linear combinations of a few dictionary elements. Subsequently, the corresponding tools have been widely adopted by several scientific communities such as neuroscience, bioinformatics, or computer vision. The goal of this monograph is to offer a self-contained view of sparse modeling for visual recognition and image processing. More specifically, we focus on applications where the dictionary is learned and adapted to data, yielding a compact representation that has been successful in various contexts.

J. Mairal, F. Bach and J. Ponce. *Sparse Modeling for Image and Vision Processing*. Foundations and Trends® in Computer Graphics and Vision, vol. 8, no. 2-3, pp. 85–283, 2012.

DOI: 10.1561/0600000058.

1

A Short Introduction to Parsimony

In its most general definition, the principle of sparsity, or parsimony, consists of representing some phenomenon with as few variables as possible. It appears to be central to many research fields and is often considered to be inspired from an early doctrine formulated by the philosopher and theologian William of Ockham in the 14th century, which essentially favors simple theories over more complex ones. Of course, the link between Ockham and the tools presented in this monograph is rather thin, and more modern views seem to appear later in the beginning of the 20th century. Discussing the scientific method, Wrinch and Jeffreys [1921] introduce indeed a simplicity principle related to parsimony as follows:

> *The existence of simple laws is, then, apparently, to be regarded as a quality of nature; and accordingly we may infer that it is justifiable to prefer a simple law to a more complex one that fits our observations slightly better.*

Remarkably, Wrinch and Jeffreys [1921] further discuss statistical modeling of physical observations and relate the concept of "simplicity" to the number of learning parameters; as a matter of fact, this concept is relatively close to the contemporary view of parsimony.

Subsequently, numerous tools have been developed by statisticians to build models of physical phenomena with good predictive power. Models are usually learned from observed data, and their generalization performance is evaluated on test data. Among a collection of plausible models, the simplest one is often preferred, and the number of underlying parameters is used as a criterion to perform model selection [Mallows, 1964, 1966, Akaike, 1973, Hocking, 1976, Barron et al., 1998, Rissanen, 1978, Schwarz, 1978, Tibshirani, 1996].

In signal processing, similar problems as in statistics arise, but a different terminology is used. Observations, or data vectors, are called "signals", and data modeling appears to be a crucial step for performing various operations such as restoration, compression, or for solving inverse problems. Here also, the sparsity principle plays an important role and has been successful [Mallat and Zhang, 1993, Pati et al., 1993, Donoho and Johnstone, 1994, Cotter et al., 1999, Chen et al., 1999, Donoho, 2006, Candès et al., 2006]. Each signal is approximated by a sparse linear combination of prototypes called dictionary elements, resulting in simple and compact models.

However, statistics and signal processing remain two distinct fields with different objectives and methodology; specifically, signals often come from the same data source, *e.g.*, natural images, whereas problems considered in statistics are unrelated to each other in general. Then, a long series of works has been devoted to finding appropriate dictionaries for signal classes of interest, leading to various sorts of wavelets [Freeman and Adelson, 1991, Simoncelli et al., 1992, Donoho, 1999, Candès and Donoho, 2002, Do and Vetterli, 2005, Le Pennec and Mallat, 2005, Mallat, 2008]. Even though statistics and signal processing have devised most of the methodology of sparse modeling, the parsimony principle was also discovered independently in other fields. To some extent, it appears indeed in the work of Markowitz [1952] about portfolio selection in finance, and also in geophysics [Claerbout and Muir, 1973, Taylor et al., 1979].

In neuroscience, Olshausen and Field [1996, 1997] proposed a significantly different approach to sparse modeling than previously established practices. Whereas classical techniques in signal

processing were using fixed off-the-shelf dictionaries, the method of Olshausen and Field [1996, 1997] consists of learning it from training data. In a pioneer exploratory experiment, they demonstrated that dictionary learning could easily discover underlying structures in natural image patches; later, their approach found numerous applications in many fields, notably in image and audio processing [Lewicki, 2002, Elad and Aharon, 2006, Mairal et al., 2009, Yang et al., 2010a] and computer vision [Raina et al., 2007, Yang et al., 2009, Zeiler et al., 2011, Mairal et al., 2012, Song et al., 2012, Castrodad and Sapiro, 2012, Elhamifar et al., 2012, Pokrass et al., 2013].

The goal of this monograph is to present basic tools of sparse modeling and their applications to visual recognition and image processing. We aim at offering a self-contained view combining pluri-disciplinary methodology, practical advice, and a large review of the literature. Most of the figures in the paper are produced with the software SPAMS[1], and the corresponding Matlab code will be provided on the first author's webpage.

The monograph is organized as follows: the current introductory section is divided into several parts providing a simple historical view of sparse estimation. In Section 1.1, we start with early concepts of parsimony in statistics and information theory from the 70's and 80's. We present the use of sparse estimation within the wavelet framework in Section 1.2, which was essentially developed in the 90's. Section 1.3 introduces the era of "modern parsimony"—that is, the ℓ_1-norm and its variants, which have been heavily used during the last two decades. Section 1.4 is devoted to the dictionary learning formulation originally introduced by Olshausen and Field [1996, 1997], which is a key component of most applications presented later in this monograph. In Sections 1.5 and 1.6, we conclude our introductory tour with some theoretical aspects, such as the concept of "compressed sensing" and sparse recovery that has attracted a large attention in recent years.

With all these parsimonious tools in hand, we discuss the use of sparse coding and related sparse matrix factorization techniques for discovering the underlying structure of natural image patches in Sec-

[1] available here `http://spams-devel.gforge.inria.fr/`.

tion 2 . Even though the task here is subjective and exploratory, it is the first successful instance of dictionary learning; the insight gained from these early experiments forms the basis of concrete applications presented in subsequent sections.

Section 3 covers numerous applications of sparse models of natural image patches in image processing, such as image denoising, super-resolution, inpainting, or demosaicking. This section is concluded with other related patch-modeling approaches.

Section 4 presents recent success of sparse models for visual recognition, such as codebook learning of visual descriptors, face recognition, or more low-level tasks such as edge detection and classification of textures and digits. We conclude the section with other computer vision applications such as visual tracking and data visualization.

Section 5 is devoted to optimization algorithms. It presents in a concise way efficient algorithms for solving sparse decomposition and dictionary learning problems.

We see our monograph as a good complement of other books and monographs about sparse estimation, which offer different perspectives, such as Mallat [2008], Elad [2010] in signal and image processing, or Bach et al. [2012a] in optimization and machine learning. We also put the emphasis on the structure of natural image patches learned with dictionary learning, and thus present an alternative view to the book of Hyvärinen et al. [2009], which is focused on independent component analysis.

Notation. In this monograph, vectors are denoted by bold lower-case letters and matrices by upper-case ones. For instance, we consider in the rest of this paragraph a vector $\mathbf{x}$ in $\mathbb{R}^n$ and a matrix $\mathbf{X}$ in $\mathbb{R}^{m\times n}$. The columns of $\mathbf{X}$ are represented by indexed vectors $\mathbf{x}_1,\ldots,\mathbf{x}_n$ such that we can write $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_n]$. The i-th entry of $\mathbf{x}$ is denoted by $\mathbf{x}[i]$, and the i-th entry of the j-th column of $\mathbf{X}$ is represented by $\mathbf{X}[i,j]$. For any subset g of $\{1,\ldots,n\}$, we denote by $\mathbf{x}[g]$ the vector in $\mathbb{R}^{|g|}$ that records the entries of $\mathbf{x}$ corresponding to indices in g. For $q \geq 1$, we define the ℓ_q-norm of $\mathbf{x}$ as $\|\mathbf{x}\|_q \triangleq \left(\sum_{i=1}^n |\mathbf{x}[i]|^q\right)^{1/q}$, and the ℓ_∞-norm as $\|\mathbf{x}\|_\infty \triangleq \lim_{q\rightarrow+\infty} \|\mathbf{x}\|_q = \max_{i=1,\ldots,n} |\mathbf{x}[i]|$.

For $q < 1$, we define the ℓ_q-penalty as $\|\mathbf{x}\|_q \triangleq \sum_{i=1}^n |\mathbf{x}[i]|^q$, which, with an abuse of terminology, is often referred to as ℓ_q-norm. The ℓ_0-penalty simply counts the number of non-zero entries in a vector: $\|\mathbf{x}\|_0 \triangleq \sharp\{i \text{ s.t. } \mathbf{x}[i] \neq 0\}$. For a matrix $\mathbf{X}$, we define the Frobenius norm $\|\mathbf{X}\|_{\mathrm{F}} = \left(\sum_{i=1}^m \sum_{j=1}^n \mathbf{X}[i,j]^2\right)^{1/2}$. When dealing with a random variable X defined on a probability space, we denote its expectation by $\mathbb{E}[X]$, assuming that there is no measurability or integrability issue.

1.1 Early concepts of parsimony in statistics

A large number of statistical procedures can be formulated as maximum likelihood estimation. Given a statistical model with parameters $\boldsymbol{\theta}$, it consists of minimizing with respect to $\boldsymbol{\theta}$ an objective function representing the negative log-likelihood of observed data. Assuming for instance that we observe independent samples $\mathbf{z}_1, \ldots, \mathbf{z}_n$ of the (unknown) data distribution, we need to solve

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left[\mathcal{L}(\boldsymbol{\theta}) \triangleq -\sum_{i=1}^n \log P_{\boldsymbol{\theta}}(\mathbf{z}_i) \right], \tag{1.1}$$

where P is some probability distribution parameterized by $\boldsymbol{\theta}$.

Simple methods such as ordinary least squares can be written as (1.1). Consider for instance data points $\mathbf{z}_i$ that are pairs $(y_i, \mathbf{x}_i)$, with y_i is an observation in $\mathbb{R}$ and $\mathbf{x}_i$ is a vector in $\mathbb{R}^p$, and assume that there exists a linear relation $y_i = \mathbf{x}_i^\top \theta + \varepsilon_i$, where ε_i is an approximation error for observation i. Under a model where the ε_i's are independent and identically normally distributed with zero-mean, Eq. (1.1) is equivalent to a least square problem:[2]

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \sum_{i=1}^n \frac{1}{2} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\theta} \right)^2 .$$

To prevent overfitting and to improve the interpretability of the learned model, it was suggested in early work that a solution involving only a

[2] Note that the Gaussian noise assumption is not necessary to justify the ordinary least square formulation. It is only sufficient to interpret it as maximum likelihood estimation. In fact, as long as the conditional expectation $\mathbb{E}[y|\mathbf{x}]$ is linear, the ordinary least square estimator is statistically consistent under mild assumptions.

few model variables could be more appropriate than an exact solution of (1.1); in other words, a sparse solution involving only—let us say—k variables might be desirable in some situations. Unfortunately, such a strategy yields two difficulties: first, it is not clear a priori how to choose k; second, finding the best subset of k variables is NP-hard in general [Natarajan, 1995]. The first issue was addressed with several criterions for controlling the trade-off between the sparsity of the solution $\boldsymbol{\theta}$ and the adequacy of the fit to training data. For the second issue, approximate computational techniques have been proposed.

Mallows's C_p, AIC, and BIC. For the ordinary least squares problem, Mallows [1964, 1966] introduced the C_p-statistics, later generalized by Akaike [1973] with the Akaike information criterion (AIC), and then by Schwarz [1978] with the Bayesian information criterion (BIC). Using C_p, AIC, or BIC is equivalent to solving the penalized ℓ_0-maximum likelihood estimation problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_0, \tag{1.2}$$

where λ depends on the chosen criterion [see Hastie et al., 2009], and $\|\boldsymbol{\theta}\|_0$ is the ℓ_0-penalty. Similar formulations have also been derived by using the minimum description length (MDL) principle for model selection [Rissanen, 1978, Barron et al., 1998]. As shown by Natarajan [1995], the problem (1.2) is NP-hard, and approximate algorithms are necessary unless p is very small, *e.g.*, $p < 30$.

Forward selection and best subset selection for least squares. To obtain an approximate solution of (1.2), a classical approach is the forward selection technique, which is a greedy algorithm that solves a sequence of maximum likelihood estimation problems computed on a subset of variables. After every iteration, a new variable is added to the subset according to the chosen sparsity criterion in a greedy manner. Some variants allow backward steps—that is, a variable can possibly exit the active subset after an iteration. The algorithm is presented in more details in Section 5.1 and seems to be due to Efroymson [1960], according to Hocking [1976]. Other approaches considered in the 70's include

also the *leaps and bounds* technique of Furnival and Wilson [1974], a branch-and-bound algorithm providing the exact solution of (1.2) with exponential worst-case complexity.

1.2 Wavelets in signal processing

In signal processing, similar problems as in statistics have been studied in the context of wavelets. In a nutshell, a wavelet basis represents a set of functions $\phi_1, \phi_2, \ldots$ that are essentially dilated and shifted versions of each other. Unlike Fourier basis, wavelets have the interesting properties to be localized both in the space and frequency domains, and to be suitable to multi-resolution analysis of signals [Mallat, 1989].

The concept of parsimony is central to wavelets. When a signal f is "smooth" in a particular sense [see Mallat, 2008], it can be well approximated by a linear combination of a few wavelets. Specifically, f is close to an expansion $\sum_i \alpha_i \phi_i$ where only a few coefficients α_i are non-zero, and the resulting compact representation has effective applications in estimation and compression. The wavelet theory is well developed for continuous signals, *e.g.*, f is chosen in the Hilbert space $L^2(\mathbb{R})$, but also for discrete signals f in $\mathbb{R}^m$, making it suitable to modern digital image processing.

Since the first wavelet was introduced by Haar [1910], much research has been devoted to designing a wavelet set that is adapted to particular signals such as natural images. After a long quest for finding good orthogonal basis such as the one proposed by Daubechies [1988], a series of works has focused on wavelet sets whose elements are not linearly independent. It resulted a large number of variants, such as steerable wavelets [Simoncelli et al., 1992], curvelets [Candès and Donoho, 2002], contourlets [Do and Vetterli, 2005], or bandlets [Le Pennec and Mallat, 2005]. For the purpose of our monograph, one concept related to sparse estimation is particularly important; it is called *wavelet thresholding*.

Sparse estimation and wavelet thresholding. Let us consider a discrete signal represented by a vector $\mathbf{x}$ in $\mathbb{R}^p$ and an orthogonal wavelet basis set $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_p]$—that is, satisfying $\mathbf{D}^\top \mathbf{D} = \mathbf{I}$ where $\mathbf{I}$ is the

identity matrix. Approximating $\mathbf{x}$ by a sparse linear combination of wavelet elements can be formulated as finding a sparse vector $\boldsymbol{\alpha}$ in $\mathbb{R}^p$, say with k non-zero coefficients, that minimizes

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2 \;\text{ s.t. }\; \|\boldsymbol{\alpha}\|_0 \leq k. \tag{1.3}$$

The sparse decomposition problem (1.3) is an instance of the best subset selection formulation presented in Section 1.1 where $\boldsymbol{\alpha}$ represents model parameters, demonstrating that similar topics arise in statistics and signal processing. However, whereas (1.3) is NP-hard for general matrices $\mathbf{D}$ [Natarajan, 1995], we have assumed $\mathbf{D}$ to be orthogonal in the context of wavelets. As such, (1.3) is equivalent to

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\left\|\mathbf{D}^\top\mathbf{x}-\boldsymbol{\alpha}\right\|_2^2 \;\text{ s.t. }\; \|\boldsymbol{\alpha}\|_0 \leq k,$$

and admits a closed form. Let us indeed define the vector $\boldsymbol{\beta} \triangleq \mathbf{D}^\top\mathbf{x}$ in $\mathbb{R}^p$, corresponding to the exact non-sparse decomposition of $\mathbf{x}$ onto $\mathbf{D}$—that is, we have $\mathbf{x} = \mathbf{D}\boldsymbol{\beta}$ since $\mathbf{D}$ is orthogonal. To obtain the best k-sparse approximation, we denote by μ the k-th largest value among the set $\{|\boldsymbol{\beta}[1]|,\ldots,|\boldsymbol{\beta}[p]|\}$, and the solution $\boldsymbol{\alpha}^{\text{ht}}$ of (1.3) is obtained by applying to $\boldsymbol{\beta}$ an operator called "hard-thresholding" and defined as

$$\boldsymbol{\alpha}^{\text{ht}}[i] = \mathbf{1}_{|\boldsymbol{\beta}[i]|\geq\mu}\boldsymbol{\beta}[i] = \begin{cases} \boldsymbol{\beta}[i] & \text{if } |\boldsymbol{\beta}[i]| \geq \mu, \\ 0 & \text{otherwise,} \end{cases} \tag{1.4}$$

where $\mathbf{1}_{|\boldsymbol{\beta}[i]|\geq\mu}$ is the indicator function, which is equal to 1 if $|\boldsymbol{\beta}[i]| \geq \mu$ and 0 otherwise. In other words, the hard-thresholding operator simply sets to zero coefficients from $\boldsymbol{\beta}$ whose magnitude is below the threshold μ. The corresponding procedure, called "wavelet thresholding", is simple and effective for image denoising, even though it does not perform as well as recent state-of-the-art techniques presented in Section 3. When an image $\mathbf{x}$ is noisy, *e.g.*, corrupted by white Gaussian noise, and μ is well chosen, the estimate $\mathbf{D}\boldsymbol{\alpha}^{\text{ht}}$ is a good estimate of the clean original image. The terminology "hard" is defined in contrast to an important variant called the "soft-thresholding operator", which was in-

troduced by Donoho and Johnstone [1994] in the context of wavelets:[3]

$$\boldsymbol{\alpha}^{\mathrm{st}}[i] \triangleq \mathrm{sign}(\boldsymbol{\beta}[i]) \max(|\boldsymbol{\beta}[i]| - \lambda, 0) = \begin{cases} \boldsymbol{\beta}[i] - \lambda & \text{if } \boldsymbol{\beta}[i] \geq \lambda, \\ \boldsymbol{\beta}[i] + \lambda & \text{if } \boldsymbol{\beta}[i] \leq -\lambda, \\ 0 & \text{otherwise,} \end{cases} \tag{1.5}$$

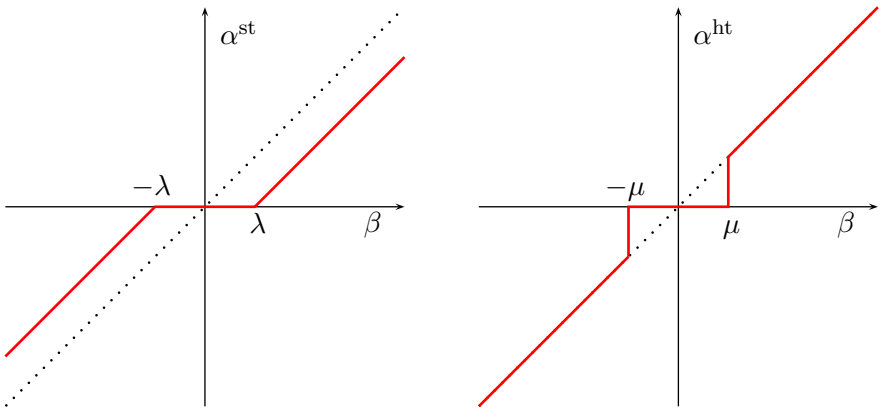
where λ is a parameter playing the same role as μ in (1.4). Not only does the operator set small coefficients of $\boldsymbol{\beta}$ to zero, but it also reduces the magnitude of the non-zero ones. Both operators are illustrated and compared to each other in Figure 1.1. Interestingly, whereas $\boldsymbol{\alpha}^{\mathrm{ht}}$ is the solution of (1.3) when μ corresponds to the entry of $\boldsymbol{\beta} = \mathbf{D}^\top \mathbf{x}$ with k-th largest magnitude, $\boldsymbol{\alpha}^{\mathrm{st}}$ is in fact the solution of the following sparse reconstruction problem with the orthogonal matrix $\mathbf{D}$:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1. \tag{1.6}$$

This formulation will be the topic of the next section for general non-orthogonal matrices. Similar to statistics where choosing the parameter k of the best subset selection was difficult, automatically selecting the best thresholds μ or λ has been a major research topic [see, *e.g.* Donoho and Johnstone, 1994, 1995, Chang et al., 2000a,b].

Structured group thresholding. Wavelets coefficients have a particular structure since the basis elements $\mathbf{d}_i$ are dilated and shifted versions of each other. It is for instance possible to define neighborhood relationships for wavelets whose spatial supports are close to each other, or hierarchical relationships between wavelets with same or similar localization but with different scales. For one-dimensional signals, we present in Figure 1.2 a typical organization of wavelet coefficients on a tree with arrows representing such relations. For two-dimensional images, the structure is slightly more involved and the coefficients are usually organized as a collection of quadtrees [see Mallat, 2008, for more details]; we present such a configuration in Figure 1.3.

[3] Note that the soft-thresholding operator appears in fact earlier in the statistics literature [see Efron and Morris, 1971, Bickel, 1984], but it was used there for a different purpose.



(a) Soft-thresholding operator, $\alpha^{\mathrm{st}} = \mathrm{sign}(\beta)\max(|\beta| - \lambda, 0)$.

(b) Hard-thresholding operator $\alpha^{\mathrm{ht}} = \mathbf{1}_{|\beta|\geq\mu}\beta$.

Figure 1.1: Soft- and hard-thresholding operators, which are commonly used for signal estimation with orthogonal wavelet basis.

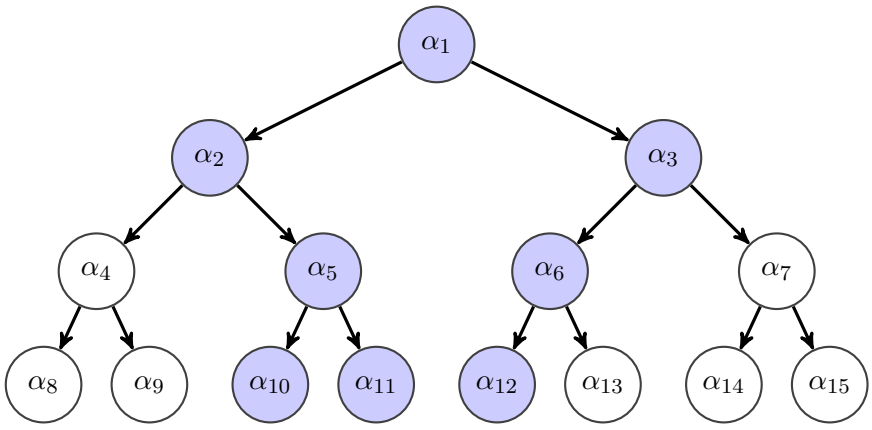


Figure 1.2: Illustration of a wavelet tree with four scales for one-dimensional signals. Nodes represent wavelet coefficients and their depth in the tree correspond to the scale parameter of the wavelet. We also illustrate the zero-tree coding scheme [Shapiro, 1993] in this figure. Empty nodes correspond to zero coefficient: according to the zero-tree coding scheme, their descendants in the tree are also zero.

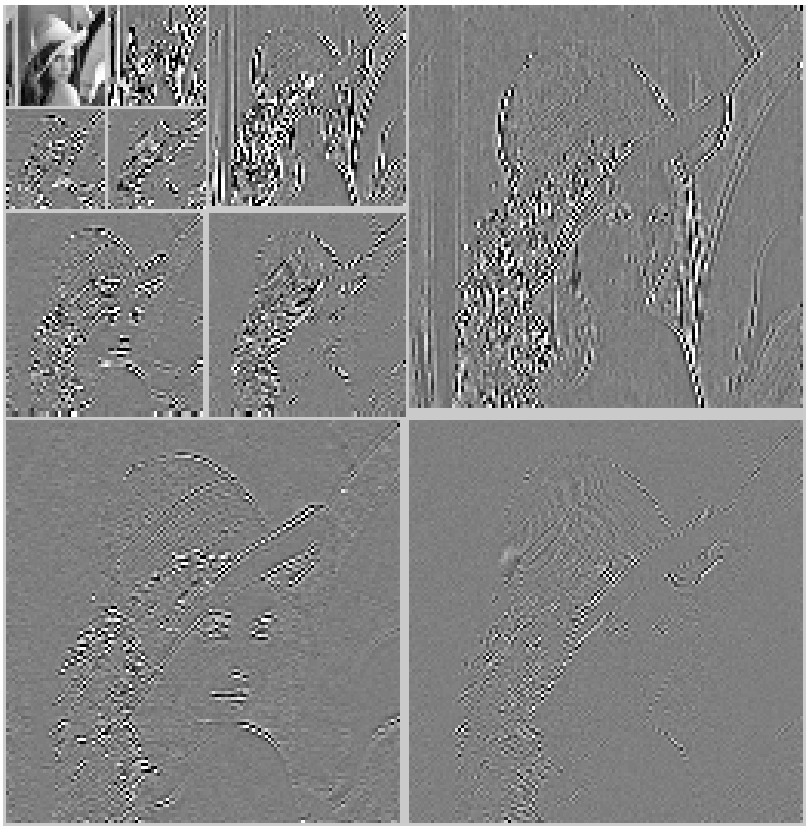


Figure 1.3: Wavelet coefficients displayed for the image lena using the orthogonal basis of Daubechies [1988]. A few coefficients representing a low-resolution version of the image are displayed on the top-left corner. Wavelets corresponding to this low-resolution image are obtained by filtering the original image with shifted versions of a low-pass filter called "scaling function" or "father wavelet". The rest of the coefficients are organized into three quadtrees (on the right, on the left, and on the diagonal). Each quadtree is obtained by filtering the original image with a wavelet at three different scales and at different positions. The value zero is represented by the grey color; negative values appear in black, and positive values in white. The wavelet decomposition and this figure have been produced with the software package matlabPyrTools developed by Eero Simoncelli and available here: `http://www.cns.nyu.edu/~lcv/software.php`.

A natural idea has inspired the recent concept of *group sparsity* that will be presented in the next section; it consists in exploiting the wavelet structure to improve thresholding estimators. Specifically, it is possible to use neighborhood relations between wavelet basis elements to define groups of coefficients that form a partition $\mathcal{G}$ of $\{1,\dots,p\}$, and use a group-thresholding operator [Hall et al., 1999, Cai, 1999] defined for every group g in $\mathcal{G}$ as

$$\boldsymbol{\alpha}^{\mathrm{gt}}[g] \triangleq \begin{cases} \left(1 - \frac{\lambda}{\|\boldsymbol{\beta}[g]\|_2}\right)\boldsymbol{\beta}[g] & \text{if} \quad \|\boldsymbol{\beta}[g]\|_2 \geq \lambda, \\ 0 & \text{otherwise,} \end{cases} \tag{1.7}$$

where $\boldsymbol{\beta}[g]$ is the vector of size $|g|$ recording the entries of $\boldsymbol{\beta}$ whose indices are in g. By using such an estimator, groups of neighbor coefficients are set to zero together when their joint ℓ_2-norm falls below the threshold λ. Interestingly, even though the next interpretation does not appear in early work about group-thresholding [Hall et al., 1999, Cai, 1999], it is possible to view $\boldsymbol{\alpha}^{\mathrm{gt}}$ with $\boldsymbol{\beta} = \mathbf{D}^\top\mathbf{x}$ as the solution of the following penalized problem

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{g\in\mathcal{G}} \|\boldsymbol{\alpha}[g]\|_2, \tag{1.8}$$

where the closed-form solution (1.7) holds because $\mathbf{D}$ is orthogonal [see Bach et al., 2012a]. Such a formulation will be studied in the next section for general matrices.

Finally, other ideas for exploiting both structure and wavelet parsimony have been proposed. One is a coding scheme called "zero-tree" wavelet coding [Shapiro, 1993], which uses the tree structure of wavelets to force all descendants of zero coefficients to be zero as well. Equivalently, a coefficient can be non-zero only if its parent in the tree is non-zero, as illustrated in Figure 1.2. This idea has been revisited later in a more general context by Zhao et al. [2009]. Other complex models have been used as well for modeling interactions between coefficients: we can mention the application of hidden Markov models (HMM) to wavelets by Crouse et al. [1998] and the Gaussian scale mixture model of Portilla et al. [2003].

1.3 Modern parsimony: the ℓ_1-norm and other variants

The era of "modern" parsimony corresponds probably to the use of convex optimization techniques for solving feature selection or sparse decomposition problems. Even though the ℓ_1-norm was introduced for that purpose in geophysics [Claerbout and Muir, 1973, Taylor et al., 1979], it was popularized in statistics with the Lasso estimator of Tibshirani [1996] and independently in signal processing with the basis pursuit formulation of Chen et al. [1999]. Given observations $\mathbf{x}$ in $\mathbb{R}^n$ and a matrix of predictors $\mathbf{D}$ in $\mathbb{R}^{n\times p}$, the Lasso consists of learning a linear model $\mathbf{x} \approx \mathbf{D}\boldsymbol{\alpha}$ by solving the following quadratic program:

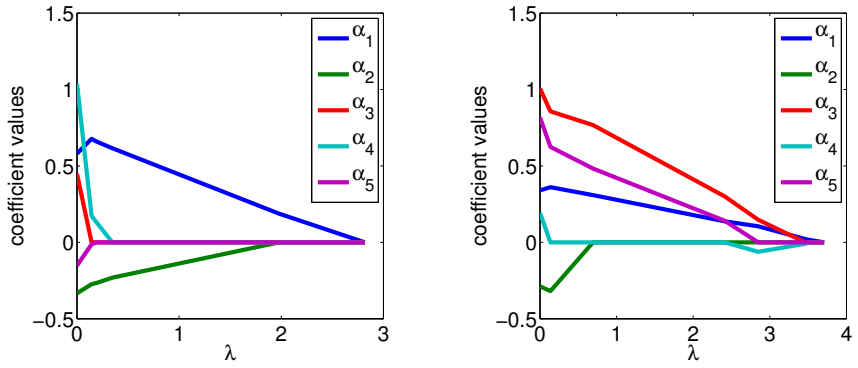
$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_1 \leq \mu. \tag{1.9}$$

As detailed in the sequel, the ℓ_1-norm encourages the solution $\boldsymbol{\alpha}$ to be sparse and the parameter μ is used to control the trade-off between data fitting and the sparsity of $\boldsymbol{\alpha}$. In practice, reducing the value of μ leads indeed to sparser solution in general, *i.e.*, with more zeroes, even though there is no formal relation between the sparsity of $\boldsymbol{\alpha}$ and its ℓ_1-norm for general matrices $\mathbf{D}$.

The basis pursuit denoising formulation of Chen et al. [1999] is relatively similar but the ℓ_1-norm is used as a penalty instead of a constraint. It can be written as

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \tag{1.10}$$

which is essentially equivalent to (1.9) from a convex optimization perspective, and in fact (1.10) is also often called "Lasso" in the literature. Given some data $\mathbf{x}$, matrix $\mathbf{D}$, and parameter $\mu > 0$, we indeed know from Lagrange multiplier theory [see, *e.g.*, Borwein and Lewis, 2006, Boyd and Vandenberghe, 2004] that for all solution $\boldsymbol{\alpha}^\star$ of (1.9), there exists a parameter $\lambda \geq 0$ such that $\boldsymbol{\alpha}^\star$ is also a solution of (1.10). We note, however, that there is no direct mapping between λ and μ, and thus the choice of formulation (1.9) or (1.10) should be made according to how easy it is to select the parameters λ or μ. For instance, one may prefer (1.9) when a priori information about the ℓ_1-norm of the solution is available. In Figure 1.4, we illustrate the effect of changing



(a) Path for dataset 1 (b) Path for dataset 2

Figure 1.4: Two examples of regularization paths for the Lasso/Basis Pursuit. The curves represent the values of the $p = 5$ entries of the solutions of (1.10) when varying the parameter λ for two datasets. On the left, the relation between λ and the sparsity of the solution is monotonic; On the right, this is not the case. Note that the paths are piecewise linear, see Section 5.2 for more details.

the value of the regularization parameter λ on the solution of (1.10) for two datasets. When $\lambda = 0$, the solution is dense; in general, increasing λ sets more and more variables to zero. However, the relation between λ and the sparsity of the solution is not exactly monotonic. In a few cases, increasing λ yields a denser solution.

Another "equivalent" formulation consists of finding a sparse decomposition under a reconstruction constraint:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\boldsymbol{\alpha}\|_1 \quad \text{s.t.} \quad \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq \varepsilon. \tag{1.11}$$

This formulation can be useful when we have a priori knowledge about the noise level and the parameter ε is easy to choose. The link between (1.10) and (1.11) is similar to the link between (1.10) and (1.9).

For noiseless problems, Chen et al. [1999] have also introduced a formulation simply called "basis pursuit" (without the terminology "denoising"), defined as

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\boldsymbol{\alpha}\|_1 \quad \text{s.t.} \quad \mathbf{x} = \mathbf{D}\boldsymbol{\alpha}, \tag{1.12}$$

which is related to (1.10) in the sense that the set of solutions of (1.10) converges to the solutions of (1.12) when λ converges to 0^+, whenever the linear system $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$ is feasible. These four formulations (1.9-1.12) have gained a large success beyond the statistics and signal processing communities. More generally, the ℓ_1-norm has been used as a regularization function beyond the least-square context, leading to problems of the form

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} f(\boldsymbol{\alpha}) + \lambda\|\boldsymbol{\alpha}\|_1, \tag{1.13}$$

where $f : \mathbb{R}^p \to \mathbb{R}$ is a loss function. In the rest of this section, we will present several variants of the ℓ_1-norm, but before that, we will try to understand why such a penalty is appropriate for sparse estimation.

Why does the ℓ_1-norm induce sparsity? Even though we have claimed that there is no rigorous relation between the sparsity of $\boldsymbol{\alpha}$ and its ℓ_1-norm in general, intuition about the sparsity-inducing effect of the ℓ_1-norm may be obtained from several viewpoints.

Analytical point of view. In the previous section about wavelets, we have seen that when $\mathbf{D}$ is orthogonal, the ℓ_1-decomposition problem (1.10) admits an analytic closed form solution (1.5) obtained by soft-thresholding. As a result, whenever the magnitude of the inner product $\mathbf{d}_i^\top \mathbf{x}$ is smaller than λ for an index i, the corresponding variable $\boldsymbol{\alpha}^\star[i]$ is equal to zero. Thus, the number of zeroes of the solution $\boldsymbol{\alpha}^\star$ monotonically increases with λ.

For non-orthogonal matrices $\mathbf{D}$, such a monotonic relation does not formally hold anymore; in practice, the sparsity-inducing property of the ℓ_1-penalty remains effective, as illustrated in Figure 1.4. Some intuition about this fact can be gained by studying optimality conditions for the general ℓ_1-regularized problem (1.13) where f is a differentiable function. The following lemma details these conditions.

Lemma 1.1 (Optimality conditions for ℓ_1-regularized problems)**.** A vector $\boldsymbol{\alpha}^\star$ in $\mathbb{R}^p$ is a solution of (1.13) if and only if

$$\forall i = 1,\dots,p \quad \begin{cases} -\nabla f(\boldsymbol{\alpha}^\star)[i] = \lambda \operatorname{sign}(\boldsymbol{\alpha}^\star[i]) & \text{if } \boldsymbol{\alpha}^\star[i] \neq 0, \\ |\nabla f(\boldsymbol{\alpha}^\star)[i]| \leq \lambda & \text{otherwise.} \end{cases} \tag{1.14}$$

Proof. A proof using the classical concept of subdifferential from convex optimization can be found in [see,*e.g.*, Bach et al., 2012a]. Here, we provide instead an elementary proof using the simpler concept of directional derivative for nonsmooth functions, defined as, when the limit exists,

$$\nabla g(\boldsymbol{\alpha}, \boldsymbol{\kappa}) \triangleq \lim_{t \to 0^+} \frac{g(\boldsymbol{\alpha} + t\boldsymbol{\kappa}) - g(\boldsymbol{\alpha})}{t},$$

for a function $g : \mathbb{R}^p \to \mathbb{R}$ at a point $\boldsymbol{\alpha}$ in $\mathbb{R}^p$ and a direction $\boldsymbol{\kappa}$ in $\mathbb{R}^p$. For convex functions g, directional derivatives always exist and a classical optimality condition for $\boldsymbol{\alpha}^\star$ to be a minimum of g is to have $\nabla g(\boldsymbol{\alpha}^\star, \boldsymbol{\kappa})$ non-negative for all directions $\boldsymbol{\kappa}$ [Borwein and Lewis, 2006]. Intuitively, this means that one cannot find any direction $\boldsymbol{\kappa}$ such that an infinitesimal move along $\boldsymbol{\kappa}$ from $\boldsymbol{\alpha}^\star$ decreases the value of the objective. When g is differentiable, the condition is equivalent to the classical optimality condition $\nabla g(\boldsymbol{\alpha}^\star) = 0$.

We can now apply the directional derivative condition to the function $g : \boldsymbol{\alpha} \mapsto f(\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|_1$, which is equivalent to

$$\forall \boldsymbol{\kappa} \in \mathbb{R}^p, \quad \nabla f(\boldsymbol{\alpha}^\star)^\top \boldsymbol{\kappa} + \lambda \sum_{i=1}^{p} \left\{ \begin{array}{ll} \operatorname{sign}(\boldsymbol{\alpha}^\star[i])\boldsymbol{\kappa}[i] & \text{if } \boldsymbol{\alpha}^\star[i] \neq 0, \\ |\boldsymbol{\kappa}[i]| & \text{otherwise} \end{array} \right\} \geq 0. \tag{1.15}$$

It is then easy to show that (1.15) holds for all $\boldsymbol{\kappa}$ if and only the inequality holds for the specific values $\boldsymbol{\kappa} = \mathbf{e}_i$ and $\boldsymbol{\kappa} = -\mathbf{e}_i$ for all i, where $\mathbf{e}_i$ is the vector in $\mathbb{R}^p$ with zeroes everywhere except for the i-th entry that is equal to one. This immediately provides an equivalence between (1.15) and (1.14). □

Lemma 1.1 is interesting from a computational point of view (see Section 5.2), but it also tells us that when $\lambda \geq \|\nabla f(0)\|_\infty$, the conditions (1.14) are satisfied for $\boldsymbol{\alpha}^\star = 0$, the sparsest solution possible.

Physical point of view. In image processing or computer vision, the word "energy" often denotes the objective function of a minimization problem; it is indeed common in physics to have complex systems that stabilize at a configuration of minimum potential energy. The negative of the energy's gradient represents a force, a terminology we will

borrow in this paragraph. Consider for instance a one-dimensional ℓ_1-regularized estimation problem

$$\min_{\alpha \in \mathbb{R}} \frac{1}{2}(\beta - \alpha)^2 + \lambda |\alpha|, \tag{1.16}$$

where β is a positive constant. Whenever α is non-zero, the ℓ_1-penalty is differentiable with derivative $\lambda \operatorname{sign}(\alpha)$. When interpreting this objective as an energy minimization problem, the ℓ_1-penalty can be seen as applying *a force driving α towards the origin with constant intensity λ*. Consider now instead the squared ℓ_2-penalty, also called regularization of Tikhonov [1963], or ridge regression regularization [Hoerl and Kennard, 1970]:

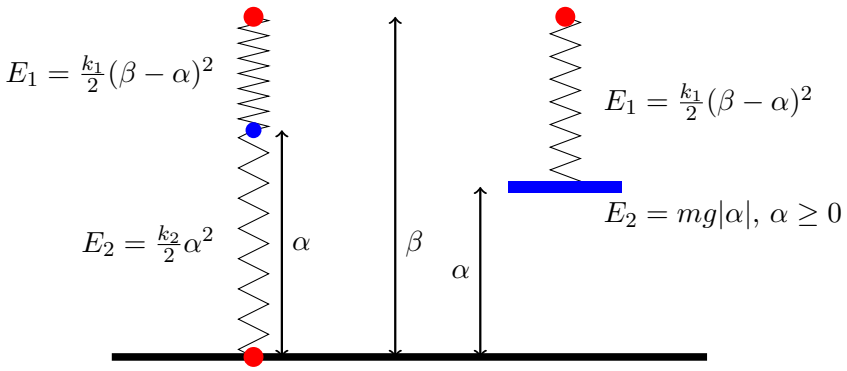
$$\min_{\alpha \in \mathbb{R}} \frac{1}{2}(\beta - \alpha)^2 + \frac{\lambda}{2}\alpha^2. \tag{1.17}$$

The derivative of the quadratic energy $(\lambda/2)\alpha^2$ is $\lambda\alpha$. It can be interpreted as *a force that also points to the origin but with linear intensity $\lambda|\alpha|$*. Therefore, the force corresponding to the ridge regularization can be arbitrarily strong when α is large, but if fades away when α gets close to zero. As a result, the squared ℓ_2-regularization does not have a sparsity-inducing effect. From an analytical point of view, we have seen that the solution of (1.16) is zero when $|\beta|$ is smaller than λ. In contrast, the solution of (1.17) admits a closed form $\alpha^\star = \beta/(1+\lambda)$. And thus, regardless of the parameter λ, the solution is never zero.

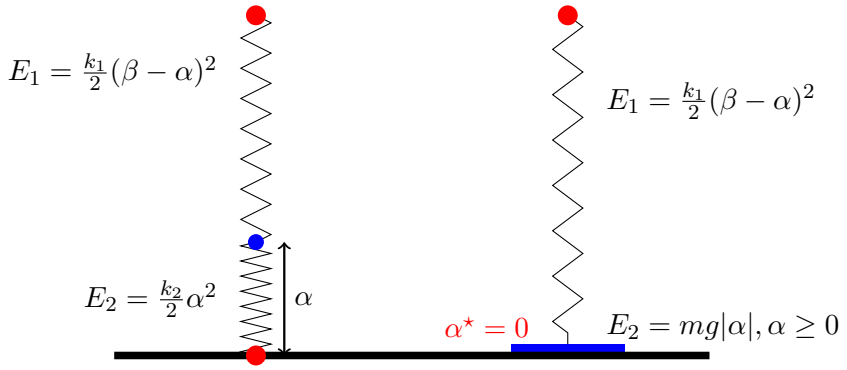
We present a physical example illustrating this phenomenon in Figure 1.5. We use springs whose potential energy is known to be quadratic, and objects with a gravitational potential energy that is approximately linear on the Earth's surface.

Geometrical point of view. The sparsity-inducing effect of the ℓ_1-norm can also be interpreted by studying the geometry of the ℓ_1-ball $\{\boldsymbol{\alpha} \in \mathbb{R}^p : \|\boldsymbol{\alpha}\|_1 \leq \mu\}$. More precisely, understanding the effect of the Euclidean projection onto this set is important: in simple cases where the design matrix $\mathbf{D}$ is orthogonal, the solution of (1.9) can indeed be obtained by the projection

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_1 \leq \mu, \tag{1.18}$$



(a) Small regularization



(b) High regularization

Figure 1.5: A physical system illustrating the sparsity-inducing effect of the ℓ_1-norm (on the right) in contrast to the Tikhonov-ridge regularization (on the left). Three springs are represented in each figure, two on the left, one on the right. Red points are fixed and cannot move. On the left, two springs are linked to each other by a blue point whose position can vary. On the right, a blue object of mass m is attached to the spring. Right and left configurations define two different dynamical systems with energies E_1+E_2; on the left, E_1 and E_2 are elastic potential energies; on the right, E_1 is the same as on the left, whereas E_2 is a gravitational potential energy, where g is the gravitational constant on the Earth's surface. Both system can evolve according to their initial positions, and stabilize for the value of $\alpha^\star$ that minimizes the energy $E_1 + E_2$, assuming that some energy can be dissipated by friction forces. On the left, it is possible to show that $\alpha^\star = \beta k_1/(k_1 + k_2)$ and thus, the solution $\alpha^\star$ is never equal to zero, regardless of the strength k_2 of the bottom spring. On the right, the solution is obtained by soft-thresholding: $\alpha^\star = \max(\beta - mg/k_1, 0)$. As shown on Figure 1.5(b), when the mass m is large enough, the blue object touches the ground and $\alpha^\star = 0$. Figure adapted from [Mairal, 2010].

where $\boldsymbol{\beta} = \mathbf{D}^\top \mathbf{x}$. When $\mathbf{D}$ is not orthogonal, a classical algorithm for solving (1.9) is the projected gradient method (see Section 5.2), which performs a sequence of projections (1.18) for different values of $\boldsymbol{\beta}$. Note that how to solve (1.18) efficiently is well studied; it can be achieved in $O(p)$ operations with a divide-and-conquer strategy [Brucker, 1984, Duchi et al., 2008].

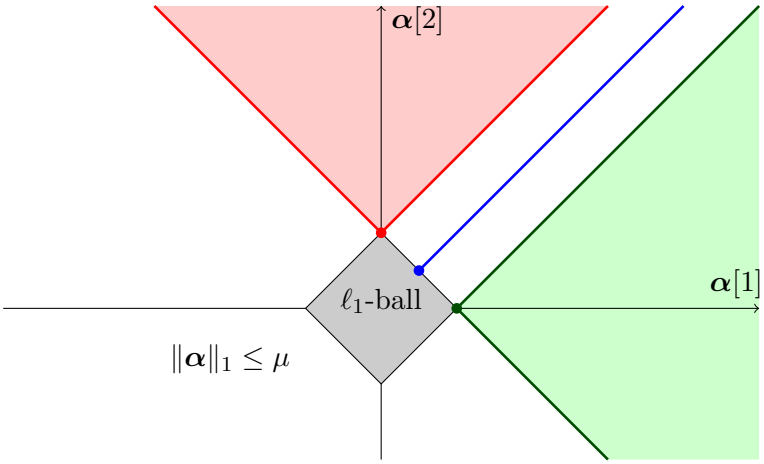
In Figure 1.6, we illustrate the effect of the ℓ_1-norm projection and compare it to the case of the ℓ_2-norm. The corners of the ℓ_1-ball are on the main axes and correspond to sparse solutions. Two of them are represented by red and green dots, with respective coordinates $(\mu, 0)$ and $(0, \mu)$. Most strikingly, a large part of the space in the figure, represented by red and green regions, ends up on these corners after projection. In contrast, the set of points that is projected onto the blue dot, is simply the blue line. The blue dot corresponds in fact to a dense solution with coordinates $(\mu/2, \mu/2)$. Therefore, the figure illustrates that the ℓ_1-ball in two dimensions encourages solutions to be on its corners. In the case of the ℓ_2-norm, the ball is isotropic, and treats every direction equally. In Figure 1.7, we represent these two balls in three dimensions, where we can make similar observations.

More formally, we can mathematically characterize our remarks about Figure 1.6. Consider a point $\mathbf{y}$ in $\mathbb{R}^p$ on the surface of the ℓ_1-ball of radius $\mu = 1$, and define the set $\mathcal{N} \triangleq \{\mathbf{z} \in \mathbb{R}^p : \pi(\mathbf{z}) = \mathbf{y}\}$, where π is the projection operator onto the ℓ_1-ball. Examples of pairs $(\mathbf{y}, \mathcal{N})$ have been presented in Figure 1.6; for instance, when $\mathbf{y}$ is the red or green dot, $\mathcal{N}$ is respectively the red or green region. It is particularly informative to study how $\mathcal{N}$ varies with $\mathbf{y}$, which is the focus of the next proposition.

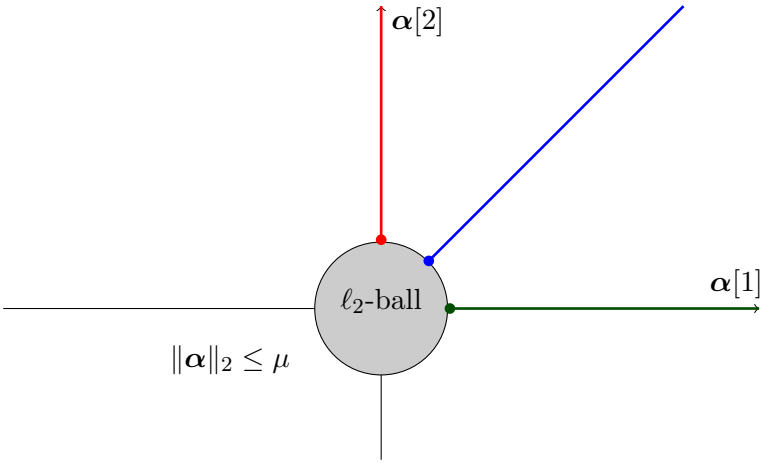
Proposition 1.1 (Characterization of the set $\mathcal{N}$)**.**
For a non-zero vector $\mathbf{y}$ in $\mathbb{R}^p$, the set $\mathcal{N}$ defined in the previous paragraph can be written as $\mathcal{N} = \mathbf{y} + \mathcal{K}$, where $\mathcal{K}$ is a polyhedral cone of dimension $p - \|\mathbf{y}\|_0 + 1$.

Proof. A classical theorem [see Bertsekas, 1999, Proposition B.11] allows us to rewrite $\mathcal{N}$ as

$$\mathcal{N} = \{\mathbf{z} \in \mathbb{R}^p \ : \ \forall \ \|\mathbf{x}\|_1 \leq 1, \ \ (\mathbf{z} - \mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \leq 0\} = \mathbf{y} + \mathcal{K},$$

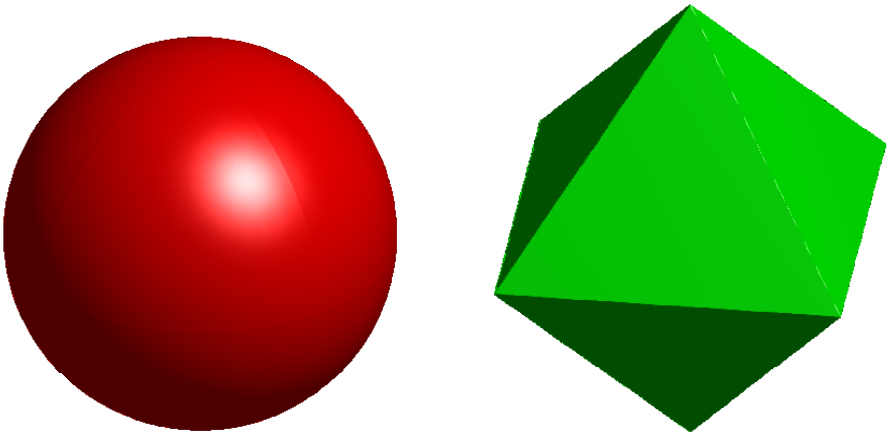


(a) Effect of the Euclidean projection onto the ℓ_1-ball.



(b) Effect of the Euclidean projection onto the ℓ_2-ball.

Figure 1.6: Illustration in two dimensions of the projection operator onto the ℓ_1-ball in Figure (a) and ℓ_2-ball in Figure (b). The balls are represented in gray. All points from the red regions are projected onto the point of coordinates $(0, \mu)$ denoted by a red dot. Similarly, the green and blue regions are projected onto the green and blue dots, respectively. For the ℓ_1-norm, a large part of the figure is filled by the red and green regions, whose points are projected to a sparse solution corresponding to a corner of the ball. For the ℓ_2-norm, this is not the case: any non-sparse point—say, for instance on the blue line—is projected onto a non-sparse solution.



(a) ℓ_2-ball in 3D (b) ℓ_1-ball in 3D

Figure 1.7: Representation in three dimensions of the ℓ_1- and ℓ_2-balls. Figure borrowed from Bach et al. [2012a], produced by Guillaume Obozinski.

where $\mathbf{y} + \mathcal{K}$ denotes the Minkowski sum $\{\mathbf{y} + \mathbf{z} : \mathbf{z} \in \mathcal{K}\}$ between the set $\{\mathbf{y}\}$ and the cone $\mathcal{K}$ defined as

$$\mathcal{K} \triangleq \{\mathbf{d} \in \mathbb{R}^p \ : \ \forall \ \|\mathbf{x}\|_1 \leq 1, \ \ \mathbf{d}^\top(\mathbf{x} - \mathbf{y}) \leq 0\}.$$

Note that in the optimization literature, $\mathcal{K}$ is often called the "normal cone" to the unit ℓ_1-ball at the point $\mathbf{y}$ [Borwein and Lewis, 2006]. Equivalently, we have

$$\begin{aligned} \mathcal{K} &= \{\mathbf{d} \in \mathbb{R}^p \ : \ \max_{\|\mathbf{x}\|_1 \leq 1} \mathbf{d}^\top \mathbf{x} \leq \mathbf{d}^\top \mathbf{y}\} \\ &= \{\mathbf{d} \in \mathbb{R}^p \ : \ \|\mathbf{d}\|_\infty \leq \mathbf{d}^\top \mathbf{y}\}, \end{aligned} \tag{1.19}$$

where we have used the fact that quantity $\max_{\|\mathbf{x}\|_1 \leq 1} \mathbf{d}^\top \mathbf{x}$, called the dual-norm of the ℓ_1-norm, is equal to $\|\mathbf{d}\|_\infty$ [see Bach et al., 2012a]. Note now that according to Hölder's inequality, we also have $\mathbf{d}^\top \mathbf{y} \leq \|\mathbf{d}\|_\infty \|\mathbf{y}\|_1 \leq \|\mathbf{d}\|_\infty$ in Eq. (1.19). Therefore, the inequalities are in fact equalities. It is then easy to characterize vectors $\mathbf{d}$ such that $\mathbf{d}^\top \mathbf{y} = \|\mathbf{d}\|_\infty \|\mathbf{y}\|_1$ and it is possible to show that $\mathcal{K}$ is simply the set of vectors $\mathbf{d}$ satisfying $\mathbf{d}[i] = \text{sign}(\mathbf{y}[i])\|\mathbf{d}\|_\infty$ for all i such that $\mathbf{y}[i] \neq 0$.

This would be sufficient to conclude the proposition, but it is also possible to pursue the analysis and exactly characterize $\mathcal{K}$ by finding

a set of generators.[4] Let us define the vector $\mathbf{s}$ in $\{-1, 0, +1\}^p$ that carries the sparsity pattern of $\mathbf{y}$, more precisely, with $\mathbf{s}[i] = \text{sign}(\mathbf{y}[i])$ for all i such that $\mathbf{y}[i] \neq 0$, and $\mathbf{s}[i] = 0$ otherwise. Let us also define the set of indices $\{i_1, \ldots, i_l\}$ corresponding to the l zero entries of $\mathbf{y}$, and $\mathbf{e}_i$ in $\mathbb{R}^p$ the binary vector whose entries are all zero but the i-th one that is equal to 1. Then, after a short calculation, we can geometrically characterize the polyhedral cone $\mathcal{K}$:

$$\mathcal{K} = \text{cone}\left(\mathbf{s}, \mathbf{s} - \mathbf{e}_{i_1}, \mathbf{s} + \mathbf{e}_{i_1}, \mathbf{s} - \mathbf{e}_{i_2}, \mathbf{s} + \mathbf{e}_{i_2}, \ldots, \mathbf{s} - \mathbf{e}_{i_l}, \mathbf{s} + \mathbf{e}_{i_l}\right),$$

where the notation "cone" is defined in footnote 4. □

It is now easy to see that the set $\mathcal{K}$ "grows" with the number l of zero entries in $\mathbf{y}$, and that $\mathcal{K}$ lives in a subspace of dimension $l + 1$ for all non-zero vector $\mathbf{y}$. For example, when $l = 0$—that is, $\mathbf{y}$ is a dense vector (*e.g.*, the blue point in Figure 1.6(a)), $\mathcal{K}$ is simply a half-line.

To conclude, the geometrical intuition to gain from this section is that *the Euclidean projection onto a convex set encourages solutions on singular points, such as edges or corners for polytopes.* Such a principle indeed applies beyond the ℓ_1-norm. For instance, we illustrate the regularization effect of the ℓ_∞-norm in Figure 1.8, whose corners coordinates have same magnitude.

Non-convex regularization. Even though it is well established that the ℓ_1-norm encourages sparse solutions, it remains only a convex proxy of the ℓ_0-penalty. Both in statistics and signal processing, other sparsity-inducing regularization functions have been proposed, in particular continuous relaxations of ℓ_0 that are non-convex [Frank and Friedman, 1993, Fan and Li, 2001, Daubechies et al., 2010, Gasso et al., 2009]. These functions are using a non-decreasing concave function $\varphi : \mathbb{R}^+ \mapsto \mathbb{R}$, and the sparsity-inducing penalty is defined as

$$\psi(\boldsymbol{\alpha}) \triangleq \sum_{i=1}^{p} \varphi\left(|\boldsymbol{\alpha}[i]|\right).$$

[4] A collection of vectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_l$ are called generators for a cone $\mathcal{K}$ when $\mathcal{K}$ consists of all positive combinations of the vectors $\mathbf{z}_i$. In other words, $\mathcal{K} = \{\sum_{i=1}^{l} \alpha_i \mathbf{z}_i : \alpha_i \geq 0\}$. In that case, we use the notation $\mathcal{K} = \text{cone}(\mathbf{z}_1, \ldots, \mathbf{z}_l)$.

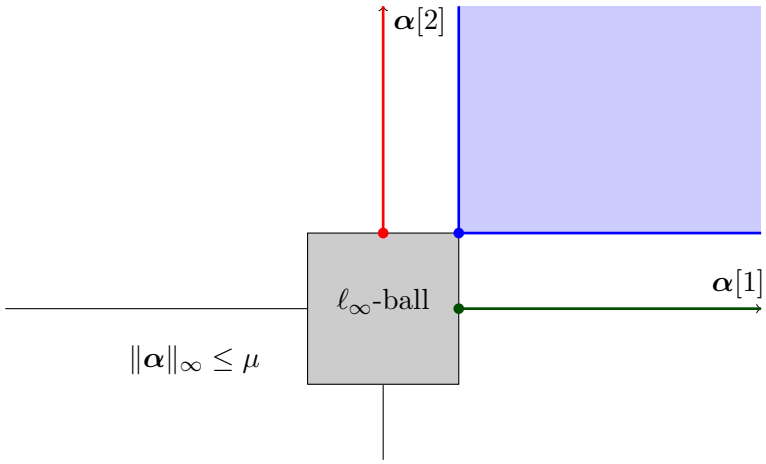
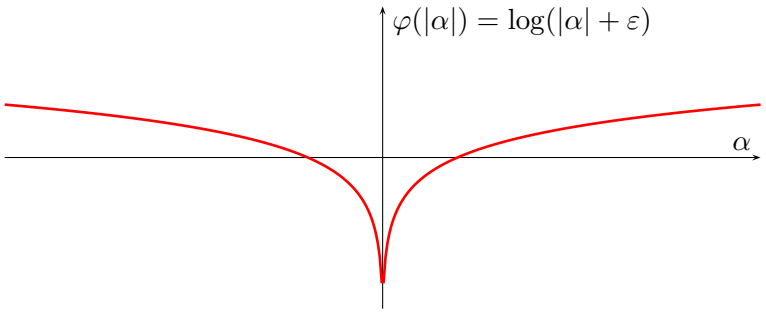


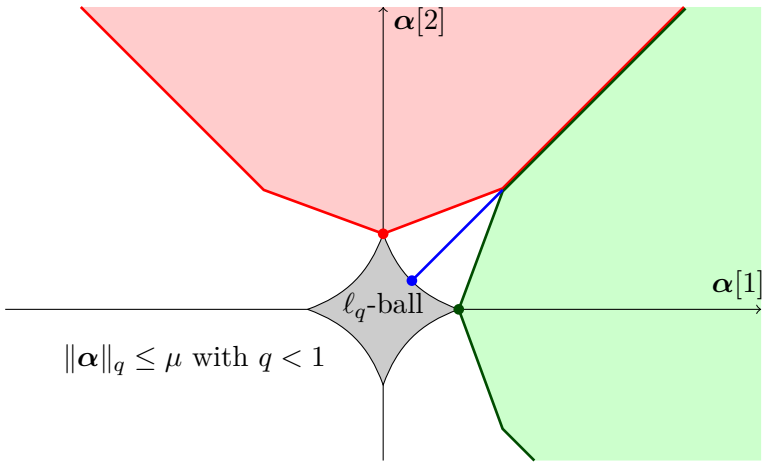
Figure 1.8: Similar illustration as Figure 1.6 for the ℓ_∞-norm. The regularization effect encourages solution to be on the corners of the ball, corresponding to points with the same magnitude $|\boldsymbol{\alpha}[1]| = |\boldsymbol{\alpha}[2]| = \mu$.

For example, the ℓ_q-penalty uses $\varphi : x \mapsto x^q$ [Frank and Friedman, 1993], or an approximation $\varphi : x \mapsto (x + \varepsilon)^q$; the reweighted-$\ell_1$ algorithm of Fazel [2002], Fazel et al. [2003], Candès et al. [2008] implicitly uses $\varphi : x \mapsto \log(x + \varepsilon)$. These penalties typically lead to intractable estimation problems, but approximate solutions can be obtained with continuous optimization techniques (see Section 5.3).

The sparsity-inducing effect of the penalties ψ is known to be stronger than ℓ_1. As shown in Figure 1.9(a), the magnitude of the derivative of φ grows when one approaches zero because of its concavity. Thus, in the one-dimensional case, ψ can be interpreted as *a force driving α towards the origin with increasing intensity when α gets closer to zero.* In terms of geometry, we also display the ℓ_q-ball in Figure 1.9(b), with the same red, blue, and green dots as in Figure 1.6. The part of the space that is projected onto the corners of the ℓ_q-ball is larger than that for ℓ_1. Interestingly, the geometrical structure of the red and green regions are also more complex. Their combinatorial nature makes the projection problem onto the ℓ_q-ball more involved when $q < 1$.



(a) Illustration of a non-convex sparsity-inducing penalty.



(b) ℓ_q-ball with $q < 1$.

Figure 1.9: Illustration of the sparsity-inducing effect of a non-convex penalty. In (a), we plot the non-convex penalty $\alpha \mapsto \log(|\alpha| + \varepsilon)$, and in (b), we present a similar figure as 1.6 for the ℓ_q-penalty, when choosing $q < 1$.

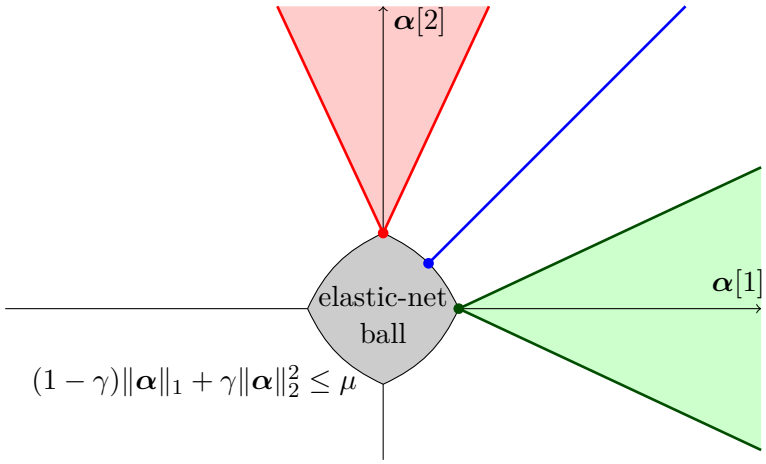


Figure 1.10: Similar figure as 1.6 for the elastic-net penalty.

The elastic-net. To cope with instability issues of estimators obtained with the ℓ_1-regularization, Zou and Hastie [2005] have proposed to combine the ℓ_1- and ℓ_2-norms with a penalty called elastic-net:

$$\psi(\boldsymbol{\alpha}) \triangleq \|\boldsymbol{\alpha}\|_1 + \gamma\|\boldsymbol{\alpha}\|_2^2.$$

The effect of this penalty is illustrated in Figure 1.10. Compared to Figure 1.6, we observe that the red and green regions are smaller for the elastic-net penalty than for ℓ_1. The sparsity-inducing effect is thus less aggressive than the one obtained with ℓ_1.

Total variation. The anisotropic total variation penalty [Rudin et al., 1992] for one dimensional signals is the ℓ_1-norm of finite differences

$$\psi(\boldsymbol{\alpha}) \triangleq \sum_{i=1}^{p-1} |\boldsymbol{\alpha}[i+1] - \boldsymbol{\alpha}[i]|,$$

which encourages piecewise constant signals. It is also known in statistics under the name of "fused Lasso" [Tibshirani et al., 2005]. The penalty can easily be extended to two-dimensional signals, and has been widely used for regularizing inverse problems in image processing [Chambolle, 2005].

Group sparsity. In some cases, variables are organized into predefined groups forming a partition $\mathcal{G}$ of $\{1, \ldots, p\}$, and one is looking for a solution $\boldsymbol{\alpha}^\star$ such that *variables belonging to the same group of $\mathcal{G}$ are set to zero together.* For example, such groups have appeared in Section 1.2 about wavelets, where $\mathcal{G}$ could be defined according to neighborhood relationships of wavelet coefficients. Then, when it is known beforehand that a problem solution only requires a few groups of variables to explain the data, a regularization function automatically selecting the relevant groups has been shown to improve the prediction performance or the interpretability of the solution [Turlach et al., 2005, Yuan and Lin, 2006, Obozinski et al., 2009, Huang and Zhang, 2010]. The group sparsity principle is illustrated in Figure 1.11(b).

An appropriate regularization function to obtain a group-sparsity effect is known as "Group-Lasso" penalty and is defined as

$$\psi(\boldsymbol{\alpha}) = \sum_{g \in \mathcal{G}} \|\boldsymbol{\alpha}[g]\|_q, \tag{1.20}$$

where $\|.\|_q$ is either the ℓ_2 or ℓ_∞-norm. To the best of our knowledge, such a penalty appears in the early work of Grandvalet and Canu [1999] and Bakin [1999] for $q = 2$, and Turlach et al. [2005] for $q = \infty$. It has been popularized later by Yuan and Lin [2006].

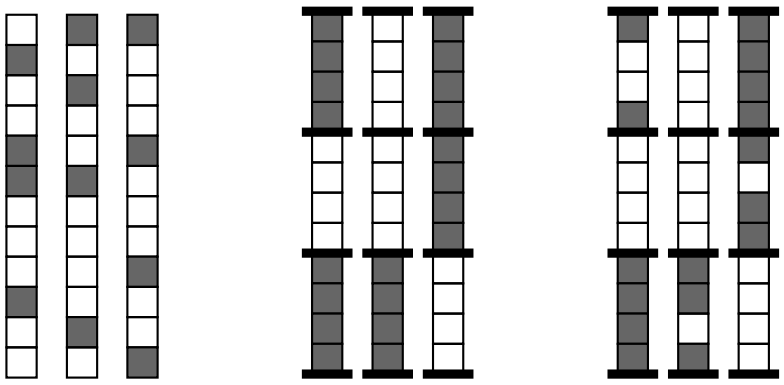
The function ψ in (1.20) is a norm, thus convex, and can be interpreted as the ℓ_1-norm of the vector $[\|\boldsymbol{\alpha}[g]\|_q]_{g \in \mathcal{G}}$ of size $|\mathcal{G}|$. Consequently, the sparsity-inducing effect of the ℓ_1-norm is applied at the group level. The penalty is highly related to the group-thresholding approach for wavelets, since the group-thresholding estimator (1.7) is linked to ψ through Eq. (1.8).

In Figure 1.13(a), we visualize the unit ball of a Group-Lasso norm obtained when $\mathcal{G}$ contains two groups $\mathcal{G} = \{\{1, 2\}, \{3\}\}$. The ball has two singularities: the top and bottom corners, corresponding to solutions where variables 1 and 2 are simultaneously set to zero, and the middle circle, corresponding to solutions where variable 3 only is set to zero. As expected, the geometry of the ball induces the group-sparsity effect.

Structured sparsity. Group-sparsity is a first step towards the more general idea that a regularization function can encourage sparse solutions with a particular structure. This notion is called *structured sparsity* and has been introduced under a large number of different point of views [Zhao et al., 2009, Jacob et al., 2009, Jenatton et al., 2011a, Baraniuk et al., 2010, Huang et al., 2011]. To some extent, it follows the concept of group-thresholding introduced in the wavelet literature, which we have presented in Section 1.2. In this paragraph, we briefly review some of these works, but for a more detailed review, we refer the reader to [Bach et al., 2012b].

Some penalties are non-convex. For instance, Huang et al. [2011] and Baraniuk et al. [2010] propose two different combinatorial approaches based on a predefined set $\mathcal{G}$ of possibly overlapping groups of variables. These penalties encourage solutions whose support is in the *union of a few number groups*, but they lead to NP-hard optimization problems. Other penalties are convex. In particular, Jacob et al. [2009] introduce a sparsity-inducing norm that is exactly a convex relaxation of the penalty of Huang et al. [2011], even though these two approaches were independently developed at the same time. As a result, the convex penalty of Jacob et al. [2009] encourages a similar structure as the one of Huang et al. [2011].

By following a different direction, the Group-Lasso penalty (1.20) has been considered when the groups are allowed to overlap [Zhao et al., 2009, Jenatton et al., 2011a]. As a consequence, variables belonging to the same groups are encouraged to be set to zero together. It was proposed for hierarchical structures by Zhao et al. [2009] with the following rule: whenever two groups g and h are in $\mathcal{G}$, they should be either disjoint, or one should be included in another. Examples of such hierarchical group structures are given in Figures 1.11 and 1.12. The effect of the penalty is to encourage sparsity patterns that are rooted subtrees. Equivalently, a variable can be non-zero only if its parent in the tree is non-zero, which is the main property of the zero-tree coding scheme introduced in the wavelet literature [Shapiro, 1993], and already illustrated in Figure 1.2.



(a) Sparsity. (b) Group sparsity. (c) Hierarchical sparsity.

Figure 1.11: Illustration of the sparsity, group sparsity, and hierarchical sparsity principles. Each column represents the sparsity pattern of a vector with 12 variables and non-zero coefficients are represented by gray squares. On the left (a), the vectors are obtained with a simple sparsity-inducing penalty, such as the ℓ_1-norm, and the non-zero variables are scattered. In the middle figure (b), a group sparsity-inducing penalty with three groups of variables is used. On the right (c), we use the hierarchical penalty consisting of the Group Lasso plus the ℓ_1-norm. Some variables within a group can be discarded.

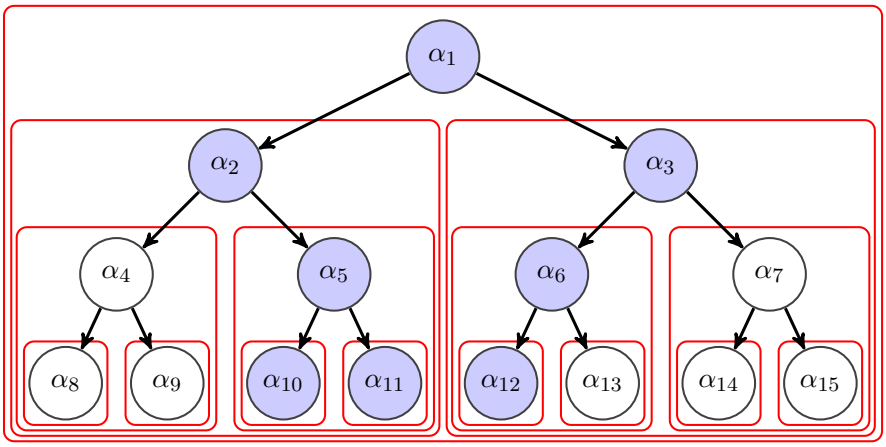


Figure 1.12: Illustration of the hierarchical sparsity of Zhao et al. [2009], which generalizes the zero-tree coding scheme of Shapiro [1993]. The groups of variables correspond to the red rectangles. The empty nodes represent variable that are set to zero. They are contained in three groups: $\{4, 8, 9\}, \{13\}, \{7, 14, 15\}$.

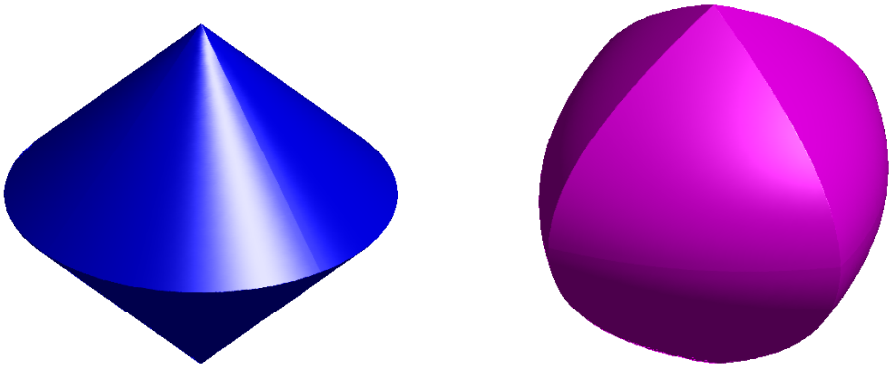
Finally, Jenatton et al. [2011a] have extended the hierarchical penalty of Zhao et al. [2009] to more general group structures, for example when variable are organized on a two-dimensional grid, encouraging neighbor variables to be simultaneously set to zero. We conclude this brief presentation of structured sparsity with Figure 1.13, where we present the unit balls of some sparsity-inducing norms. Each of them exhibits singularities and encourages particular sparsity patterns.

Spectral sparsity. Another form of parsimony has been devised in the spectral domain [Fazel et al., 2001, Srebro et al., 2005]. For estimation problems where model parameters are matrices, the rank has been used as a natural regularization function. The rank of a matrix is equal to the number of non-zero singular values, and thus, it can be interpreted as the ℓ_0-penalty of the matrix spectrum. Unfortunately, due to the combinatorial nature of ℓ_0, the rank penalization typically leads to intractable optimization problems.

A natural convex relaxation has been introduced in the control theory literature by Fazel et al. [2001] and consists of computing the ℓ_1-norm of the spectrum—that is, simply the sum of the singular values. The resulting penalty appears under different names, the most common ones being the trace, nuclear, or Schatten norm. It is defined for a matrix $\mathbf{A}$ in $\mathbb{R}^{p\times k}$ with $k \geq p$ as

$$\|\mathbf{A}\|_* \triangleq \sum_{i=1}^{p} s_i(\mathbf{A}),$$

where $s_i(\mathbf{A})$ is the i-th singular value of $\mathbf{A}$. Traditional applications of the trace norm in machine learning are matrix completion or collaborative filtering [Pontil et al., 2007, Abernethy et al., 2009]. These problems have become popular with the need of scalable recommender systems for video streaming providers. The goal is to infer movie preferences for each customer, based on their partial movie ratings. Typically, the matrix is of size $p \times k$, where p is the number of movies and k is the number of users. Each user gives a score for a few movies, corresponding to some entries of the matrix, and the recommender system tries to infer the missing values. Similar techniques have also recently



(a) Group-Lasso penalty $\psi(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}[1,2]\|_2 + |\boldsymbol{\alpha}[3]|$.

(b) Hierarchical penalty $\psi(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_2 + |\boldsymbol{\alpha}[1]| + |\boldsymbol{\alpha}[2]|$.

(c) Structured sparse penalty.

(d) Structured sparse penalty.

Figure 1.13: Visualization in three dimensions of unit balls corresponding to various sparsity-inducing norms. (a): Group Lasso penalty; (b): hierarchical penalty of Zhao et al. [2009]; (c) and (d): examples of structured sparsity-inducing penalties of Jacob et al. [2009]. Figure borrowed from Bach et al. [2012a], produced by Guillaume Obozinski.

been used in other fields, such as in genomics to infer missing genetic information [Chi et al., 2013].

1.4 Dictionary learning

We have previously presented various formulations where a signal $\mathbf{x}$ in $\mathbb{R}^m$ is approximated by a sparse linear combination of a few columns of a matrix $\mathbf{D}$ in $\mathbb{R}^{m\times p}$. In the context of signal and image processing, this matrix is often called *dictionary* and its columns *atoms*. As seen in Section 1.2, a large amount of work has been devoted in the wavelet literature for designing a good dictionary adapted to natural images.

In neuroscience, Olshausen and Field [1996, 1997] have proposed a significantly different approach to sparse modeling consisting of adapting the dictionary to training data. Because the size of natural images is too large for learning a full matrix $\mathbf{D}$, they have chosen to learn the dictionary on natural image patches, *e.g.*, of size $m = 16 \times 16$ pixels, and have demonstrated that their method could automatically discover interpretable structures. We discuss this topic in more details in Section 2.

The motivation of Olshausen and Field [1996, 1997] was to show that the structure of natural images is related to classical theories of the mammalian visual cortex. Later, dictionary learning found numerous applications in image restoration, and was shown to significantly outperform off-the-shelf bases for signal reconstruction [see, *e.g.*, Elad and Aharon, 2006, Mairal et al., 2008c, 2009, Protter and Elad, 2009, Yang et al., 2010a].

Concretely, given a dataset of n training signals $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, dictionary learning can be formulated as the following minimization problem

$$\min_{\mathbf{D}\in\mathcal{C},\mathbf{A}\in\mathbb{R}^{p\times n}} \sum_{i=1}^{n} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda\psi(\boldsymbol{\alpha}_i), \tag{1.21}$$

where $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]$ carries the decomposition coefficients of the signals $\mathbf{x}_1, \ldots, \mathbf{x}_n$, ψ is sparsity-inducing regularization function, and $\mathcal{C}$ is typically chosen as the following set:

$$\mathcal{C} \triangleq \{\mathbf{D} \in \mathbb{R}^{m\times p} : \forall j \quad \|\mathbf{d}_j\|_2 \leq 1\}.$$

To be more precise, Olshausen and Field [1996] proposed several choices for ψ; their experiments were for instance conducted with the ℓ_1-norm, or with the smooth function $\psi(\boldsymbol{\alpha}) \triangleq \sum_{j=1}^{p} \log(\varepsilon + \boldsymbol{\alpha}[j]^2)$, which has an approximate sparsity-inducing effect. The constraint $\mathbf{D} \in \mathcal{C}$ was also not explicitly modeled in the original dictionary learning formulation; instead, the algorithm of Olshausen and Field [1996] includes a mechanism to control and rescale the ℓ_2-norm of the dictionary elements. Indeed, without such a mechanism, the norm of $\mathbf{D}$ would arbitrarily go to infinity, leading to small values for the coefficients $\boldsymbol{\alpha}_i$ and making the penalty ψ ineffective.

The number of samples n is typically large, whereas the signal dimension m is small. The number of dictionary elements p is often chosen larger than m—in that case, the dictionary is said to be *overcomplete*—even though a choice $p < m$ often leads to reasonable results in many applications. For instance, a typical setting would be to have $m = 10 \times 10$ pixels for natural image patches, a dictionary of size $p = 256$, and more than 100 000 training patches.

A large part of this monograph is related to dictionary learning and thus we only briefly discuss this matter in this introduction. Section 2 is indeed devoted to unsupervised learning techniques for natural image patches, including dictionary learning; Sections 3 and 4 present a large number of applications in image processing and computer vision; how to solve (1.21) is explained in Section 5.5 about optimization.

Matrix factorization point of view. An equivalent representation of (1.21) is the following regularized matrix factorization problem

$$\min_{\mathbf{D}\in\mathcal{C},\mathbf{A}\in\mathbb{R}^{p\times n}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2 + \lambda\Psi(\mathbf{A}), \tag{1.22}$$

where $\Psi(\mathbf{A}) = \sum_{i=1}^{n} \psi(\boldsymbol{\alpha}_i)$. Even though the reformulation is a matter of using different notation, seeing dictionary learning as a matrix factorization problem opens up interesting perspectives. In particular, it makes obvious some links with other unsupervised learning approaches such as non-negative matrix factorization [Paatero and Tapper, 1994], clustering techniques, and others [see Mairal et al., 2010a]. These links will be further developed in Section 2.

Risk minimization point of view. Dictionary learning can also be seen from a machine learning point of view. Indeed, dictionary learning can be written as

$$\min_{\mathbf{D}\in\mathcal{C}} \left\{ f_n(\mathbf{D}) \triangleq \frac{1}{n}\sum_{i=1}^{n} L(\mathbf{x}_i, \mathbf{D}) \right\},$$

where $L : \mathbb{R}^m \times \mathbb{R}^{m\times p}$ is a loss function defined as

$$L(\mathbf{x}, \mathbf{D}) \triangleq \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\psi(\boldsymbol{\alpha}).$$

The quantity $L(\mathbf{x}, \mathbf{D})$ should be small if $\mathbf{D}$ is "good" at representing the signal $\mathbf{x}$ in a sparse fashion, and large otherwise. Then, $f_n(\mathbf{D})$ is called the *empirical cost*.

However, as pointed out by Bottou and Bousquet [2008], one is usually not interested in the exact minimization of the empirical cost $f_n(\mathbf{D})$ for a fixed n, which may lead to overfitting on the training data, but instead in the minimization of the *expected cost*, which measures the quality of the dictionary on new unseen data:

$$f(\mathbf{D}) \triangleq \mathbb{E}_{\mathbf{x}}[L(\mathbf{x}, \mathbf{D})] = \lim_{n\to\infty} f_n(\mathbf{D}) \quad \text{a.s.},$$

where the expectation is taken relative to the (unknown) probability distribution of the data.[5]

The expected risk minimization formulation is interesting since it paves the way to stochastic optimization techniques when a large amount of data is available [Mairal et al., 2010a] and to theoretical analysis [Maurer and Pontil, 2010, Vainsencher et al., 2011, Gribonval et al., 2013], which are developed in Sections 5.5 and 1.6, respectively.

Constrained variants. Following the original formulation of Olshausen and Field [1996, 1997], we have chosen to present dictionary learning where the regularization function is used as a penalty, even though it can also be used as a constraint as in (1.9). Then, natural variants of (1.21) are

$$\min_{\mathbf{D}\in\mathcal{C},\mathbf{A}\in\mathbb{R}^{p\times n}} \sum_{i=1}^{n} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 \quad \text{s.t.} \quad \psi(\boldsymbol{\alpha}_i) \leq \mu. \tag{1.23}$$

[5] We use "a.s." to denote almost sure convergence.

or

$$\min_{\mathbf{D}\in\mathcal{C},\mathbf{A}\in\mathbb{R}^{p\times n}} \sum_{i=1}^{n} \psi(\boldsymbol{\alpha}_i) \quad \text{s.t.} \quad \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 \leq \varepsilon. \tag{1.24}$$

Note that (1.23) and (1.24) are not equivalent to (1.21). For instance, problem (1.23) can be reformulated using a Lagrangian function [Boyd and Vandenberghe, 2004] as

$$\min_{\mathbf{D}\in\mathcal{C}} \sum_{i=1}^{n} \left(\max_{\lambda_i \geq 0} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda_i(\psi(\boldsymbol{\alpha}_i) - \mu) \right),$$

where the optimal λ_i's are not necessarily equal to each other, and their relation with the constraint parameter μ is unknown in advance. A similar discussion can be conducted for (1.24) and it is thus important in practice to choose one of the formulations (1.21), (1.23), or (1.24); the best one depends on the problem at hand and there is no general rule for preferring one instead of another.

1.5 Compressed sensing and sparse recovery

Finally, we conclude our historical tour of parsimony with recent theoretical results obtained in signal processing and statistics. We focus on methods based on the ℓ_1-norm, *i.e.*, the basis pursuit formulation of (1.9)—more results on structured sparsity-inducing norms are presented by Bach et al. [2012b].

Most analyses rely on particular assumptions regarding the problem. We start this section with a cautionary note from Hocking [1976]:

> *The problem of selecting a subset of independent or predictor variables is usually described in an idealized setting. That is, it is assumed that (a) the analyst has data on a large number of potential variables which include all relevant variables and appropriate functions of them plus, possibly, some other extraneous variables and variable functions and (b) the analyst has available "good" data on which to base the eventual conclusions. In practice, the lack of satisfaction of these assumptions may make a detailed subset selection analysis a meaningless exercise.*

In this section, we present such theoretical results where the assumptions are often not met in practice, but also results that either (1) can have an impact on the practice of sparse recovery or (2) do not need strong assumptions.

From support recovery to signal denoising. Given a signal $\mathbf{x}$ in $\mathbb{R}^m$ and a dictionary $\mathbf{D}$ in $\mathbb{R}^{m \times p}$ with ℓ_2-normalized columns, throughout this section, we assume that $\mathbf{x}$ is generated as $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}^\star + \boldsymbol{\varepsilon}$ with a sparse vector $\boldsymbol{\alpha}^\star$ in $\mathbb{R}^p$ and an additive noise $\boldsymbol{\varepsilon}$ in $\mathbb{R}^m$. For simplicity, we consider $\boldsymbol{\alpha}^\star$ and $\mathbf{D}$ as being deterministic while the noise is random, independent and identically distributed, with zero mean and finite variance σ^2.

The different formulations presented earlier in Section 1.3, for instance basis pursuit, provide estimators $\hat{\boldsymbol{\alpha}}$ of the "true" vector $\boldsymbol{\alpha}^\star$. Then, the three following goals have been studied in sparse recovery, typically in decreasing order of hardness:

- **support recovery and sign consistency**: we want the support of $\hat{\boldsymbol{\alpha}}$ (*i.e.*, the set of non-zero elements) to be the same or to be close to the one of $\boldsymbol{\alpha}^\star$. The problem is often called "model selection" in statistics and "support recovery" in signal processing; it is often refined to the estimation of the full sign pattern—that is, among the non-zero elements, we also want the correct sign to be estimated.

- **code estimation**: the distance $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\star\|_2$ should be small. In statistical terms, this correspond to the "estimation" of $\boldsymbol{\alpha}^\star$.

- **signal denoising**: regardless of code estimation, we simply want the distance $\|\mathbf{D}\hat{\boldsymbol{\alpha}} - \mathbf{D}\boldsymbol{\alpha}^\star\|_2$ to be small; the goal is not to obtain exactly $\boldsymbol{\alpha}^\star$, but simply to obtain a good denoised version of the signal $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}^\star + \boldsymbol{\varepsilon}$.

A good code estimation performance does imply a good denoising performance but the converse is not true in general. In most analyses, support recovery is harder than code estimation. As detailed below, the sufficient conditions for good support recovery lead indeed to good estimation.

High-dimensional phenomenon. Without sparsity assumptions, even the simplest denoising task can only achieve denoising errors of the order $\frac{1}{n}\|\mathbf{D}\hat{\boldsymbol{\alpha}} - \mathbf{D}\boldsymbol{\alpha}^\star\|_2^2 \approx \frac{\sigma^2 p}{m}$, which is attained for ordinary least-squares, and is the best possible [Tsybakov, 2003]. Thus, in order to have at least a good denoising performance (prediction performance in statistics), either the noise σ is small, or the signal dimension m (the number of samples) is much larger than the number of atoms p (the number of variables to select from).

When making the assumption that the true code $\boldsymbol{\alpha}^\star$ is sparse with at most k non zeros, smaller denoising errors can be obtained. In that case, it is possible indeed to replace the scaling $\frac{\sigma^2 p}{m}$ by $\frac{\sigma^2 k \log p}{m}$. Thus, even when p is much larger than m, as long as $\log p$ is much smaller than m, we may have good prediction performance. However, this high-dimensional phenomenon currently[6] comes at a price: (1) either an exhaustive search over the subsets of size k needs to be performed [Massart, 2003, Bunea et al., 2007, Raskutti et al., 2011] or (2) some assumptions have to be made regarding the dictionary $\mathbf{D}$, which we now describe.

Sufficient conditions for high-dimensional fast rates. Most sufficient conditions have the same flavor. A dictionary behaves well if the off-diagonal elements of $\mathbf{D}^\top\mathbf{D}$ are small, in other words, if there is little correlation between atoms. However, the notion of coherence (the maximal possible correlation between two atoms) was the first to emerge [see, *e.g.*, Elad and Bruckstein, 2002, Gribonval and Nielsen, 2003], but it is not sufficient to obtain a high-dimensional phenomenon.

In the noiseless setting, Candes and Tao [2005] and Candès et al. [2006] introduced the *restricted isometry property* (RIP), which states that all submatrices of size $k \times k$ of $\mathbf{D}^\top\mathbf{D}$ should be close to isometries, that is, should have all of their eigenvalues sufficiently close to one. With such an assumption, the Lasso behaves well: it recovers the true support and estimates the code $\boldsymbol{\alpha}^\star$ and the signal $\mathbf{D}\boldsymbol{\alpha}^\star$ with an error of order $\frac{\sigma^2 k \log p}{m}$.

[6] Note that recent research suggests that this fast rate of $\frac{\sigma^2 k \log p}{m}$ cannot be achieved by polynomial-time algorithms [Zhang et al., 2014].

The main advantage of the RIP assumption is that one may exhibit dictionaries for which it is satisfied, usually obtained by normalizing a matrix $\mathbf{D}$ obtained from independent Gaussian entries, which may satisfy the condition that $(k \log p)/m$ remains small. Thus, the sufficient conditions are not vacuous. However, the RIP assumption has two main drawbacks: first, it cannot be checked on a given dictionary $\mathbf{D}$ without checking all $O(p^k)$ submatrices of size k; second, it may be weakened if the goal is support recovery or simply estimation (code recovery).

There is therefore a need for sufficient conditions that can be checked in polynomial time while ensuring sparse recovery. However, none currently exists with the same scalings between k, p and m [see, *e.g.*, Juditsky and Nemirovski, 2011, d'Aspremont and El Ghaoui, 2011]. When refining to support recovery, Fuchs [2005], Tropp [2004], Wainwright [2009] provide sufficient and necessary conditions of a similar flavor than requiring that all submatrices of size k are sufficiently close to orthogonal. For the tightest conditions, see, *e.g.*, Bühlmann and Van De Geer [2011]. Note that these conditions are also typically sufficient for algorithms that are not based explicitly on convex optimization [Tropp, 2004].

Finally, it is important to note that (a) most of the theoretical results advocate a value for the regularization parameter λ proportional to $\sigma\sqrt{m \log p}$, which unfortunately depends on the noise level σ (which is typically unknown in practice), and that (b) for orthogonal dictionaries, all of these assumptions are met; however, this imposes $p = m$.

Compressed sensing vs. statistics. Our earlier quote from Hocking [1976] applies to sparse estimation as used in statistics for least-squares regression, where the dictionary $\mathbf{D}$ is simply the input data and $\mathbf{x}$ the output data. In most situations, there are some variables, represented by columns of $\mathbf{D}$, that are heavily correlated. Therefore, in most practical situations, the assumptions do not apply. However, it does not mean that the high-dimensional phenomenon does not apply in a weaker sense (see the next paragraph for slow rates); moreover it is important to remark that there are other scenarios, beyond statistical variable selection, where the dictionary $\mathbf{D}$ may be chosen.

In particular, in signal processing, the dictionary $\mathbf{D}$ may be seen as *measurements*—that is, we want to encode $\boldsymbol{\alpha}^\star$ in $\mathbb{R}^p$ using m linear measurements $\mathbf{D}\boldsymbol{\alpha}^\star$ in $\mathbb{R}^m$ for m much larger than p. What the result of Candes and Tao [2005] alluded to earlier shows is that for random measurements, one can recover a k-sparse $\boldsymbol{\alpha}^\star$ from (a potentially noisy version of) $\mathbf{D}\boldsymbol{\alpha}^\star$, with overwhelming probability, as long as $(k \log p)/m$ remains small. This is the core idea behind compressive sensing. See more details from Donoho [2006], Candès and Wakin [2008].

High-dimensional slow rates. While sufficient conditions presented earlier are often not met beyond random dictionaries, for the basis pursuit/Lasso formulation, the high-dimensional phenomenon may still be observed, but only for the denoising situation and with a weaker result. As shown by Greenshtein [2006] and Bühlmann and Van De Geer [2011, Corollary 6.1], *without assumptions regarding correlations*, we have $\frac{1}{m}\|\mathbf{D}\hat{\boldsymbol{\alpha}} - \mathbf{D}\boldsymbol{\alpha}^*\|_2^2 \approx \sqrt{\frac{\sigma^2 k^2 \log p}{m}}$. Note that this slower rate does not readily extend to non-convex formulations.

Impact on dictionary learning. The dictionary learning framework which we describe in this monograph relies on sparse estimation, that is, given the dictionary $\mathbf{D}$, the estimation of the code $\boldsymbol{\alpha}$ may be analyzed using the tools we have presented in this section. However, the dictionaries that are learned do not exhibit low correlations between atoms and thus theoretical results do not apply (see dedicated results in the next section). However, they suggest that (a) the codes $\boldsymbol{\alpha}$ may not be unique in general and caution has to be observed when representing a signal $\mathbf{x}$ by its code $\boldsymbol{\alpha}$, (b) methods based on ℓ_1-penalization are more robust as they still provably perform denoising in presence of strong correlations and (c) incoherence promoting may be used in order to obtain better-behaved dictionaries [see Ramirez et al., 2009].

1.6 Theoretical results about dictionary learning

Dictionary learning, as formulated in Eq. (1.21), may be seen from several perspectives, mainly as an unsupervised learning or a matrix

factorization problem. While the supervised learning problem from the previous section (sparse estimation of a single signal given the dictionary) comes with many theoretical analyses, there are still few theoretical results of the same kind for dictionary learning. In this section, we present some of them. For simplicity, we assume that we penalize with the ℓ_1-norm and consider the minimization of

$$\min_{\mathbf{D}\in\mathcal{C},\mathbf{A}\in\mathbb{R}^{p\times n}} \sum_{i=1}^{n} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1, \tag{1.25}$$

where $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]$ carries the decomposition coefficients of the signals $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and $\mathcal{C}$ is chosen as the following set:

$$\mathcal{C} \triangleq \{\mathbf{D} \in \mathbb{R}^{m\times p} : \forall j \quad \|\mathbf{d}_j\|_2 \leq 1\}.$$

Non-convex optimization problem. After imposing parsimony through the ℓ_1-norm, given $\mathbf{D}$ the objective function is convex in $\boldsymbol{\alpha}$, given $\boldsymbol{\alpha}$ the objective and constraints are convex in $\mathbf{D}$. However, the objective function is not jointly convex, which is typical of unsupervised learning formulations. Hence, we consider an optimization problem for which it is not possible in general to guarantee that we are going to obtain the global minimum; the same applies to EM-based approaches [Dempster et al., 1977] or K-means [see, *e.g.*, Bishop, 2006].

Symmetries. Worse, the problem in Equation (1.25) exhibits several symmetries and admits multiple global optima, and the descent methods that are described in Section 5 will also have the same invariance property. For example, the columns of $\mathbf{D}$ and rows of $\mathbf{A}$ can be submitted to $p!$ arbitrary (but consistent) permutations. There are also sign ambiguities: in fact, if $(\mathbf{D}, \mathbf{A})$ is solution of (1.25), so is $(\mathbf{D}\mathrm{diag}(\boldsymbol{\varepsilon}), \mathrm{diag}(\boldsymbol{\varepsilon})\mathbf{A})$, where $\boldsymbol{\varepsilon}$ is a vector in $\{-1, +1\}^p$ that carries a sign pattern. Therefore, for every one of the $p!$ possible atom orders, the dictionary learning problem admits 2^p equivalent solutions. In other words, for a solution $(\mathbf{D}, \mathbf{A})$, the pair $(\mathbf{D}\boldsymbol{\Gamma}, \boldsymbol{\Gamma}^{-1}\mathbf{A})$ is also solution, where $\boldsymbol{\Gamma}$ is a *generalized permutation* formed by the product of a diagonal matrix with $+1$ and -1's on its diagonal with a permutation matrix (in particular, $\boldsymbol{\Gamma}$ is thus orthogonal).

The fact that there are no other transformations $\mathbf{\Gamma}$ such that $(\mathbf{D}\mathbf{\Gamma}, \mathbf{\Gamma}^{-1}\mathbf{A})$ is also solution of Eq. (1.25) for *all* solutions $(\mathbf{D}, \mathbf{A})$ of this problem follows from a general property of isometries of the ℓ_q norm for finite values of q such that $q \geq 1$ and $q \neq 2$ [Li and So, 1994].

A manifold interpretation of sparse coding with projective geometry. The interpretation of sparse coding as a *locally* linear representation of a non-linear "manifold" is problematic because certain signals/features are best thought of as "points" in some space rather than vectors. For example, what does it mean to "add" two natural image patches? The simplest point structure that one can think of is affine or projective, and we show below that sparse coding indeed admits a natural interpretation in this setting, at least for normalized signals.

Indeed, Let us restrict our attention from now on to unit-norm signals, as is customary in image processing after the usual centering and normalization steps, which will be studied in Section 2.1.[7] Note that the dictionary elements $\mathbf{d}_j$ in a solution $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_p]$ of Eq. (1.25) also have unit norm by construction.

Let us now consider the "half sphere"

$$\mathbb{S}_+^{m-1} \triangleq \left\{ \mathbf{d} \in \mathbb{S}^{m-1} : \text{the first non-zero coefficient of } \mathbf{d} \text{ is positive} \right\}, \tag{1.26}$$

where $\mathbb{S}^{m-1}$ is the unit sphere of dimension $m-1$ formed by the unit vectors of $\mathbb{R}^m$.[8] A direct consequence of the sign ambiguities of dictionary learning discussed in the previous paragraph is that, for any solution $(\mathbf{D}, \mathbf{A})$ of Eq. (1.25), there is an equivalent solution $(\mathbf{D}', \mathbf{A}')$ with all columns of $\mathbf{D}'$ in $\mathbb{S}_+^{m-1}$. Indeed, suppose some column $\mathbf{d}_j$ is not in $\mathbb{S}_+^{m-1}$, and let $\mathbf{d}_j[i]$ be its first non-zero coefficient (which is necessarily negative since $\mathbf{d}_j \notin \mathbb{S}_+^{m-1}$). We can replace $\mathbf{d}_j$ by $-\mathbf{d}_j$ and the corresponding row of the matrix $\mathbf{A}$ by its opposite to construct an equivalent minimum of the dictionary learning problem in $\mathbb{S}_+^{m-1} \times \mathbb{R}^{p \times n}$.

[7] Note that the fact that the individual signals are centered does not imply that the dictionary elements are.

[8] Similarly, one may define the set $\mathbb{S}_-^{m-1}$ by replacing "positive" by "non-negative" in (1.26). The two sets $\mathbb{S}_+^{m-1}$ and $\mathbb{S}_-^{m-1}$ form a partition of $\mathbb{S}^{m-1}$ with equal volume, and indeed, each one geometrically corresponds to a half sphere.

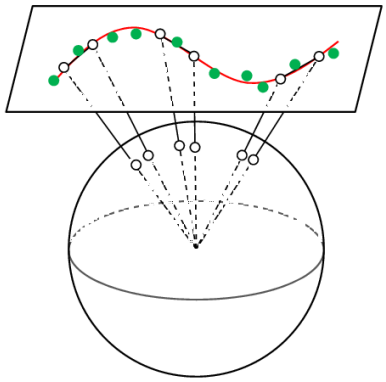


Figure 1.14: An illustration of the projective interpretation of sparse coding.

Likewise, we can restrict the signals $\mathbf{x}_i$ to lie in $\mathbb{S}_+^{m-1}$ since replacing $\mathbf{x}_i$ by its opposite for a given dictionary simply amounts to replacing the code $\boldsymbol{\alpha}_i$ by its opposite. Note that this identifies a patch with its "negative", but remember that the sign of its code elements is not uniquely defined in the first place in conventional dictionary learning settings (it is uniquely defined if we insist that the dictionary elements belong to $\mathbb{S}_+^{m-1}$). This allows us to identify both the dictionary elements and the signals with points in the projective space $\mathbb{P}^{m-1} = P(\mathbb{R}^m)$. Any k independent column vectors of $\mathbf{D}$ define a $(k-1)$-dimensional projective subspace of $\mathbb{P}^{m-1}$ (see Figure 1.14).

In particular, if the data signals are assumed to be sampled from a "noisy manifold" of dimension $k-1$ embedded in $\mathbb{P}^{m-1}$, an approximation of some sample $\mathbf{x}$ by a sparse linear combination of k elements of $\mathbf{D}$ can be thought of as lying in (or near) the $k-1$ dimensional "tangent plane" there.

Consistency results. Given the dictionary learning problem from a finite number of signals, there are several interesting theoretical questions to be answered. The first natural question is to understand the

properties of the cost function that is minimized when the number of signals tends to infinity, and in particular how it converges to the expectation under the signal generating distribution [Vainsencher et al., 2011, Maurer and Pontil, 2010]. Then, given the non-convexity of the optimization problems, local consistency results may be obtained, by showing that the cost function which is minimized has a local minimum around the pairs $(\mathbf{D}^\star, \mathbf{A}^\star)$ that has generated the data. Given RIP-based assumptions on the dictionary $\mathbf{D}^\star$ and number of non zero elements in the columns of $\mathbf{A}^\star$, and the noise level, Gribonval et al. [2014] show that the cost function defined in Eq. (1.25) has a local minimum around $(\mathbf{D}^\star, \mathbf{A}^\star)$ with high probability, as long as the number of signals n is greater than a constant times mp^3. In the noiseless case, earlier results have been also obtained Gribonval and Schnass [2010], Geng et al. [2011], and recently it has been shown that under additional assumptions, a good initializer could be found so that the previous type of local consistency results can be applied [Agarwal et al., 2013].

Finally, recent algorithms have emerged in the theoretical science community, which are not explicitly based on optimization [see, *e.g.*, Spielman et al., 2013, Recht et al., 2012, Arora et al., 2014]. These come with global convergence guarantees (with additional assumptions regarding the signals), but their empirical performance on concrete signal and image processing problems have not yet been demonstrated.

References

J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. A new approach to collaborative filtering: operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10:803–826, 2009.

A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries via alternating minimization. *preprint arXiv:1310.7991*, 2013.

M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, 100(1):90–93, 1974.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, volume 1, pages 267–281, 1973.

S. Arora, R. Ge, and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2014.

S. P. Awate and R. T. Whitaker. Unsupervised, information-theoretic, adaptive image filtering for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):364–376, 2006.

D. Baby, T. Virtanen, T. Barker, and H. Van hamme. Coupled dictionary training for exemplar-based speech enhancement. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundation and Trends in Machine Learning*, 4:1–106, 2012a.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012b.

S. Bakin. *Adaptive regression and model selection in data mining problems*. PhD thesis, 1999.

R. G. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

E. Barillot, L. Calzone, P. Hupe, J.-P. Vert, and A. Zinovyev. *Computational systems biology of cancer*. CRC Press, 2012.

A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6): 2743–2760, 1998.

M. J. Bayarri and J. O Berger. The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine learning*, 56(1-3):209–239, 2004.

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

A. J. Bell and T. J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the ACM SIGGRAPH Conference*, 2000.

D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999. 2nd edition.

P. J. Bickel. Parametric robustness: small biases can be worthwhile. *The Annals of Statistics*, 12(4):864–879, 1984.

C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

L. Bo, X. Ren, and D. Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

L. Bo, X. Ren, and D. Fox. Multipath sparse coding using hierarchical matching pursuit. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

J.M. Borwein and A.S. Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer, 2006.

T. Bossomaier and A. W. Snyder. Why spatial frequency processing in the visual cortex? *Vision Research*, 26(8):1307–1309, 1986.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

Y-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

Y-L. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun. Ask the locals: Multi-way local pooling for image recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

D. M. Bradley and J. A. Bagnell. Differential sparse coding. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

P. Brucker. An $O(n)$ algorithm for quadratic knapsack problems. *Operations Research Letters*, 3(3):163–166, 1984.

J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013.

O. Bryt and M. Elad. Compression of facial images using the K-SVD algorithm. *Journal of Visual Communication and Image Representation*, 19(4): 270–282, 2008.

A. Buades, B. Coll, and J.-M. Morel. A review of image denoising algorithms, with a new one. *SIAM Journal on Multiscale Modeling and Simulation*, 4 (2):490–530, 2005.

A. Buades, B. Coll, J.-M. Morel, and C. Sbert. Self-similarity driven color demosaicking. *IEEE Transactions on Image Processing*, 18(6):1192–1202, 2009.

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

T. T. Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Annals of Statistics*, 27(3):898–924, 1999.

E. J. Candès and D. L. Donoho. Recovering edges in ill-posed inverse problems: Optimality of curvelet frames. *Annals of Statistics*, 30(3):784–842, 2002.

E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

E.J. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.

M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust. Do we know what the early visual system does? *The Journal of Neuroscience*, 25(46):10577–10597, 2005.

J.-F. Cardoso. Dependence, correlation and Gaussianity in independent component analysis. *Journal of Machine Learning Research*, 4:1177–1203, 2003.

A. Castrodad and G. Sapiro. Sparse modeling of human actions from motion imagery. *International Journal of Computer Vision*, 100(1):1–15, 2012.

A. Chambolle. Total variation minimization and a class of binary MRF models. In *Proceedings of the 5th International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2005.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

S. G. Chang, B. Yu, and M. Vetterli. Spatially adaptive wavelet thresholding with context modeling for image denoising. *IEEE Transactions on Image Processing*, 9(9):1522–1531, 2000a.

S. G. Chang, B. Yu, and M. Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing*, 9(9): 1532–1546, 2000b.

A. S. Charles, B. A. Olshausen, and C. J. Rozell. Learning sparse codes for hyperspectral imagery. *IEEE Journal of Selected Topics in Signal Processing*, 5(5):963–978, 2011.

P. Chatterjee and P. Milanfar. Patch-based near-optimal image denoising. *IEEE Transactions on Image Processing*, 21(4):1635–1649, 2012.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1999.

Y. Chen, J. Mairal, and Z. Harchaoui. Fast and robust archetypal analysis for representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

E. C. Chi, H. Zhou, G. K. Chen, D. O. Del Vecchyo, and K. Lange. Genotype imputation via matrix completion. *Genome research*, 23(3):509–518, 2013.

J. F. Claerbout and F. Muir. Robust modeling with erratic data. *Geophysics*, 38(5):826–844, 1973.

A. Coates, A. Y. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

R. R. Coifman and D. L. Donoho. Translation-invariant de-noising. In *Lecture Notes in Statistics*, volume 103, pages 125–150. 1995.

P. L. Combettes and J.-C. Pesquet. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, chapter Proximal Splitting Methods in Signal Processing. Springer, 2011.

L. Condat. A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.

S. F. Cotter, J. Adler, B. Rao, and K. Kreutz-Delgado. Forward sequential algorithms for best basis selection. In *IEEE Proceedings of Vision Image and Signal Processing*, pages 235–244, 1999.

F. Couzinie-Devy, J. Mairal, F. Bach, and J. Ponce. Dictionary learning for deblurring and digital zoom. *preprint arXiv:1110.0957*, 2011.

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley and Sons, 2006. 2nd edition.

A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004.

M. S. Crouse, R. D. Nowak, and R. G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing*, 46(4):886–902, 1998.

G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proceedings of the workshop on statistical learning in computer vision, ECCV*, 2004.

A. Cutler and L. Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007a.

K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Color image denoising via sparse 3D collaborative filtering with grouping constraint in luminance-chrominance space. In *IEEE International Conference on Image Processing (ICIP)*, 2007b.

K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. BM3D image denoising with shape-adaptive principal component analysis. In *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

A. d'Aspremont and L. El Ghaoui. Testing the nullspace property using semidefinite programming. *Mathematical Programming*, 127(1):123–144, 2011.

I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 41(7):909–996, 1988.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010.

J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169, 1985.

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

M. N. Do and M. Vetterli. The contourlet transform: an efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14(12):2091–2106, 2005.

W. Dong, L. Zhang, and G. Shi. Centralized sparse representation for image restoration. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011a.

W. Dong, L. Zhang, G. Shi, and X. Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011b.

W. Dong, L. Zhang, G. Shi, and X. Li. Nonlocally centralized sparse representation for image restoration. *IEEE Transactions on Image Processing*, 22(4):1620–1630, 2013.

D. L. Donoho. Wedgelets: Nearly minimax estimation of edges. *Annals of Statistics*, 27(3):859–897, 1999.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432): 1200–1224, 1995.

D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l1-ball for learning in high dimensions. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

B. Efron and C. Morris. Limiting the risk of Bayes and empirical Bayes estimators — part I: the Bayes case. *Journal of the American Statistical Association*, 66(336):807–815, 1971.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 1999.

M. A. Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, 9(1):191–203, 1960.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868, 1998.

M. Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 48(9):2558–2567, 2002.

E. Elhamifar and R. Vidal. Sparse subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

K. Engan, S. O. Aase, and J. H. Husoy. Method of optimal directions for frame design. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96 (456):1348–1360, 2001.

O. D. Faugeras. Digital color image processing within the framework of a human visual model. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(4):380–393, 1979.

M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference*, volume 6, pages 4734–4739, 2001.

M. Fazel, H. Hindi, and S. P. Boyd. Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. In *Proceedings of the American Control Conference*, volume 3, pages 2156–2162, 2003.

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4 (12):2379–2394, 1987.

M. A. T. Figueiredo and R. D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8):906–916, 2003.

M. A. T. Figueiredo and R. D. Nowak. A bound optimization approach to wavelet-based image deconvolution. In *Proceedings of the International Conference on Image Processing (ICIP)*, 2005.

M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak. Majorization-minimization algorithms for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 16(12):2980–2991, 2007.

R. W. Floyd and L. Steinberg. An adaptive algorithm for spatial grey scale. In *Proceedings of the Society of Information Display*, volume 17, pages 75–77, 1976.

D. A. Forsyth and J. Ponce. *Computer vision: a modern approach*. Prentice Hall, 2012. 2nd edition.

I. E Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9): 891–906, 1991.

W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65, 2002.

J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981.

W.J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

J. J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Image Processing*, 51(10):3601–3608, 2005.

G. M. Furnival and R. W. Wilson. Regressions by leaps and bounds. *Technometrics*, 16(4):499–511, 1974.

D. Gabor. Theory of communication. *Journal of the Institution of Electrical Engineers*, 93(26):429–441, 1946.

S. Gao, I. W.-H. Tsang, L.-T. Chia, and P. Zhao. Local features are not lonely–Laplacian sparse coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

S. Gao, W.-H Tsang, and L.-T. Chia. Laplacian sparse coding, hypergraph Laplacian sparse coding, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):92–104, 2013.

P. Garrigues and B. A. Olshausen. Group sparse coding with a Laplacian scale mixture prior. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

G. Gasso, A. Rakotomamonjy, and S. Canu. Recovering sparse signals with non-convex penalties and DC programming. *IEEE Transactions on Signal Processing*, 57(12):4686–4698, 2009.

Q. Geng, H. Wang, and J. Wright. On the local correctness of ℓ_1-minimization for dictionary learning. *preprint arXiv:1101.5672*, 2011.

A. Gersho and R. M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, 1992.

M. Gharavi-Alkhansari and T. S. Huang. A fast orthogonal matching pursuit algorithm. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.

R. Giryes and M. Elad. Sparsity based Poisson denoising with dictionary learning. *IEEE Transactions on Image Processing*, 2014. to appear.

D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.

G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. Johns Hopkins University Press, 2012. 4th edition.

Y. Grandvalet and S. Canu. Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In *Advances in Neural Information Processing Systems (NIPS)*, 1999.

H. Grassmann. LXXXVII. On the theory of compound colours. *Philosophical Magazine Series 4*, 7(45):254–264, 1854.

K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2005.

E. Greenshtein. Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1 constraint. *The Annals of Statistics*, 34(5):2367–2386, 2006.

K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

K. Gregor, A. Szlam, and Y. LeCun. Structured sparse coding via lateral inhibition. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12):3320–3325, 2003.

R. Gribonval and K. Schnass. Dictionary identification–sparse matrix-factorization via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

R. Gribonval, V. Cevher, and M. E. Davies. Compressible distributions for high-dimensional statistics. *IEEE Transactions on Information Theory*, 58 (8):5016–5034, 2012.

R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert. Sample complexity of dictionary learning and other matrix factorizations. *preprint arXiv:1312.3790*, 2013.

R. Gribonval, R. Jenatton, and F. Bach. Sparse and spurious: dictionary learning with noise and outliers. *preprint arXiv:1407.5155*, 2014.

T. Guha and R. K. Ward. Learning sparse representations for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1576–1588, 2012.

A. Haar. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 69(3):331–371, 1910.

P. Hall, G. Kerkyacharian, and D. Picard. On the minimax optimality of block thresholded wavelet estimators. *Statistica Sinica*, 9(1):33–49, 1999.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2009. 2nd edition.

S. Hawe, M. Seibert, and M. Kleinsteuber. Separable dictionary learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

J. Hérault, C. Jutten, and B. Ans. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In *Actes du Xème Colloque GRETSI*, 1985.

K. K. Herrity, A. C. Gilbert, and J. A. Tropp. Sparse approximation via iterative thresholding. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2006.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

R. R. Hocking. A Biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*, 32:1–49, 1976.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

J. Huang and T. Zhang. The benefit of group sparsity. *Annals of Statistics*, 38(4):1978–2004, 2010.

J. Huang, Z. Zhang, and D. Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research*, 12:3371–3412, 2011.

K. Huang and S. Aviyente. Sparse representation for signal classification. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural computation*, 13(7):1527–1558, 2001.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. John Wiley and Sons, 2004.

A. Hyvärinen, J. Hurri, and P. O. Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer, 2009.

L. Jacob, G. Obozinski, and J.-P. Vert. Group Lasso with overlaps and graph Lasso. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.

H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, 87(3):316–336, 2010.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010a.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *Proceedings of International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2010b.

R. Jenatton, J-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011a.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011b.

A. Juditsky and A. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via ℓ_1-minimization. *Mathematical Programming*, 127(1): 57–88, 2011.

V. Katkovnik, A. Foi, K. Egiazarian, and J. Astola. From local kernel to nonlocal multiple-model image denoising. *International Journal of Computer Vision*, 86(1):1–32, 2010.

K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun. Learning invariant features through topographic filter maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

K. Kavukcuoglu, M. Ranzato, and Y. LeCun. Fast inference in sparse coding algorithms with applications to object recognition. *preprint arXiv:1010.3467*, 2010a.

K. Kavukcuoglu, P. Sermanet, Y-L. Boureau, K. Gregor, M. Mathieu, and Y. LeCun. Learning convolutional feature hierarchies for visual recognition. In *Advances in Neural Information Processing Systems (NIPS)*, 2010b.

K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

C. Kervrann and J. Boulanger. Optimal spatial adaptation for patch-based image denoising. *IEEE Transactions on Image Processing*, 15(10):2866–2878, 2006.

R. Kimmel. Demosaicing: image reconstruction from color ccd samples. *IEEE Transactions on Image Processing*, 8(9):1221–1228, 1999.

J. Koenderink and A. Van Doorn. The structure of locally orderless images. *International Journal of Computer Vision*, 31(2/3):159–168, 1999.

P. Koniusz and K. Mikolajczyk. Spatial coordinate coding to reduce histogram representations, dominant angle and colour pyramid match. In *Proceedings of the International Conference on Image Processing (ICIP)*, 2011.

P. Koniusz, F. Yan, and K. Mikolajczyk. Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection. *Computer Vision and Image Understanding*, 117(5):479–492, 2013.

H. J. Kushner and G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.

K. Lange, D. R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9 (1):1–20, 2000.

I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

E. Le Pennec and S. Mallat. Sparse geometric image representations with bandelets. *IEEE Transactions on Image Processing*, 14(4):423–438, 2005.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998a.

Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop. In G. Orr and Muller K., editors, *Neural Networks: Tricks of the trade*. Springer, 1998b.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

H. Lee, A. Battle, R. Raina, and A.Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

A. Levin, B. Nadler, F. Durand, and W. T. Freeman. Patch complexity, finite pixel correlations and optimal denoising. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.

M. S. Lewicki. Efficient coding of natural sounds. *Nature neuroscience*, 5(4): 356–363, 2002.

M. S. Lewicki and B. A. Olshausen. Probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A*, 16(7):1587–1601, 1999.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

C.-K. Li and W. So. Isometries of the ℓ_p norm. *The American Mathematical Monthly*, 101(5):452–453, 1994.

Y. Li and D. P. Huttenlocher. Sparse long-range random field and its application to image denoising. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008.

M. S. Livingstone and D. H. Hubel. Anatomy and physiology of a color system in the primate visual cortex. *The Journal of Neuroscience*, 4(1):309–356, 1984.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th Joint Conference on Artificial Intelligence (IJCAI)*, 1981.

D. J. C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge university press, 2003.

M. Mahmoudi and G. Sapiro. Fast image and video denoising via nonlocal means of similar neighborhoods. *IEEE Signal Processing Letters*, 12(12): 839–842, 2005.

J. Mairal. *Sparse coding for machine learning, image processing and computer vision*. PhD thesis, Ecole Normale Supérieure de Cachan, 2010. `http://tel.archives-ouvertes.fr/tel-00595312`.

J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

J. Mairal and B. Yu. Complexity analysis of the Lasso regularization path. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008a.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008b.

J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, 2008c.

J. Mairal, M. Leordeanu, F. Bach, M. Hebert, and J. Ponce. Discriminative sparse image models for class-specific edge detection and image interpretation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008d.

J. Mairal, G. Sapiro, and M. Elad. Learning multiscale sparse representations for image and video restoration. *SIAM Multiscale Modeling and Simulation*, 7(1):214–241, 2008e.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010a.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems (NIPS)*, 2010b.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Convex and network flow optimization for structured sparsity. *Journal of Machine Learning Research*, 12:2681–2720, 2011.

J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012.

S. Mallat. *A wavelet tour of signal processing*. Academic press, 2008. 3rd edition.

S. Mallat and Z. Zhang. Matching pursuit in a time-frequency dictionary. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

C. L. Mallows. Choosing variables in a linear regression: A graphical aid. unpublished paper presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas, 1964.

C. L. Mallows. Choosing a subset regression. unpublished paper presented at the Joint Statistical Meeting, Los Angeles, California, 1966.

H. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

D. Martin, C. Fowlkes, Doron Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2001.

D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004.

P. Massart. *Concentration Inequalities and Model Selection: Ecole d'été de Probabilités de Saint-Flour 23*. Springer, 2003.

I. Matthews, T. Ishikawa, and S. Baker. The template update problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):810–815, 2004.

A. Maurer and M. Pontil. k-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

J. C. Maxwell. On the theory of compound colours, and the relations of the colours of the spectrum. *Philosophical Transactions of the Royal Society of London*, pages 57–84, 1860.

X. Mei and H. Ling. Robust visual tracking using ℓ_1-minimization. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

D. Menon and G. Calvagno. Color image demosaicking: an overview. *Signal Processing: Image Communication*, 26(8):518–533, 2011.

C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. In *Journal of Machine Learning Research*, volume 6, pages 1099–1125, 2005.

K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10): 1615–1630, 2005.

N. Murray and F. Perronnin. Generalized max pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141, 1964.

I. Naseem, R. Togneri, and M. Bennamoun. Linear regression for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):2106–2112, 2010.

N. M. Nasrabadi and R. A. King. Image coding using vector quantization: A review. *IEEE Transactions on Communications*, 36(8):957–971, 1988.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.

J. Nathans, D. Thomas, and D. S. Hogness. Molecular genetics of human color vision: the genes encoding blue, green, and red pigments. *Science*, 232(4747):193–202, 1986.

Y. Nesterov. Gradient methods for minimizing composite objective function. *Mathematical Programming*, 140(1):125–161, 2013.

I. Newton. Hypothesis explaining the properties of light. In *The History of the Royal Society*, volume 3, pages 247–269. T. Birch, 1675. text published in 1757.

S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011.

R. D. Nowak and M. A. T. Figueiredo. Fast wavelet-based image deconvolution using the EM algorithm. In *Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems and Computers.*, 2001.

G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2009.

T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

B. A. Olshausen and D. J. Field. How close are we to understanding V1? *Neural computation*, 17(8):1665–1699, 2005.

M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20 (3):389–403, 2000.

P. Paatero and U. Tapper. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, 1993.

J. Pearl. On coding and filtering stationary signals by discrete fourier transforms (corresp.). *IEEE Transactions on Information Theory*, 19(2):229–232, 1973.

F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.

G. Peyré. Sparse modeling of textures. *Journal of Mathematical Imaging and Vision*, 34(1):17–31, 2009.

D.-S. Pham and S. Venkatesh. Joint learning and dictionary construction for pattern recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

D.-T. Pham. Fast algorithms for mutual information based independent component analysis. *IEEE Transactions on Signal Processing*, 52(10):2690–2700, 2004.

N. Pinto, D. D. Cox, and J. J. DiCarlo. Why is real-world visual object recognition hard? *PLoS computational biology*, 4(1):151–156, 2008.

J. Pokrass, A. M. Bronstein, M. M. Bronstein, P. Sprechmann, and G. Sapiro. Sparse modeling of intrinsic correspondences. In *Computer Graphics Forum*, volume 32, pages 459–468, 2013.

M. Pontil, A. Argyriou, and T. Evgeniou. Multi-task feature learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, 12(11):1338–1351, 2003.

W. Pratt. Spatial transform coding of color images. *IEEE Transactions on Communication Technology*, 19(6):980–992, 1971.

M. Protter and M. Elad. Image sequence denoising via sparse and redundant representations. *IEEE Transactions on Image Processing*, 18(1):27–35, 2009.

H. Raguet, J. Fadili, and G. Peyré. A generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, 6(3):1199–1226, 2013.

R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2007.

I. Ramirez, F. Lecumberry, and G. Sapiro. Sparse modeling with universal priors and learned incoherent dictionaries. In *Proceedings of the 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2009.

I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

M. Ranzato, F. J. Huang, Y-L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

B. Recht, C. Re, J. Tropp, and V. Bittorf. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

X. Ren and D. Ramanan. Histograms of sparse codes for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

R. Rigamonti, M. A. Brown, and V. Lepetit. Are sparse representations really relevant for image classification? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

R. Rigamonti, A. Sironi, V. Lepetit, and P. Fua. Learning separable filters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

K. Ritter. Ein verfahren zur lösung parameterabhängiger, nichtlinearer maximum-probleme. *Mathematical Methods of Operations Research*, 6(4): 149–166, 1962.

F. Rodriguez and G. Sapiro. Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries. Technical report, University of Minnesota, 2008.

Y. Romano, M. Protter, and M. Elad. Single image interpolation via adaptive non-local sparsity-based modeling. *IEEE Transactions on Image Processing*, 23(7):3085–3098, 2014.

S. Roth and M. J. Black. Fields of experts. *International Journal of Computer Vision*, 82(2):205–229, 2009.

R. Rubinstein, M. Zibulevsky, and M. Elad. Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit. Technical report, Technion - Computer Science Department, 2008.

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, May 1997.

M. Schmidt, D. Kim, and S. Sra. Projected Newton-type methods in machine learning. In S. Sra, S. Nowozin, and S.J. Wright, editors, *Optimization for Machine Learning*. MIT Press, 2011.

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

M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.

T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

J. M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, 1993.

G. Sharma and H. J. Trussell. Digital color imaging. *IEEE Transactions on Image Processing*, 6(7):901–932, 1997.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, 38(2): 587–607, 1992.

J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2003.

K. Skretting and K. Engan. Recursive least squares dictionary learning algorithm. *IEEE Transactions on Signal Processing*, 58(4):2121–2130, 2010.

H. O. Song, S. Zickler, T. Althoff, R. Girshick, M. Fritz, C. Geyer, P. Felzenszwalb, and T. Darrell. Sparselet models for efficient multiclass object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.

D. A. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

J.-L. Starck, D. L. Donoho, and E. J. Candès. Astronomical image representation by the curvelet transform. *Astronomy and Astrophysics*, 398(2): 785–800, 2003.

A. Szlam, M. Maggioni, and R.R. Coifman. Regularization on graphs with function-adapted diffusion processes. *Journal of Machine Learning Research*, 2007.

H. Takeda, S. Farsiu, and P. Milanfar. Kernel regression for image processing and reconstruction. *IEEE Transactions on Image Processing*, 16(2):349–366, 2007.

H. L. Taylor, S. C. Banks, and J. F. McCoy. Deconvolution with the ℓ_1 norm. *Geophysics*, 44(1):39–52, 1979.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society Series B*, 67(1):91–108, 2005.

A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. In *Soviet Math. Doklady.*, volume 4, pages 1035–1038, 1963.

E. Tola, V. Lepetit, and P. Fua. DAISY: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, 2010.

I. Tosic, B. A. Olshausen, and B. J. Culpepper. Learning sparse representations of depth. *IEEE Journal of Selected Topics in Signal Processing*, 5(5): 941–952, 2011.

J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Signal Processing*, 50(10):2231–2242, 2004.

J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. part I: Greedy pursuit. *Signal Processing, special issue "sparse approximations in signal and image processing"*, 86:572–588, 2006.

P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, 2009.

D. Y. Ts'o and C. D. Gilbert. The organization of chromatic and spatial interactions in the primate striate cortex. *The Journal of Neuroscience*, 8 (5):1712–1727, 1988.

A. B. Tsybakov. Optimal rates of aggregation. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2003.

B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.

D. Vainsencher, S. Mannor, and A. M. Bruckstein. The sample complexity of dictionary learning. *Journal of Machine Learning Research*, 12:3259–3281, 2011.

J. C. van Gemert, C. J. Venman, A. W. M. Smeulders, and J. M. Geusebrock. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.

G. Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, and B. Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Information Processing in Medical Imaging*, pages 562–573, 2011.

G. Vinci, P. Freeman, J. Newman, L. Wasserman, and C. Genovese. Estimating the distribution of galaxy morphologies on a continuous space. In *Statistical Challenges in 21st Century Cosmology. Proceedings IAU Symposium No. 306*, 2014.

H. von Helmholtz. LXXXI. on the theory of compound colours. *Philosophical Magazine Series 4*, 4(28):519–534, 1852.

A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: robust alignment and illumination by sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):372–386, 2012.

M.J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming. *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2003.

J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

S. Wang, D. Zhang, Y. Liang, and Q. Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009.

L. Wasserman. *All of nonparametric statistics*. Springer, 2006.

G. S. Watson. Smooth regression analysis. *Sankhya, The Indian Journal of Statistics, Series A*, 26:359–372, 1964.

J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

D. Wrinch and H. Jeffreys. XLII. On certain fundamental principles of scientific inquiry. *Philosophical Magazine Series 6*, 42(249):369–390, 1921.

T. T. Wu and K. Lange. Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, 2(1):224–244, 2008.

Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 9:2543–2596, 2010.

R. Xiaofeng and L. Bo. Discriminatively trained sparse code gradients for contour detection. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Z. Xing, M. Zhou, A. Castrodad, G. Sapiro, and L. Carin. Dictionary learning for noisy and incomplete hyperspectral images. *SIAM Journal on Imaging Sciences*, 5(1):33–56, 2012.

J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010a.

J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010b.

J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang. Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing*, 21(8):3467–3478, 2012a.

J. Yang, K. Yu, and T. Huang. Efficient highly overcomplete sparse coming using a mixture model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012b.

M. Yang, D. Zhang, and X. Feng. Fisher discrimination dictionary learning for sparse representation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

T. Young. *A course of lectures on natural philosophy and the mechanical art*. Taylor and Watson, 1845.

G. Yu, G. Sapiro, and S. Mallat. Solving inverse problems with piecewise linear estimators: from Gaussian mixture models to structured sparsity. *IEEE Transactions on Image Processing*, 21(5):2481–2499, 2012.

K. Yu and T. Zhang. Improved local coordinate coding using local tangents. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68: 49–67, 2006.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse representations. In *Curves and Surfaces*, pages 711–730. Springer, 2012.

Q. Zhang and B. Li. Discriminative K-SVD for dictionary learning in face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *preprint arXiv:1402.1918*, 2014.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A): 3468–3497, 2009.

M. Zhou, H. Chen, L. Ren, G. Sapiro, L. Carin, and J. W. Paisley. Non-parametric bayesian dictionary learning for sparse image representations. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin. Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images. *IEEE Transactions on Image Processing*, 21 (1):130–144, 2012.

X. Zhou, K. Yu, T. Zhang, and T. Huang. Image classification using super-vector coding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.

S. C. Zhu and D. Mumford. Prior learning and gibbs reaction-diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(11):1236–1250, 1997.

S.-C. Zhu, C.-E. Guo, Y. Wang, and Z. Xu. What are textons? *International Journal of Computer Vision*, 62(1-2):121–143, 2005.

D. Zoran and Y. Weiss. From learning models of natural image patches to whole image restoration. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

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