Sparse Estimation for Image and Vision Processing

Julien Mairal

Inria, Grenoble

Ohrid, September 2016
Course material (freely available on arXiv)


Part I: A Short Introduction to Parcimony
Early thoughts

(a) Dorothy Wrinch
1894–1980

(b) Harold Jeffreys
1891–1989

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.

[Wrinch and Jeffreys, 1921]. Philosophical Magazine Series.
Historical overview of parsimony

- 14th century: Ockham’s razor;
- 1921: Wrinch and Jeffreys’ simplicity principle;
- 1952: Markowitz’s portfolio selection;
- 60 and 70’s: best subset selection in statistics;
- 70’s: use of the $\ell_1$-norm for signal recovery in geophysics;
- 90’s: wavelet thresholding in signal processing;
- 1996: Olshausen and Field’s dictionary learning;
- 1996–1999: Lasso (statistics) and basis pursuit (signal processing);
- 2006: compressed sensing (signal processing) and Lasso consistency (statistics);
- 2006–now: applications of dictionary learning in various scientific fields such as image processing and computer vision.
The modern parsimony and the $\ell_1$-norm
Sparse linear models in signal processing

Let $x$ in $\mathbb{R}^n$ be a signal.

Let $D = [d_1, \ldots, d_p] \in \mathbb{R}^{n \times p}$ be a set of elementary signals. We call it dictionary.

$D$ is “adapted” to $x$ if it can represent it with a few elements—that is, there exists a sparse vector $\alpha$ in $\mathbb{R}^p$ such that $x \approx D\alpha$. We call $\alpha$ the sparse code.
The modern parsimony and the $\ell_1$-norm
Sparse linear models in signal processing

How do we find $\alpha$?
We try to solve the sparse approximation problem

$$\min_{\alpha \in \mathbb{R}^p} \| x - D\alpha \|^2_2 \quad \text{s.t.} \quad \| \alpha \|_0 \leq k,$$

but...
The modern parsimony and the $\ell_1$-norm

Sparse linear models in signal processing

How do we find $\alpha$?

We try to solve the sparse approximation problem

$$\min_{\alpha \in \mathbb{R}^p} \| x - D\alpha \|_2^2 \quad \text{s.t.} \quad \| \alpha \|_0 \leq k,$$

but... the problem is NP-hard [Natarajan, 1995].

Strategy 1: try anyway

use greedy algorithm to find an approximate solution.

Strategy 2: use a convex relaxation

replace $\ell_0$ by $\ell_1$. 
The modern parsimony and the $\ell_1$-norm

Sparse linear models: machine learning/statistics point of view

Let $(y_i, x_i)_{i=1}^n$ be a training set, where the vectors $x_i$ are in $\mathbb{R}^p$ and are called features. The scalars $y_i$ are in

- $\{-1, +1\}$ for **binary** classification problems.
- $\mathbb{R}$ for **regression** problems.

We assume there exists a relation $y \approx \beta^T x$, and solve

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} L(y_i, \beta^T x_i) + \lambda \psi(\beta).$$

The term $\frac{1}{n} \sum_{i=1}^{n} L(y_i, \beta^T x_i)$ is the **empirical risk** and $\lambda \psi(\beta)$ is the **regularization**.
The modern parsimony and the $\ell_1$-norm

Sparse linear models: machine learning/statistics point of view

A few examples:

**Ridge regression:**
$$
\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} (y_i - \beta^T x_i)^2 + \lambda \|\beta\|_2^2.
$$

**Linear SVM:**
$$
\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i \beta^T x_i) + \lambda \|\beta\|_2^2.
$$

**Logistic regression:**
$$
\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \log \left(1 + e^{-y_i \beta^T x_i}\right) + \lambda \|\beta\|_2^2.
$$
The modern parsimony and the $\ell_1$-norm
Sparse linear models: machine learning/statistics point of view

A few examples:

**Ridge regression:**
\[
\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} (y_i - \beta^\top x_i)^2 + \lambda \|\beta\|_2^2.
\]

**Linear SVM:**
\[
\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i \beta^\top x_i) + \lambda \|\beta\|_2^2.
\]

**Logistic regression:**
\[
\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \log \left(1 + e^{-y_i \beta^\top x_i}\right) + \lambda \|\beta\|_2^2.
\]

The **squared $\ell_2$-norm** induces “smoothness” in $\beta$. When one knows in advance that $\beta$ should be sparse, one should use a **sparsity-inducing** regularization such as the $\ell_1$-norm. [Chen et al., 1999, Tibshirani, 1996]
The modern parsimony and the $\ell_1$-norm

Originally used to induce sparsity in geophysics [Claerbout and Muir, 1973, Taylor et al., 1979], the $\ell_1$-norm became popular in statistics with the Lasso [Tibshirani, 1996] and in signal processing with the Basis pursuit [Chen et al., 1999].

Three “equivalent” formulations

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

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

3. $$\min_{\alpha \in \mathbb{R}^p} \| \alpha \|_1 \quad \text{s.t.} \quad \| x - D\alpha \|_2^2 \leq \varepsilon.$$
The modern parsimony and the $\ell_1$-norm

And some variants...

For noiseless problems

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

Beyond least squares

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

where $f : \mathbb{R}^p \to \mathbb{R}$ is convex.
The modern parsimony and the $\ell_1$-norm

And some variants...

For noiseless problems

$$\min_{\alpha \in \mathbb{R}^p} \|\alpha\|_1 \; \text{s.t.} \; x = D\alpha.$$ 

Beyond least squares

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

where $f : \mathbb{R}^p \to \mathbb{R}$ is convex.

An important question remains:

why does the $\ell_1$-norm induce sparsity?
The modern parsimony and the $\ell_1$-norm

Why does the $\ell_1$-norm induce sparsity?

Can we get some intuition from the simplest isotropic case?

\[
\hat{\alpha}(\lambda) = \arg \min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \| x - \alpha \|^2_2 + \lambda \| \alpha \|_1,
\]
or equivalently the Euclidean projection onto the $\ell_1$-ball?

\[
\tilde{\alpha}(\mu) = \arg \min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \| x - \alpha \|^2_2 \quad \text{s.t.} \quad \| \alpha \|_1 \leq \mu.
\]

“equivalent” means that for all $\lambda > 0$, there exists $\mu \geq 0$ such that

\[
\tilde{\alpha}(\mu) = \hat{\alpha}(\lambda).
\]
The modern parsimony and the $\ell_1$-norm

Why does the $\ell_1$-norm induce sparsity?

Can we get some intuition from the simplest isotropic case?

\[ \hat{\alpha}(\lambda) = \arg \min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \| x - \alpha \|_2^2 + \lambda \| \alpha \|_1, \]

or equivalently the Euclidean projection onto the $\ell_1$-ball?

\[ \tilde{\alpha}(\mu) = \arg \min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \| x - \alpha \|_2^2 \quad \text{s.t.} \quad \| \alpha \|_1 \leq \mu. \]

“equivalent” means that for all $\lambda > 0$, there exists $\mu \geq 0$ such that $\tilde{\alpha}(\mu) = \hat{\alpha}(\lambda)$.

The relation between $\mu$ and $\lambda$ is unknown a priori.
Why does the $\ell_1$-norm induce sparsity?

Regularizing with the $\ell_1$-norm

The projection onto a convex set is “biased” towards singularities.
Why does the $\ell_1$-norm induce sparsity?

Regularizing with the $\ell_2$-norm

The $\ell_2$-norm is isotropic.
Why does the $\ell_1$-norm induce sparsity?

In 3D. (images produced by G. Obozinski)
Why does the $\ell_1$-norm induce sparsity?

Regularizing with the $\ell_\infty$-norm

$$\|\alpha\|_\infty \leq \mu$$

The $\ell_\infty$-norm encourages $|\alpha[1]| = |\alpha[2]|$. 

$$\alpha[2]$$

$$\alpha[1]$$
Why does the $\ell_1$-norm induce sparsity?
Analytical point of view: 1D case

$$\min_{\alpha \in \mathbb{R}} \frac{1}{2} (x - \alpha)^2 + \lambda |\alpha|$$

Piecewise quadratic function with a kink at zero.

Derivative at $0_+$: $g_+ = -x + \lambda$ and $0_-$: $g_- = -x - \lambda$.

Optimality conditions. $\alpha$ is optimal iff:

- $|\alpha| > 0$ and $(x - \alpha) + \lambda \text{sign}(\alpha) = 0$
- $\alpha = 0$ and $g_+ \geq 0$ and $g_- \leq 0$

The solution is a soft-thresholding:

$$\alpha^* = \text{sign}(x)(|x| - \lambda)^+.$$
Why does the $\ell_1$-norm induce sparsity?

Analytical point of view: 1D case

(c) Soft-thresholding operator,
\[ \alpha^{st} = \text{sign}(\beta) \max(|\beta| - \lambda, 0). \]

(d) Hard-thresholding operator
\[ \alpha^{ht} = \delta_{|\beta| \geq \mu} \beta. \]
Why does the $\ell_1$-norm induce sparsity?
Comparison with $\ell_2$-regularization in 1D

The gradient of the $\ell_2$-penalty vanishes when $\alpha$ get close to 0. On its differentiable part, the norm of the gradient of the $\ell_1$-norm is constant.
Why does the $\ell_1$-norm induce sparsity?

Physical illustration

\[ E_1 = 0 \]
Why does the $\ell_1$-norm induce sparsity?

Physical illustration

$$E_1 = \frac{k_1}{2} (x - \alpha)^2$$

$$E_2 = \frac{k_2}{2} \alpha^2$$
Why does the $\ell_1$-norm induce sparsity?

Physical illustration

$$E_1 = \frac{k_1}{2} (x - \alpha)^2$$

$$E_2 = \frac{k_2}{2} \alpha^2$$

$$E_1 = \frac{k_1}{2} (x - \alpha)^2$$

$$E_2 = mg\alpha$$

$\alpha = 0$!!
Non-convex sparsity-inducing penalties

Figure: Open balls in 2-D corresponding to several $\ell_q$-norms and pseudo-norms.

$$\|\alpha\|_q^q = \sum_{j=1}^{p} |\alpha[j]|^q.$$
Non-convex sparsity-inducing penalties

\[ \|\alpha\|_q \leq \mu \text{ with } q < 1 \]
Elastic-net

The elastic net introduced by [Zou and Hastie, 2005]

\[ \psi(\alpha) = \|\alpha\|_1 + \gamma \|\alpha\|_2^2, \]

The penalty provides more stable (but less sparse) solutions.
The elastic-net
vs other penalties

\[ \|\alpha\|_q \leq \mu \text{ with } q < 1 \]
The elastic-net
vs other penalties

\[ \| \alpha \|_1 \leq \mu \]
The elastic-net
vs other penalties

\[(1 - \gamma)\|\mathbf{\alpha}\|_1 + \gamma\|\mathbf{\alpha}\|_2^2 \leq \mu\]
The elastic-net vs other penalties

\[ \alpha \leq \mu \]

\[ \ell_2\text{-ball} \]

\[ \alpha[2] \]

\[ \alpha[1] \]
Structured sparsity
Structured sparsity
Part II: Discovering the structure of natural images
Dictionary learning

Neuroscientists were the first to automatically learn local structures in natural images.

The model of Olshausen and Field [1996] looks for a dictionary $D$ adapted to a training set of natural image patches $x_i$, $i = 1, \ldots, n$:

$$
\min_{D \in C, A \in \mathbb{R}^{p \times n}} \frac{1}{n} \sum_{i=1}^{n} \|x_i - D\alpha_i\|^2_2 + \lambda \psi(\alpha_i),
$$

where $A = [\alpha_1, \ldots, \alpha_n]$ and $C \triangleq \{D \in \mathbb{R}^{m \times p} : \forall j, \|d_j\|_2 \leq 1\}$.

Typical settings

- $n \approx 100\,000$;
- $m = 10 \times 10$ pixels;
- $p = 256$. 
Dictionary learning

Figure: with centering
Dictionary learning

Figure: with whitening
Dictionary learning

Why was it found impressive by neuroscientists?

- since Hubel and Wiesel [1968], it is known that some visual neurons are responding to particular image features, such as oriented edges.
- Later, Daugman [1985] demonstrated that fitting a linear model to neuronal responses given a visual stimuli may produce filters that can be well approximated by a two-dimensional Gabor function.
- The original motivation of Olshausen and Field [1996] was to establish a relation between the statistical structure of natural images and the properties of neurons from area V1.

The results provided some “support” for classical models of V1 based on Gabor filters.
Dictionary learning

Why was it found impressive by neuroscientists?

- since Hubel and Wiesel [1968], it is known that some visual neurons are responding to particular image features, such as oriented edges.
- Later, Daugman [1985] demonstrated that fitting a linear model to neuronal responses given a visual stimuli may produce filters that can be well approximated by a two-dimensional Gabor function.
- the original motivation of Olshausen and Field [1996] was to establish a relation between the statistical structure of natural images and the properties of neurons from area V1.

Warning

In fact, little is known about the early visual cortex [Olshausen and Field, 2005, Carandini et al., 2005].
Dictionary learning

Point of views

Matrix factorization

It is useful to see dictionary learning as a matrix factorization problem

\[
\min_{D \in \mathbb{C}, A \in \mathbb{R}^{p \times n}} \frac{1}{2n} \left\| X - DA \right\|_F^2 + \lambda \Psi(A).
\]

This is simply a matter of notation:

\[
\min_{D \in \mathbb{C}, A \in \mathbb{R}^{p \times n}} \frac{1}{n} \sum_{i=1}^{n} \left\| x_i - D \alpha_i \right\|_2^2 + \lambda \psi(\alpha_i),
\]

but the matrix factorization point of view allows us to make connections with numerous other unsupervised learning techniques, such as K-means, PCA, NMF, ICA...
Dictionary learning

Constrained variants

The formulations below are not equivalent

\[
\min_{D \in \mathcal{C}, A \in \mathbb{R}^{p \times n}} \sum_{i=1}^{n} \frac{1}{2} \| x_i - D\alpha_i \|_2^2 \quad \text{s.t.} \quad \psi(\alpha_i) \leq \mu.
\]

or

\[
\min_{D \in \mathcal{C}, A \in \mathbb{R}^{p \times n}} \sum_{i=1}^{n} \psi(\alpha_i) \quad \text{s.t.} \quad \| x_i - D\alpha_i \|_2^2 \leq \varepsilon.
\]

Using one instead of another depends on the problem at hand.
Pre-processing of natural image patches

Centering (also called removing the DC component)

\[
x_i \leftarrow x_i - \left( \frac{1}{m} \sum_{j=1}^{m} x_i[j] \right) 1_m,
\]

(a) Without pre-processing.

(b) After centering.
Pre-processing of natural image patches

Contrast (variance) normalization

\[ x_i \leftarrow \frac{1}{\text{max}(\|x_i\|_2, \eta)} x_i. \]

ex: \( \eta \) can be 0.2 times the mean value of the \( \|x_i\|_2 \).

(a) After centering.  
(b) After contrast normalization.
Pre-processing of natural image patches

Whitening after centering

\[ x_i \leftarrow US^\dagger U^\top x_i, \]

where \((1/\sqrt{n})X = USV^\top\) (SVD). Sometimes, small singular values are also set to zero. The resulting covariance \((1/n)XX^\top\) is close to identity.
Dictionary learning on color image patches

(c) With centering - RGB. (d) With whitening - RGB.

Figure: Dictionaries learned on RGB patches.
Dictionary learning with structured sparsity

Hierarchical dictionary learning

(a) Tree structure 1.
(b) Tree structure 2.

Figure: Hierarchical dictionaries learned on natural image patches of size $16 \times 16$ pixels.
Dictionary learning with structured sparsity

Topographic dictionary learning

(a) With $3 \times 3$ neighborhoods.
(b) With $4 \times 4$ neighborhood.

Figure: Topographic dictionaries learned on whitened natural image patches of size $12 \times 12$ pixels.
Part III: Sparse models for image processing
Image denoising

\[ y = x_{\text{orig}} + w \]

- \( y \) measurements
- \( x_{\text{orig}} \) original image
- \( w \) noise
Image denoising
Classical image models

\[ \mathbf{y} = \mathbf{x}_{\text{orig}} + \mathbf{w}. \]

Energy minimization problem - MAP estimation

\[ E(\mathbf{x}) = \frac{1}{2} \| \mathbf{y} - \mathbf{x} \|_2^2 + \psi(\mathbf{x}). \]

Some classical priors

- Smoothness \( \lambda \| \mathbf{L} \mathbf{x} \|_2^2 \);
- total variation \( \lambda \| \nabla \mathbf{x} \|_1^2 \) [Rudin et al., 1992];
- Markov random fields [Zhu and Mumford, 1997];
- wavelet sparsity \( \lambda \| \mathbf{W} \mathbf{x} \|_1 \).
Image denoising

The method of Elad and Aharon [2006]

Given a fixed dictionary $\mathbf{D}$, a patch $\mathbf{y}_i$ is denoised as follows:

1. center $\mathbf{y}_i$,

   $$\mathbf{y}_i^c \triangleq \mathbf{y}_i - \mu_i \mathbf{1}_m$$

   with

   $$\mu_i \triangleq \frac{1}{n} \mathbf{1}_m^\top \mathbf{y}_i;$$

2. find a sparse linear combination of dictionary elements that approximates $\mathbf{y}_i^c$ up to the noise level:

   $$\min_{\alpha_i \in \mathbb{R}^p} \| \alpha_i \|_0 \quad \text{s.t.} \quad \| \mathbf{y}_i^c - \mathbf{D} \alpha_i \|_2^2 \leq \varepsilon,$$

   where $\varepsilon$ is proportional to the noise variance $\sigma^2$;

3. add back the mean component to obtain the clean estimate $\hat{\mathbf{x}}_i$:

   $$\hat{\mathbf{x}}_i \triangleq \mathbf{D} \alpha_i^* + \mu_i \mathbf{1}_m,$$
Image denoising
The method of Elad and Aharon [2006]

An **adaptive** approach

1. extract all overlapping $\sqrt{m} \times \sqrt{m}$ patches $y_i$.
2. **dictionary learning**: learn $D$ on the set of centered noisy patches $[y_{c1}^i, \ldots, y_{cn}^i]$.
3. **final reconstruction**: find an estimate $\hat{x}_i$ for every patch using the approach of the previous slide;
4. **patch averaging**:

$$\hat{x} = \frac{1}{m} \sum_{i=1}^{n} R_i^\top \hat{x}_i,$$

**Remark**

Like other state-of-the-art denoising approaches, it is patch-based [Buades et al., 2005, Dabov et al., 2007].
Practical tricks

- use larger patches when the noise level is high;
- choose $\varepsilon = m(1.15\sigma)^2$ or take the 0.9-quantile of the $\chi^2_m$-distribution.
- always use the $\ell_0$ regularization for the final reconstruction;
- using $\ell_1$ for learning the dictionary seems to yield better results.
Image inpainting

[Mairal et al., 2008a,b]

For removing small holes in the image, a natural extension consists in introducing a **binary mask** $M_i$ in the formulation:

$$
\min_{D \in \mathbb{C}, A \in \mathbb{R}^{p \times n}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \| M_i(y_i - D \alpha_i) \|_2^2 + \lambda \psi(\alpha_i),
$$

The approach assumes that

- the noise is not structured;
- the holes are smaller than the patch size.

The problem is called inpainting [Bertalmio et al., 2000].
Image inpainting
[Mairal et al., 2008a,b]
Since 1699, when French explorers landed at the great bend of the Mississippi River and celebrated the first Mardi Gras in North America, New Orleans has brewed a fascinating melange of cultures. It was French, then Spanish, then French again, then sold to the United States. Through all these years, and even into the 1900s, others arrived from everywhere: Acadians (Cajuns), Africans, indige—
Image inpainting

[Mairal et al., 2008a,b]
Image inpainting

Inpainting a 12-Mpixel photograph [Mairal et al., 2009a]

The Salinas Valley is in Northern California. It is a long narrow seacoast between two ranges of mountains, and the Salinas River winds and twists up the center until it falls at last into Monterey Bay.

I remember my childhood names for grasses and sweet flowers. I remember where a toad may live and what time the birds gather in the summer and what trees and bushes smelt like how people looked and walked and smelled even. The memory of odors is very rich.

I remember that the Salinas Mountains to the east of the valley were light gray mountains full of sun and loveliness and a kind of invitation, so that you wanted to climb into their warm foothills almost as you want to climb into the lap of a beloved mother. They were beckoning mountains with a brown grass love. The Santa Lucias stood so against the sky to the west and kept the valley from the open sea, and they were dark and brooding-unfriendly and dangerous. I always found in myself a sense of west and a love of east. Where I ever got such an idea I cannot say, unless it could be that the morning came over the peaks of the Salinas and the night drifted back from the ridges of the Santa Lucias. It may be that the birth and death of the day had some part in my feeling about the two ranges of mountains.

From both sides of the valley little streams slipped out of the hill canyons and fell into the bed of the Salinas River. In the winter of wet years the streams ran full-freshet, and they swelled the river until sometimes it raged and boiled, bank full, and then it was a destroyer. The river tore the edges of the farm lands and washed whole acres down; it toppled barns and houses into itself, to go floating and bobbing away. It trapped cows and pigs and sheep and drowned them in its muddy brown water and carried them to the sea. Then when the late spring came, the river drew up from its banks and the sand banks appeared. And in the summer the river didn't run at all above ground. Some pools would be left in the deep swale places under a high bank. The fules and grasses grew thick, and willows straightened up with the flood debris in their upper branches. The Salinas was only a rain-time river. The summer sun drove it underground. It was not a fast river at all, but it was the only one we had and so we boasted about it, how dangerous it was in a wet winter and how dry it was in a dry summer. You can boast about anything if it's all you have. Maybe the less you have, the more you are required to boast.

The floor of the Salinas Valley, between the ranges and below the foothills, is level because this valley used to be the bottom of a hundred-mile inlet from the sea. The river mouth at Moss Landing was centuries ago the entrance to this long inland water. Once, fifty miles down the valley, my father boated a well. The draft came up first with topsoil and then with gravel and then with white sea sand full of shells and even pi...
Image inpainting

Inpainting a 12-Mpixel photograph [Mairal et al., 2009a]

![Image of a seagull on a rocky beach]

Julien Mairal

Sparse Estimation for Image and Vision Processing
Image inpainting

Inpainting a 12-Mpixel photograph [Mairal et al., 2009a]
Image inpainting

Inpainting a 12-Mpixel photograph [Mairal et al., 2009a]
Problem

The noise pattern is very structured: the previous inpainting scheme needs to be modified [Mairal et al., 2008a].
Image demosaicking

(a) Mosaicked image  (b) Demosaicked image A  (c) Demosaicked image B

Figure: Demosaicked image A is with the approach previously described; image B is with an extension called non-local sparse model [Mairal et al., 2009b].
Image demoising

Figure: Demosaicked image A is with the approach previously described; image B is with an extension called non-local sparse model [Mairal et al., 2009b].
Video processing

Extension developed by Protter and Elad [2009]:

Key ideas for video processing

- Using a 3D dictionary.
- Processing of many frames at the same time.
- Dictionary propagation.
Video processing
Inpainting, [Mairal et al., 2008b]

Figure: Inpainting results.
Video processing
Inpainting, [Mairal et al., 2008b]

Figure: Inpainting results.
Video processing
Inpainting, [Mairal et al., 2008b]

Figure: Inpainting results.
Video processing

Inpainting, [Mairal et al., 2008b]

**Figure:** Inpainting results.
Video processing

Inpainting, [Mairal et al., 2008b]

**Figure:** Inpainting results.
Video processing

Color video denoising, [Mairal et al., 2008b]

Figure: Inpainting results.
Video processing
Color video denoising, [Mairal et al., 2008b]

Figure: Inpainting results.
Video processing

Color video denoising, [Mairal et al., 2008b]

Figure: Inpainting results.
Video processing

Color video denoising, [Mairal et al., 2008b]

**Figure:** Inpainting results.
Video processing
Color video denoising, [Mairal et al., 2008b]

Figure: Inpainting results.
Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

Figure: Original
Inverting nonlinear local transformations
Inverse half-toning [Mairal et al., 2012]

Figure: Binary image
Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

Figure: Reconstructed.
Inverting nonlinear local transformations
Inverse half-toning [Mairal et al., 2012]

Figure: Original
Inverting nonlinear local transformations
Inverse half-toning [Mairal et al., 2012]

Figure: Binary image
Inverting nonlinear local transformations
Inverse half-toning [Mairal et al., 2012]

Figure: Reconstructed.
Inverting nonlinear local transformations
Inverse half-toning [Mairal et al., 2012]
Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]
Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]
Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]
Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]
Inverting nonlinear local transformations
Inverse half-toning [Mairal et al., 2012]
Inverting nonlinear local transformations
Inverse half-toning [Mairal et al., 2012]
Inverting nonlinear local transformations
Inverse half-toning [Mairal et al., 2012]
Other patch modeling approaches

Non-local means and non-parametric approaches

Image pixels are well explained by a Nadaraya-Watson estimator:

\[
\hat{x}[i] = \sum_{j=1}^{n} \frac{K_h(y_i - y_j)}{\sum_{l=1}^{n} K_h(y_i - y_l)} y[j],
\]

(2)

with successful application to

- texture synthesis: [Efros and Leung, 1999]
- image denoising (Non-local means): [Buades et al., 2005]
- image demosaicking: [Buades et al., 2009].
Other patch modeling approaches

BM3D

state-of-the-art image denoising approach [Dabov et al., 2007]:

- **block matching**: for each patch, find similar ones in the image;
- **3D wavelet filtering**: denoise blocks of patches with 3D-DCT;
- **patch averaging**: average estimates of overlapping patches;
- **second step with Wiener filtering**: use the first estimate to perform again and improve the previous steps.

Further refined by Dabov et al. [2009] with shape-adaptive patches and PCA filtering.
Other patch modeling approaches

Non-local sparse models [Mairal et al., 2009b]

Exploit some ideas of BM3D to combine the non-local means principle with dictionary learning.

The main idea is that similar patches should admit similar decompositions by using group sparsity:

The approach uses a block matching/clustering step, followed by group sparse coding and patch averaging.
Other patch modeling approaches
Non-local sparse image models
Other patch modeling approaches
Non-local sparse image models

Restored
Part IV: Optimization for sparse estimation
Sparse reconstruction with the $\ell_0$-penalty

Matching pursuit [Mallat and Zhang, 1993]

$\alpha = (0, 0, 0)$
Sparse reconstruction with the \( \ell_0 \)-penalty

Matching pursuit [Mallat and Zhang, 1993]

\[ \alpha = (0, 0, 0) \]
Sparse reconstruction with the $\ell_0$-penalty

Matching pursuit [Mallat and Zhang, 1993]

$$\alpha = (0, 0, 0)$$
Sparse reconstruction with the $\ell_0$-penalty

Matching pursuit [Mallat and Zhang, 1993]

$$\alpha = (0, 0, 0.75)$$
Sparse reconstruction with the $\ell_0$-penalty
Matching pursuit [Mallat and Zhang, 1993]

$\alpha = (0, 0, 0.75)$
Sparse reconstruction with the $\ell_0$-penalty

Matching pursuit [Mallat and Zhang, 1993]

$\alpha = (0, 0, 0.75)$
Sparse reconstruction with the $\ell_0$-penalty

Matching pursuit [Mallat and Zhang, 1993]

$\alpha = (0, 0.24, 0.75)$
Sparse reconstruction with the $\ell_0$-penalty

Matching pursuit [Mallat and Zhang, 1993]

$\alpha = (0, 0.24, 0.75)$
Sparse reconstruction with the $\ell_0$-penalty

Matching pursuit [Mallat and Zhang, 1993]

$\alpha = (0, 0.24, 0.75)$
Sparse reconstruction with the $\ell_0$-penalty

Matching pursuit [Mallat and Zhang, 1993]

$$\alpha = (0, 0.24, 0.65)$$
Sparse reconstruction with the $\ell_0$-penalty

Matching pursuit [Mallat and Zhang, 1993]

$$
\min_{\alpha \in \mathbb{R}^p} \| x - D\alpha \|_2^2 \quad \text{s.t.} \quad \| \alpha \|_0 \leq k.
$$

1: $\alpha \leftarrow 0$
2: $r \leftarrow x$ (residual).
3: while $\| \alpha \|_0 < k$ do
4: Select the predictor with maximum inner-product with the residual
   $$
   \hat{j} \leftarrow \arg \max_{j=1,\ldots,p} |d_j^\top r|
   $$
5: Update the residual and the coefficients
   $$
   \alpha[\hat{j}] \leftarrow \alpha[\hat{j}] + d_{\hat{j}}^\top r
   $$
   $$
   r \leftarrow r - (d_{\hat{j}}^\top r)d_{\hat{j}}
   $$
6: end while
Sparse reconstruction with the $\ell_0$-penalty

Matching pursuit [Mallat and Zhang, 1993]

Remarks

- Matching pursuit is a **coordinate descent** algorithm. It greedily selects one coordinate at a time and optimizes the cost function with respect to that coordinate.

$$\alpha[\hat{j}] \leftarrow \arg\min_{\alpha \in \mathbb{R}} \left\| x - \sum_{l \neq \hat{j}} \alpha[l]d_l - \alpha d_{\hat{j}} \right\|_2^2.$$  

- Each coordinate can be selected several times during the process.

- The roots of this algorithm can be found in the statistics literature [Efroymson, 1960].
Sparse reconstruction with the $\ell_0$-penalty
Orthogonal matching pursuit [Pati et al., 1993]

$\alpha = (0, 0, 0)$
$\Gamma = \emptyset$
Sparse reconstruction with the $\ell_0$-penalty

Orthogonal matching pursuit [Pati et al., 1993]

$\alpha = (0, 0, 0.75)$

$\Gamma = \{3\}$
Sparse reconstruction with the $\ell_0$-penalty

Orthogonal matching pursuit [Pati et al., 1993]

$\alpha = (0, 0.29, 0.63)$

$\Gamma = \{3, 2\}$
Sparse reconstruction with the $\ell_0$-penalty

Orthogonal matching pursuit [Pati et al., 1993]

$$\min_{\alpha \in \mathbb{R}^p} \| x - D\alpha \|_2^2 \quad \text{s.t.} \quad \| \alpha \|_0 \leq k$$

1: $\Gamma = \emptyset$.

2: for $\text{iter} = 1, \ldots , k$ do

3: Select the variable that most reduces the objective

$$(\hat{j}, \hat{\beta}) \leftarrow \arg \min_{j \in \Gamma^c, \beta} \| x - D_{\Gamma \cup \{j\}} \beta \|_2^2.$$

4: Update the active set: $\Gamma \leftarrow \Gamma \cup \{\hat{j}\}$.

5: Update the coefficients:

$$\alpha[\Gamma] \leftarrow \hat{\beta} \quad \text{and} \quad \alpha[\Gamma^c] \leftarrow 0.$$  

6: end for
Sparse reconstruction with the $\ell_0$-penalty
Orthogonal matching pursuit [Pati et al., 1993]

Remarks

- this is an **active-set** algorithm.
- when a new variable is selected, the coefficients for the full set $\Gamma$ are re-optimized:

$$\alpha[\Gamma] = (D_\Gamma^T D_\Gamma)^{-1} D_\Gamma^T x,$$

and the residual is always orthogonal to the matrix $D_\Gamma$ of previously selected dictionary elements:

$$D_\Gamma^T (x - D\alpha) = D_\Gamma^T (x - D_\Gamma \alpha[\Gamma]) = 0.$$

- several variants of OMP exist regarding the selection rule of $\hat{j}$. The one we use appears in Cotter et al. [1999].
Sparse reconstruction with the $\ell_0$-penalty

Orthogonal matching pursuit [Pati et al., 1993]

**Keys for a fast implementation**

- If available, use the Gram matrix $\mathbf{G} = \mathbf{D}^\top \mathbf{D}$;
- Maintain the computation of $\mathbf{D}^\top (\mathbf{x} - \mathbf{D} \alpha)$,
- Update the Cholesky decomposition of $(\mathbf{D}_\Gamma^\top \mathbf{D}_\Gamma)^{-1}$.

The total complexity for decomposing $n$ $k$-sparse signals of size $m$ with a dictionary of size $p$ is

$$O(p^2 m) + O(nk^3) + O(n(pm + pk^2)) = O(np(m + k^2))$$

It is also possible to use the matrix inversion lemma instead of a Cholesky decomposition.
Sparse reconstruction with the $\ell_0$-penalty
Orthogonal matching pursuit [Pati et al., 1993]

Example with the software SPAMS

```matlab
>> I=double(imread('data/lena.eps'))/255;
>> %extract all patches of I
>> X=im2col(I,[8 8],’sliding’);
>> %load a dictionary of size 64 x 256
>> D=load(’dict.mat’);
>>
>> %set the sparsity parameter L to 10
>> param.L=10;
>> alpha=mexOMP(X,D,param);
```

On this dual-core laptop: **150000 signals processed per second!**
Sparse reconstruction with the $\ell_1$-norm

Coordinate descent for the Lasso [Fu, 1998]

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

The coordinate descent method consists of iteratively fixing all variables and optimizing with respect to one:

$$
\alpha[j] \leftarrow \arg \min_{\alpha \in \mathbb{R}} \frac{1}{2} \| x - \sum_{l \neq j} \alpha[l] d[l] \|_2^2 + \lambda |\alpha|.
$$

Assume the columns of $D$ to have unit $\ell_2$-norm,

$$
\alpha_j \leftarrow \text{sign}(d_j^T r)(|d_j^T r| - \lambda)^+.
$$

This involves again the \textbf{soft-thresholding} operator.

Optimization for Dictionary Learning

\[ \min_{\alpha \in \mathbb{R}^{p \times n}} \sum_{i=1}^{n} \frac{1}{2} \| x_i - D \alpha_i \|_2^2 + \lambda \psi(\alpha_i) \]

\[ \mathcal{C} \triangleq \{ D \in \mathbb{R}^{m \times p} \text{ s.t. } \forall j = 1, \ldots, p, \| d_j \|_2 \leq 1 \}. \]

Classical approach

- Alternate minimization between $D$ and $\alpha$ (MOD with $\psi = \ell_0$ [Engan et al., 1999], K-SVD with $\psi = \ell_0$ [Aharon et al., 2006], [Lee et al., 2007] with $\psi = \ell_1$);

- good results, reliable, but can be slow when $n$ is large!
Conclusion

What we have seen:
- why the $\ell_1$-norm induce sparsity (part I);
- the classical dictionary learning formulations on natural image patches (part II);
- a few applications to image restoration (part III);
- a few algorithms (part IV).
Conclusion

What we have NOT seen:

- structured sparsity, theory of sparse estimation.
- other matrix factorization formulations;
- applications in computer vision;
- many algorithms including stochastic optimization for dictionary learning.
References I


References II


References III


References IV


References VI


References VII


