Topographic Dictionary Learning
with Structured Sparsity

Julien Mairal\textsuperscript{1} Rodolphe Jenatton\textsuperscript{2}
Guillaume Obozinski\textsuperscript{2} Francis Bach\textsuperscript{2}

\textsuperscript{1}UC Berkeley \textsuperscript{2}INRIA - SIERRA Project-Team

San Diego, Wavelets and Sparsity XIV, August 2011
What this work is about

- Group sparsity with overlapping groups.
- Hierarchical, topographic dictionary learning,
- More generally: structured dictionaries of natural image patches.

Related publications:


Part I: Introduction to Dictionary Learning
What is a Sparse Linear Model?

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

Let $D = [d^1, \ldots, d^p] \in \mathbb{R}^{m \times p}$ be a set of normalized “basis vectors”. We call it dictionary.

$D$ is “adapted” to $x$ if it can represent it with a few basis vectors—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 Sparse Decomposition Problem

\[
\min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \| x - D\alpha \|^2_2 + \lambda \psi(\alpha)
\]

\(\psi\) induces sparsity in \(\alpha\):
- the \(\ell_0\) “pseudo-norm”. \(\| \alpha \|_0 \triangleq \# \{ i \text{ s.t. } \alpha_i \neq 0 \}\) (NP-hard)
- the \(\ell_1\) norm. \(\| \alpha \|_1 \triangleq \sum_{i=1}^p |\alpha_i|\) (convex),
- \(\ldots\)

This is a selection problem. When \(\psi\) is the \(\ell_1\)-norm, the problem is called Lasso [Tibshirani, 1996] or basis pursuit [Chen et al., 1999]
Sparse representations for image restoration

**Designed dictionaries**

[Haar, 1910], [Zweig, Morlet, Grossman ∼70s], [Meyer, Mallat, Daubechies, Coifman, Donoho, Candes ∼80s-today]...

Wavelets, Curvelets, Wedgelets, Bandlets, ...lets

**Learned dictionaries of patches**

[Olshausen and Field, 1997, Engan et al., 1999, Lewicki and Sejnowski, 2000, Aharon et al., 2006], ...

\[
\min_{\alpha^i, D \in \mathcal{D}} \sum_{i=1}^{n} \frac{1}{2} \| x^i - D \alpha^i \|_2^2 + \lambda \psi(\alpha^i)
\]

- \(\psi(\alpha) = \| \alpha \|_0 \) ("\(\ell_0\) pseudo-norm")
- \(\psi(\alpha) = \| \alpha \|_1 \) (\(\ell_1\) norm)
Sparse representations for image restoration
Grayscale vs color image patches

Figure: Left: learned on grayscale image patches. Right: learned on color image patches (after removing the mean color from each patch)
Algorithms

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

How do we optimize that?

- alternate between \( D \) and \( \alpha \) [Engan et al., 1999], or other variants [Elad and Aharon, 2006]
- online learning [Olshausen and Field, 1997, Mairal et al., 2009, Skretting and Engan, 2010]

**Code SPAMS available:** [http://www.di.ens.fr/willow/SPAMS/](http://www.di.ens.fr/willow/SPAMS/), now open-source!
Part II: Introduction to Structured Sparsity
(Let us play with $\psi$)
Group Sparsity-Inducing Norms

\[
\min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \|x - D\alpha\|_2^2 + \lambda \psi(\alpha)
\]

sparsity-inducing norm

The most popular choice for \(\psi\):

- The \(l_1\) norm, \(\psi(\alpha) = \|\alpha\|_1\).
- However, the \(l_1\) norm encodes poor information, just cardinality!

Another popular choice for \(\Omega\):

- The \(l_1-l_q\) norm [Turlach et al., 2005], with \(q = 2\) or \(q = \infty\)

\[
\sum_{g \in G} \|\alpha_g\|_q \quad \text{with } G \text{ a partition of } \{1, \ldots, p\}.
\]

- The \(l_1-l_q\) norm sets to zero groups of non-overlapping variables (as opposed to single variables for the \(l_1\) norm).
Structured Sparsity with Overlapping Groups

Warning: Under the name “structured sparsity” appear in fact significantly different formulations!

1. non-convex
   - zero-tree wavelets [Shapiro, 1993]
   - sparsity patterns are in a predefined collection: [Baraniuk et al., 2010]
   - select a union of groups: [Huang et al., 2009]
   - structure via Markov Random Fields: [Cehver et al., 2008]

2. convex
   - tree-structure: [Zhao et al., 2009]
   - non-zero patterns are a union of groups: [Jacob et al., 2009]
   - zero patterns are a union of groups: [Jenatton et al., 2009]
   - other norms: [Micchelli et al., 2010]
Structured Sparsity with Overlapping Groups

\[ \psi(\alpha) = \sum_{g \in G} \| \alpha_g \|_q \]

What happens when the groups overlap? [Jenatton et al., 2009]

- Inside the groups, the \( \ell_2 \)-norm (or \( \ell_\infty \)) does not promote sparsity.
- Variables belonging to the same groups are encouraged to be set to zero together.
Examples of set of groups $\mathcal{G}$

[Jenatton et al., 2009]

Selection of contiguous patterns on a sequence, $p = 6$.

- $\mathcal{G}$ is the set of blue groups.
- Any union of blue groups set to zero leads to the selection of a contiguous pattern.
Hierarchical Norms

[Zhao et al., 2009]

A node can be active only if its ancestors are active. The selected patterns are rooted subtrees.
Algorithms/Difficulties

[Jenatton et al., 2010, Mairal et al., 2011]

\[ \min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \| x - D\alpha \|_2^2 + \lambda \sum_{g \in G} \| \alpha_g \|_q. \]

The function is convex non-differentiable; the sum is a sum of simple non-separable regularizers.

How do we optimize that?

- hierarchical norms: same complexity as \( \ell_1 \) with proximal methods.
- general case: Augmenting Lagrangian Techniques.
- general case with \( \ell_{\infty} \)-norms: proximal methods combine with network flow optimization.

Also implemented in the toolbox SPAMS
Part III: Learning Structured Dictionaries
Topographic Dictionary Learning

- [Kavukcuoglu et al., 2009]: organize the dictionary elements on a 2D-grids and use $\psi$ with $e \times e$ overlapping groups.
- [Garrigues and Olshausen, 2010]: sparse coding + probabilistic model to model lateral interactions.
- topographic ICA by Hyvärinen et al. [2001]:

![Diagram of topographic dictionary learning](image-url)
Topographic Dictionary Learning

[Mairal, Jenatton, Obozinski, and Bach, 2011], $3 \times 3$-neighborhoods
Hierarchical Dictionary Learning

[Jenatton, Mairal, Obozinski, and Bach, 2010]
Conclusion / Discussion

- Structured sparsity is a natural framework for learning structured dictionaries...
- ...and has efficient optimization tools.
- other applications in natural language processing, bio-informatics, neuroscience...
SPAMS toolbox (open-source)

- C++ interfaced with Matlab.
- proximal gradient methods for $\ell_0$, $\ell_1$, elastic-net, fused-Lasso, group-Lasso, tree group-Lasso, tree-$\ell_0$, sparse group Lasso, overlapping group Lasso...
- ...for square, logistic, multi-class logistic loss functions.
- handles sparse matrices,
- provides duality gaps.
- also coordinate descent, block coordinate descent algorithms.
- fastest available implementation of OMP and LARS.
- dictionary learning and matrix factorization (NMF, sparse PCA).
- fast projections onto some convex sets.

Try it! http://www.di.ens.fr/willow/SPAMS/
References I


References IV


References VI


First-order/proximal methods

\[ \min_{\alpha \in \mathbb{R}^p} f(\alpha) + \lambda \Omega(\alpha) \]

- \( f \) is strictly convex and differentiable with a Lipshitz gradient.
- Generalizes the idea of gradient descent

\[
\alpha^{k+1} \leftarrow \arg\min_{\alpha \in \mathbb{R}^p} f(\alpha^k) + \nabla f(\alpha^k)^\top (\alpha - \alpha^k) + \frac{L}{2} \|\alpha - \alpha^k\|_2^2 + \lambda \Omega(\alpha)
\]

\[
\leftarrow \arg\min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \|\alpha - (\alpha^k - \frac{1}{L} \nabla f(\alpha^k))\|_2^2 + \frac{\lambda}{L} \Omega(\alpha)
\]

When \( \lambda = 0 \), \( \alpha^{k+1} \leftarrow \alpha^k - \frac{1}{L} \nabla f(\alpha^k) \), this is equivalent to a classical gradient descent step.
First-order/proximal methods

- They require solving efficiently the **proximal operator**

\[
\min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \|u - \alpha\|_2^2 + \lambda \Omega(\alpha)
\]

- For the $\ell_1$-norm, this amounts to a soft-thresholding:

\[
\alpha_i^* = \text{sign}(u_i)(u_i - \lambda)^+
\]


- suited for large-scale experiments.
Tree-structured groups

Proposition [Jenatton, Mairal, Obozinski, and Bach, 2010]

- If $\mathcal{G}$ is a tree-structured set of groups, i.e., $\forall g, h \in \mathcal{G}$,

\[ g \cap h = \emptyset \text{ or } g \subset h \text{ or } h \subset g \]

- For $q = 2$ or $q = \infty$, we define $\text{Prox}_g$ and $\text{Prox}_\Omega$ as

\[
\text{Prox}_g : u \rightarrow \arg \min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \| u - \alpha \| + \lambda \| \alpha_g \|_q,
\]

\[
\text{Prox}_\Omega : u \rightarrow \arg \min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \| u - \alpha \| + \lambda \sum_{g \in \mathcal{G}} \| \alpha_g \|_q,
\]

- If the groups are sorted from the leaves to the root, then

\[ \text{Prox}_\Omega = \text{Prox}_{g_m} \circ \ldots \circ \text{Prox}_{g_1}. \]

→ Tree-structured regularization: Efficient linear time algorithm.
General Overlapping Groups for $q = \infty$
[Mairal, Jenatton, Obozinski, and Bach, 2011]

Dual formulation

The solutions $\alpha^*$ and $\xi^*$ of the following optimization problems

\[
\min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \|u - \alpha\| + \lambda \sum_{g \in \mathcal{G}} \|\alpha_g\|_{\infty}, \quad \text{(Primal)}
\]

\[
\min_{\xi \in \mathbb{R}^{p \times |\mathcal{G}|}} \frac{1}{2} \|u - \sum_{g \in \mathcal{G}} \xi_g\|_2^2 \quad \text{s.t.} \quad \forall g \in \mathcal{G}, \|\xi^g\|_1 \leq \lambda \quad \text{and} \quad \xi^g_j = 0 \text{ if } j \notin g, \quad \text{(Dual)}
\]

satisfy

\[
\alpha^* = u - \sum_{g \in \mathcal{G}} \xi^g. \quad \text{(Primal-dual relation)}
\]

The dual formulation has more variables, but is equivalent to quadratic min-cost flow problem.