Recent Advances in Structured Sparse Models

Julien Mairal

Willow group - INRIA - ENS - Paris

21 September 2010

LEAR seminar

At Grenoble, September 21^{st}, 2010

Acknowledgements

Francis Bach

Rodolphe Jenatton

Guillaume Obozinski

What this talk is about?

- **Sparse** Linear Models
- Not only sparse, but also **structured**!
- Solving challenging optimization problems
- Developping new applications of sparse models in computer vision and machine learning

Related publications:

[1] J. Mairal, R. Jenatton, G. Obozinski and F. Bach. Network Flow Algorithms for Structured Sparsity. NIPS, 2010

[2] R. Jenatton, J. Mairal, G. Obozinski and F. Bach. Proximal Methods for Hierarchical Sparse Coding. arXiv:1009.2139v1

[3] R. Jenatton, J. Mairal, G. Obozinski and F. Bach. Proximal Methods for Sparse Hierarchical Dictionary Learning. ICML, 2010

Sparse Linear Model: Machine Learning Point of View

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

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

In a linear model, on assumes a relation $y \approx \mathbf{w}^\top \mathbf{x}$, and solves

$$\min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(y^i, \mathbf{w}^\top \mathbf{x}^i)}_{\text{data-fitting}} + \underbrace{\lambda \Omega(\mathbf{w})}_{\text{regularization}} .$$

Sparse Linear Models: Machine Learning Point of View

A few examples:

Ridge regression:
$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\sum_{i=1}^{n}(y^i - \mathbf{w}^\top \mathbf{x}^i)^2 + \lambda\|\mathbf{w}\|_2^2.$$

Linear SVM:
$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y^i\mathbf{w}^\top \mathbf{x}^i) + \lambda\|\mathbf{w}\|_2^2.$$

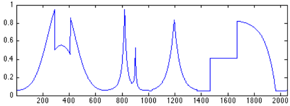
Logistic regression:
$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}\log\left(1 + e^{-y^i\mathbf{w}^\top \mathbf{x}^i}\right) + \lambda\|\mathbf{w}\|_2^2.$$

The squared ℓ_2-norm induces **smoothness** in $\mathbf{w}$. When one knows in advance that $\mathbf{w}$ should be sparse, one should use a **sparsity-inducing** regularization such as the ℓ_1-norm. [Chen et al., 1999, Tibshirani, 1996]

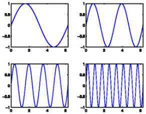
The purpose of the talk is to add **additional a-priori knowledge** in the regularization.

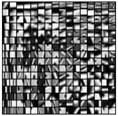
Sparse Linear Models: Signal Processing Point of View

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



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





$\mathbf{D}$ is "adapted" to $\mathbf{y}$ if it can represent it with a few basis vectors—that is, there exists a **sparse vector** $\mathbf{w}$ in $\mathbb{R}^p$ such that $\mathbf{x} \approx \mathbf{D}\mathbf{w}$. We call $\mathbf{w}$ the **sparse code**.

$$\underbrace{\begin{pmatrix} \mathbf{y} \end{pmatrix}}_{\mathbf{y}\in\mathbb{R}^n} \approx \underbrace{\begin{pmatrix} \mathbf{d}^1 & \mathbf{d}^2 & \cdots & \mathbf{d}^p \end{pmatrix}}_{\mathbf{D}\in\mathbb{R}^{n\times p}} \underbrace{\begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_p \end{pmatrix}}_{\mathbf{w}\in\mathbb{R}^p,\textbf{sparse}}$$

Sparse Linear Models: the Lasso

- Signal processing: $\mathbf{D}$ is a dictionary in $\mathbb{R}^{n \times p}$,

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

- Machine Learning:

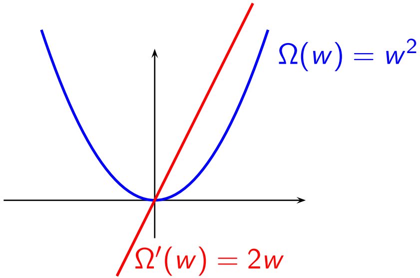
$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y^i - \mathbf{x}^{i\top}\mathbf{w})^2 + \lambda \|\mathbf{w}\|_1 = \min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}^\top \mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1,$$

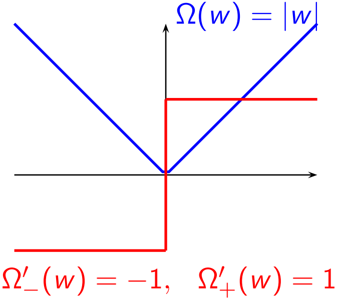
with $\mathbf{X} \triangleq [\mathbf{x}^1, \ldots, \mathbf{x}^n]$, and $\mathbf{y} \triangleq [y^1, \ldots, y^n]^\top$.

Useful tool in signal processing, machine learning, statistics, neuroscience,... as long as one wishes to **select** features.

Why does the ℓ_1-norm induce sparsity?

Analysis of the norms in 1D

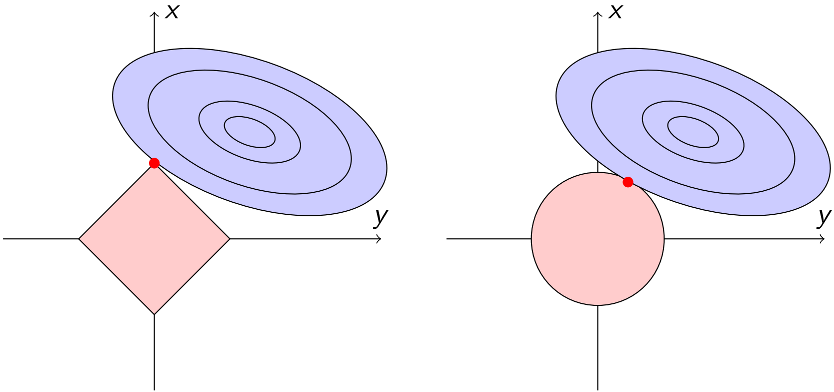




The gradient of the ℓ_2-norm vanishes when w get close to 0. On its differentiable part, the norm of the gradient of the ℓ_1-norm is constant.

Why does the ℓ_1-norm induce sparsity?

Geometric explanation



$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y}-\mathbf{D}\mathbf{w}\|_2^2+\lambda\|\mathbf{w}\|_1$$

$$\min_{\mathbf{w}\in\mathbb{R}^p} \|\mathbf{y}-\mathbf{D}\mathbf{w}\|_2^2 \ \text{ s.t. } \ \|\mathbf{w}\|_1 \leq T.$$

Other Sparsity-Inducing Norms

$$\min_{\mathbf{w}\in\mathbb{R}^p} \quad \overbrace{f(\mathbf{w})}^{\text{data fitting term}} + \lambda \underbrace{\Omega(\mathbf{w})}_{\text{sparsity-inducing norm}}$$

The most popular choice for Ω:

- The ℓ_1 norm, $\|\mathbf{w}\|_1 = \sum_{j=1}^{p} |\mathbf{w}_j|$.
- However, the ℓ_1 norm encodes poor information, just **cardinality**!

Another popular choice for Ω:

- The ℓ_1-ℓ_q norm [Yuan and Lin, 2006], with $q = 2$ or $q = \infty$

$$\sum_{g\in\mathcal{G}} \|\mathbf{w}_g\|_q \quad \text{with } \mathcal{G} \text{ a } \textbf{partition} \text{ of } \{1,\ldots,p\}.$$

- The ℓ_1-ℓ_q norm sets to zero **groups of non-overlapping variables** (as opposed to single variables for the ℓ_1 norm).

Sparsity-Inducing Norms

$$\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_q$$

Applications of group sparsity:

- Selecting groups of features instead of individual variables.
- Multi-task learning.
- Multiple kernel learning.

Drawbacks:

- Requires a **partition** of the features.
- Encodes fixed/static information.

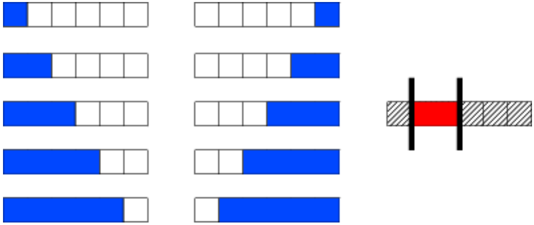
What happens when the groups overlap? [Jenatton et al., 2009]

- Inside the groups, the ℓ_2-norm (or ℓ_∞) does not promote sparsity.
- Variables belonging to the same groups are encouraged to be set to zero together.

Examples of set of groups $\mathcal{G}$ (1/3)

[Jenatton et al., 2009]

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

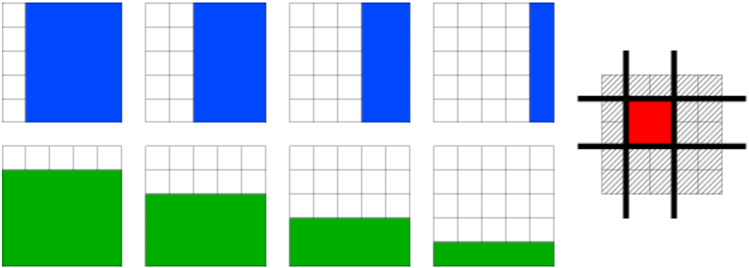


- $\mathcal{G}$ is the set of blue groups.
- Any union of blue groups set to zero leads to the selection of a contiguous pattern.

Examples of set of groups $\mathcal{G}$ (2/3)

[Jenatton et al., 2009]

Selection of rectangles on a 2-D grids, $p = 25$.

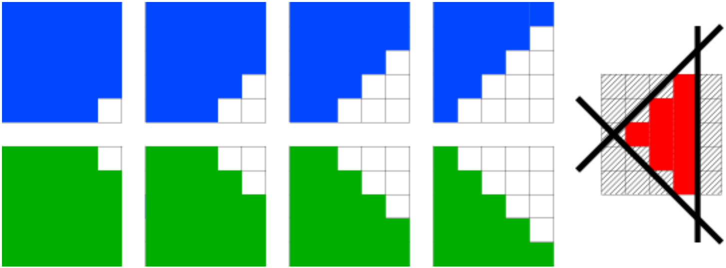


- $\mathcal{G}$ is the set of blue/green groups (with their not displayed complements).
- Any union of blue/green groups set to zero leads to the selection of a rectangle.

Examples of set of groups $\mathcal{G}$ (3/3)

[Jenatton et al., 2009]

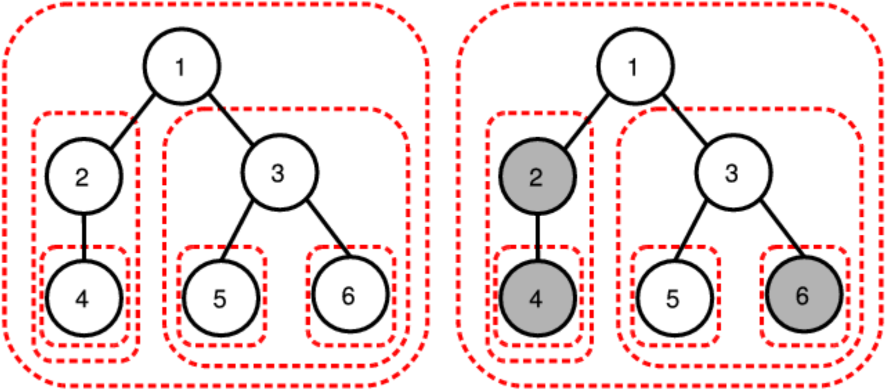
Selection of diamond-shaped patterns on a 2-D grids, $p = 25$.



- It is possible to extent such settings to 3-D space, or more complex topologies.

Hierarchical Norms

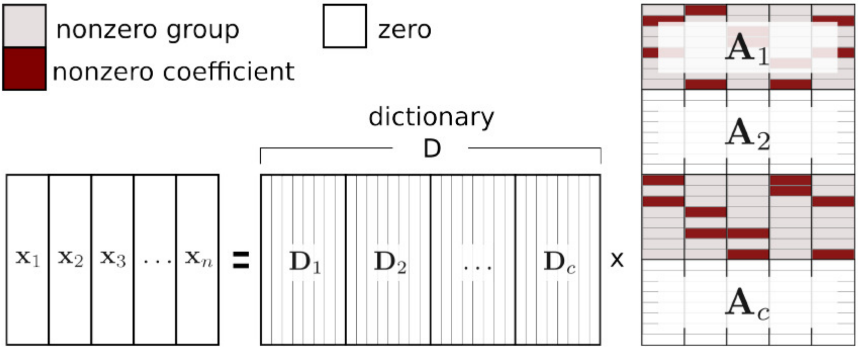
[Zhao et al., 2009, Bach, 2009]



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

Group Lasso + Sparsity

[Sprechmann et al., 2010]



Application 1:
Wavelet denoising with hierarchical norms

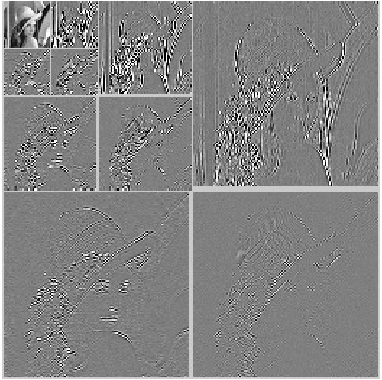
Wavelet denoising with hierarchical norms

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

Classical wavelet denoising [Donoho and Johnstone, 1995]:

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

When $\mathbf{D}$ is orthogonal, the solution is obtained via **soft-thresholding**.



Wavelet denoising with hierarchical norms

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

Wavelet with hierarchical norm: Add **a-priori knowledge** that the coefficients are embedded in a tree.

(a) Barb., $\sigma = 50$, ℓ_1

(b) Barb., $\sigma = 50$, tree

Wavelet denoising with hierarchical norms

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

Benchmark on a database of 12 **standard images:**

		Haar			
	σ	ℓ_0	ℓ_1	Ω_{ℓ_2}	Ω_{ℓ_∞}
PSNR	5	34.48	35.52	**35.89**	35.79
	10	29.63	30.74	**31.40**	31.23
	25	24.44	25.30	**26.41**	26.14
	50	21.53	20.42	**23.41**	23.05
	100	19.27	19.43	**20.97**	20.58
IPSNR	5	-	$1.04 \pm .31$	$\mathbf{1.41} \pm \mathbf{.45}$	$1.31 \pm .41$
	10	-	$1.10 \pm .22$	$\mathbf{1.76} \pm \mathbf{.26}$	$1.59 \pm .22$
	25	-	$.86 \pm .35$	$\mathbf{1.96} \pm \mathbf{.22}$	$1.69 \pm .21$
	50	-	$.46 \pm .28$	$\mathbf{1.87} \pm \mathbf{.20}$	$1.51 \pm .20$
	100	-	$.15 \pm .23$	$\mathbf{1.69} \pm \mathbf{.19}$	$1.30 \pm .19$

Application 2:
Hierarchical Dictionary Learning

Hierarchical Dictionary Learning

[Olshausen and Field, 1997, Elad and Aharon, 2006, Mairal et al., 2010a]

We now consider a sequence $\{\mathbf{y}^i\}_{i=1}^m$, of signals in $\mathbb{R}^n$.

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times m},\mathbf{D}\in\mathcal{C}} \sum_{i=1}^{m} \frac{1}{2}\|\mathbf{y}^i - \mathbf{D}\mathbf{w}^i\|_2^2 + \lambda\|\mathbf{w}^i\|_1,$$

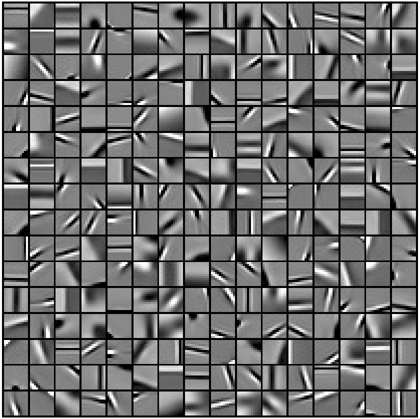
This can be rewritten as a **matrix factorization** problem

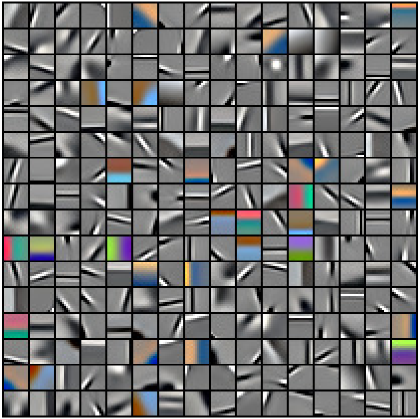
$$\min_{\mathbf{W}\in\mathbb{R}^{p\times m},\mathbf{D}\in\mathcal{C}} \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_{1,1}.$$

[Jenatton, Mairal, Obozinski, and Bach, 2010a]:
What about replacing the ℓ_1-norm by a hierarchical norm?

Hierarchical Dictionary Learning

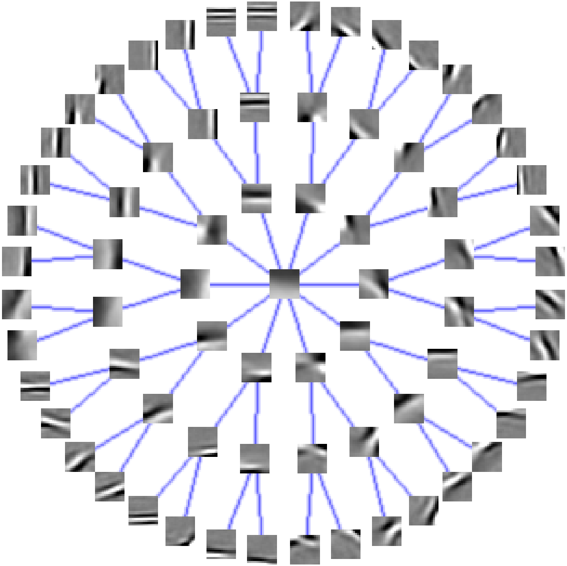
Dictionaries learned with the ℓ_1-norm





Hierarchical Dictionary Learning

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

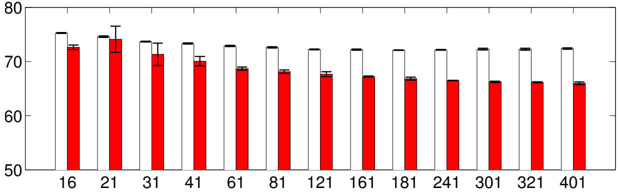


Application to patch reconstrution

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

- Reconstruction of 100,000 8×8 natural images patches
  - Remove randomly subsampled pixels
  - Reconstruct with matrix factorization and structured sparsity

noise	50 %	60 %	70 %	80 %	90 %
flat	19.3 ± 0.1	26.8 ± 0.1	36.7 ± 0.1	50.6 ± 0.0	72.1 ± 0.0
tree	18.6 ± 0.1	25.7 ± 0.1	35.0 ± 0.1	48.0 ± 0.0	65.9 ± 0.3

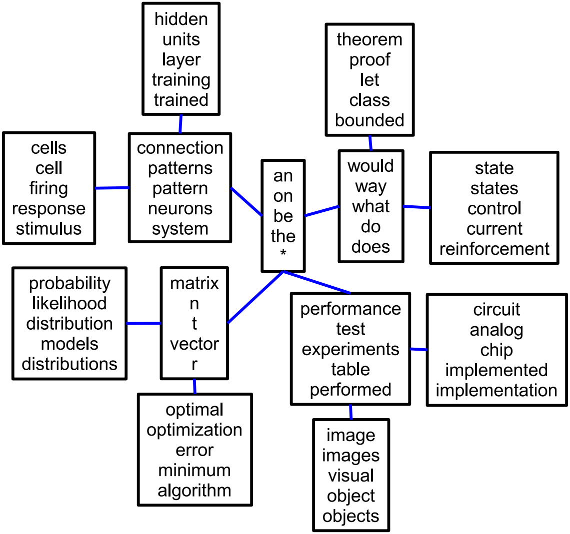


Hierarchical Topic Models for text corpora

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

- Each document is modeled through word counts
- Low-rank matrix factorization of word-document matrix
- Probabilistic topic models such as Latent Dirichlet Allocation [Blei et al., 2003]
- Organise the topics in a tree.
- Previously approached using non-parametric Bayesian methods (Hierarchical Chinese Restaurant Process and nested Dirichlet Process): [Blei et al., 2010]
- **Can we achieve similar performance with simple matrix factorization formulation?**

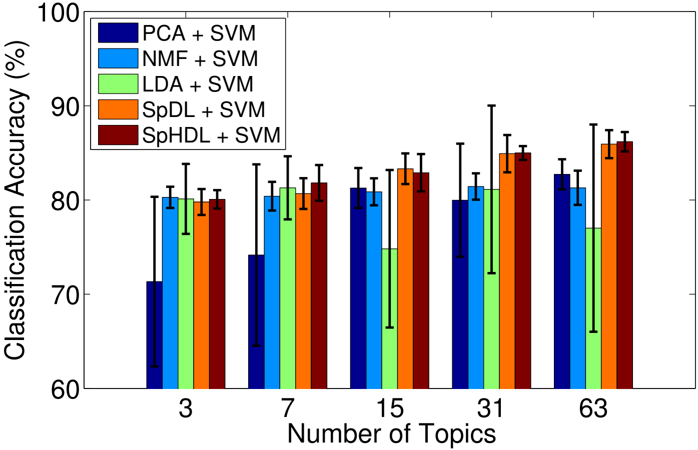
Tree of Topics



Classification based on topics

Comparison on predicting newsgroup article subjects

- 20 newsgroup articles (1425 documents, 13312 words)



Application 3:
Background Subtraction

Background Subtraction

Given a video sequence, how can we remove foreground objects?

video sequence 1

video sequence 2

Background Subtraction

$$\underbrace{\mathbf{x}}_{\text{frame}} \approx \underbrace{\mathbf{D}\mathbf{w}}_{\text{linear combination of background frames}} + \underbrace{\mathbf{e}}_{\text{error term}}.$$

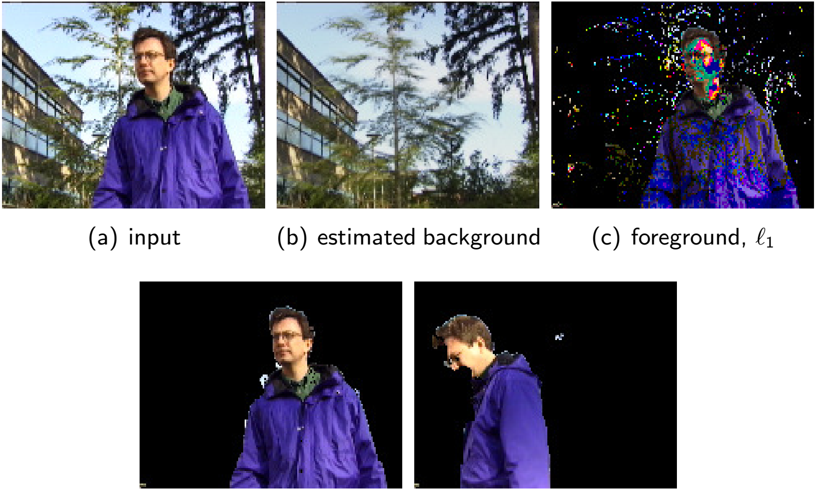
Solved by

$$\min_{\mathbf{w}\in\mathbb{R}^p,\mathbf{e}\in\mathbb{R}^m} \frac{1}{2}\|\mathbf{x}-\mathbf{D}\mathbf{w}-\mathbf{e}\|_2^2 + \lambda_1\|\mathbf{w}\| + \lambda_2\Omega(\mathbf{e}).$$

Same idea as Wright et al. [2009] for robust face recognition, where $\Omega = \ell_1$.

We are going to use overlapping groups with 3×3 neighborhoods to add spatial consistency. See also Cehver et al. [2008] for structured sparsity + background subtraction.

Background Subtraction

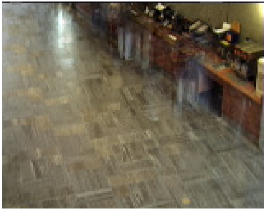


(a) input (b) estimated background (c) foreground, ℓ_1

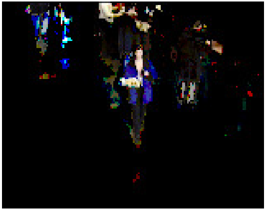
(d) foreground, ℓ_1+struct (e) other example

Background Subtraction

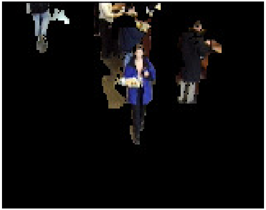
(a) input



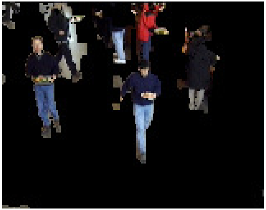
(b) estimated background



(c) foreground, ℓ_1



(d) foreground, ℓ_1+struct



(e) other example

How do we optimize all that?

First-order/proximal methods

$$\min_{\mathbf{w}\in\mathbb{R}^p} f(\mathbf{w}) + \lambda\Omega(\mathbf{w})$$

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

$$\mathbf{w}^{k+1} \leftarrow \arg\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{f(\mathbf{w}^k) + \nabla f(\mathbf{w}^k)^\top(\mathbf{w} - \mathbf{w}^k)}_{\text{linear approximation}} + \underbrace{\frac{L}{2}\|\mathbf{w} - \mathbf{w}^k\|_2^2}_{\text{quadratic term}} + \lambda\Omega(\mathbf{w})$$

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

When $\lambda = 0$, $\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k - \frac{1}{L}\nabla f(\mathbf{w}^k)$, this is equivalent to a classical gradient descent step.

First-order/proximal methods

- They require solving efficiently the proximal operator

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{w}\|_2^2 + \lambda\Omega(\mathbf{w})$$

- For the ℓ_1-norm, this amounts to a soft-thresholding:

$$\mathbf{w}_i^\star = \mathrm{sign}(\mathbf{u}_i)(\mathbf{u}_i - \lambda)^+.$$

- There exists accelerated versions based on Nesterov optimal first-order method (gradient method with "extrapolation") [Beck and Teboulle, 2009, Nesterov, 2007, 1983]
- suited for large-scale experiments.

Tree-structured groups

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

- 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 Prox_g and Prox_Ω as

$$\text{Prox}_g : \mathbf{u} \rightarrow \arg\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{u} - \mathbf{w}\| + \lambda \|\mathbf{w}_g\|_q,$$

$$\text{Prox}_\Omega : \mathbf{u} \rightarrow \arg\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{u} - \mathbf{w}\| + \lambda \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_q,$$

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

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

$\rightarrow$ Tree-structured regularization : Efficient linear time algorithm.

General Overlapping Groups for $q = \infty$

Dual formulation [Jenatton, Mairal, Obozinski, and Bach, 2010a]

The solutions $\mathbf{w}^\star$ and $\boldsymbol{\xi}^\star$ of the following optimization problems

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{u} - \mathbf{w}\| + \lambda \|\mathbf{w}_g\|_\infty, \tag{Primal}$$

$$\min_{\boldsymbol{\xi} \in \mathbb{R}^{p \times |\mathcal{G}|}} \frac{1}{2}\|\mathbf{u} - \sum_{g \in \mathcal{G}} \boldsymbol{\xi}^g\|_2^2 \ \text{ s.t. } \ \forall g \in \mathcal{G}, \ \|\boldsymbol{\xi}^g\|_1 \leq \lambda \ \text{ and } \ \boldsymbol{\xi}_j^g = 0 \text{ if } j \notin g, \tag{Dual}$$

satisfy

$$\mathbf{w}^\star = \mathbf{u} - \sum_{g \in \mathcal{G}} \boldsymbol{\xi}^{\star g}. \tag{Primal-dual relation}$$

The dual formulation has more variables, but **no overlapping constraints**.

General Overlapping Groups for $q = \infty$

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

First Step: Flip the signs of $\mathbf{u}$

The dual is equivalent to a **quadratic min-cost flow problem**.

$$\min_{\boldsymbol{\xi} \in \mathbb{R}_+^{p \times |\mathcal{G}|}} \frac{1}{2} \|\mathbf{u} - \sum_{g \in \mathcal{G}} \boldsymbol{\xi}^g\|_2^2 \ \text{ s.t. } \ \forall g \in \mathcal{G}, \sum_{j \in g} \boldsymbol{\xi}_j^g \leq \lambda \ \text{ and } \ \boldsymbol{\xi}_j^g = 0 \ \text{ if } j \notin g,$$

General Overlapping Groups for $q = \infty$

Example: $\mathcal{G} = \{g = \{1, \ldots, p\}\}$

$$\min_{\boldsymbol{\xi}^g \in \mathbb{R}_+^p} \frac{1}{2}\|\mathbf{u} - \boldsymbol{\xi}^g\|_2^2 \ \text{ s.t. } \ \sum_{j=1}^{p} \xi_j^g \leq \lambda.$$

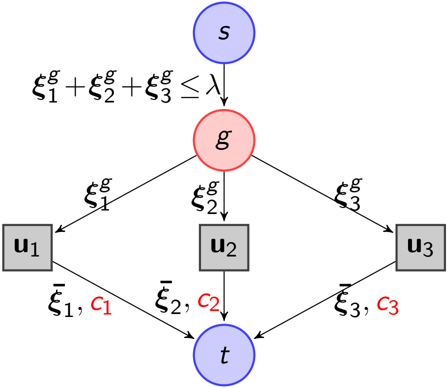


Figure: $\mathcal{G} = \{g = \{1, 2, 3\}\}$, $\forall j$, $c_j = \frac{1}{2}(\mathbf{u}_j - \bar{\xi}_j)^2$.

General Overlapping Groups for $q = \infty$

Example with two overlapping groups

$$\min_{\boldsymbol{\xi} \in \mathbb{R}_+^{p \times |\mathcal{G}|}} \frac{1}{2} \|\mathbf{u} - \sum_{g \in \mathcal{G}} \boldsymbol{\xi}^g \|_2^2 \ \text{ s.t. } \ \forall g \in \mathcal{G}, \sum_{j \in g} \boldsymbol{\xi}_j^g \leq \lambda \ \text{ and } \ \boldsymbol{\xi}_j^g = 0 \text{ if } j \notin g,$$

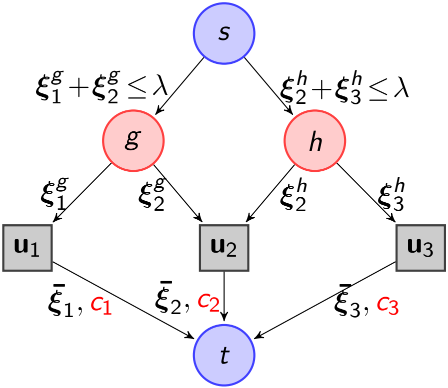


Figure: $\mathcal{G} = \{g = \{1,2\}, h = \{2,3\}\}$, $\forall j$, $c_j = \frac{1}{2}(\mathbf{u}_j - \bar{\xi}_j)^2$.

General Overlapping Groups for $q = \infty$

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

Main ideas of the algorithm: Divide and conquer

1. Solve a relaxed problem in linear time.
2. Test the feasability of the solution for the “non-relaxed” problem with a max-flow.
3. If the solution is feasible, it is optimal and stop the algorithm.
4. If not, find a minimum cut and removes the arcs along the cut.
5. Recursively process each part of the graph.

The algorithm is guaranteed to converge to the solution.
See more details in the paper

Conclusions

- We have developed efficient and large-scale algorithmic tools for solving structured sparse decomposition problems.
- These tools are related to network flow optimization.
- The hierarchical case can be solved at the same cost as ℓ_1.
- There are preliminary applications in computer vision, there should be more!

References I

F. Bach. High-dimensional non-linear variable selection through hierarchical kernel learning. Technical report, arXiv:0909.0844, 2009.

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.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

D. Blei, T. Griffiths, and M. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2): 1–30, 2010.

V. Cehver, M. F. Duarte, C. Hegde, and R. G. Baraniuk. Sparse signal recovery usingmarkov random fields. In *Advances in Neural Information Processing Systems*, 2008.

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

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.

References II

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 54(12): 3736–3745, December 2006.

R. Jenatton, J-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. Technical report, 2009. preprint arXiv:0904.3523v1.

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, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. Technical report, 2010b. submitted, arXiv:1009.3139.

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

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

Y. Nesterov. A method for solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Dokl.*, 27:372–376, 1983.

Y. Nesterov. Gradient methods for minimizing composite objective function. Technical report, CORE, 2007.

References III

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

P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar. Collaborative hierarchical sparse modeling. Technical report, 2010. Preprint arXiv:1003.0400v1.

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

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*, pages 210–227, 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.

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