Sparse Coding and Dictionary Learning in Image Analysis

Presentation uses slides by Julien Mairal

http://www.di.ens.fr/willow/events/cvml2010/materials/INRIA_summer_school_2010_Julien.pdf

19.10.2017

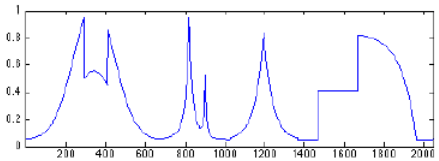
The Image Denoising Problem

$$\underbrace{\mathbf{y}}_{\text{measurements}} = \underbrace{\mathbf{x}_{orig}}_{\text{original image}} + \underbrace{\mathbf{w}}_{\text{noise}}$$

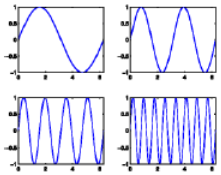
http://di.ens.fr/w...2010_Julien.pdf

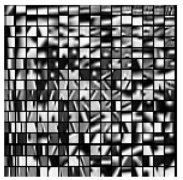
What is a Sparse Linear Model?

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



Let $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_p] \in \mathbb{R}^{m \times p}$ 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** α in $\mathbb{R}^p$ such that $\mathbf{y} \approx \mathbf{D}\alpha$. We call α the **sparse code**.

$$\underbrace{\begin{pmatrix} \mathbf{y} \end{pmatrix}}_{\mathbf{y} \in \mathbb{R}^m} \approx \underbrace{\left(\begin{array}{c|c|c|c} \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_p \end{array} \right)}_{\mathbf{D} \in \mathbb{R}^{m \times p}} \underbrace{\begin{pmatrix} \alpha[1] \\ \alpha[2] \\ \vdots \\ \alpha[p] \end{pmatrix}}_{\alpha \in \mathbb{R}^p, \text{sparse}}$$

http://di.ens.fr/w...2010_Julien.pdf

The Sparse Decomposition Problem

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \underbrace{\frac{1}{2}\|\mathbf{y}-\mathbf{D}\boldsymbol{\alpha}\|_2^2}_{\text{data fitting term}} + \underbrace{\lambda\psi(\boldsymbol{\alpha})}_{\substack{\text{sparsity-inducing}\\ \text{regularization}}}$$

ψ induces sparsity in $\boldsymbol{\alpha}$. It can be

- the ℓ_0 "pseudo-norm". $\|\boldsymbol{\alpha}\|_0 \triangleq \#\{i \text{ s.t. } \boldsymbol{\alpha}[i] \neq 0\}$ (NP-hard)
- the ℓ_1 norm. $\|\boldsymbol{\alpha}\|_1 \triangleq \sum_{i=1}^p |\boldsymbol{\alpha}[i]|$ (convex),
- ...

http://di.ens.fr/w...2010_Julien.pdf

Solving the denoising problem

[Elad and Aharon, 2006]

- Extract all overlapping 8×8 patches $\mathbf{y}_i$.
- Solve a matrix factorization problem:

$$\min_{\boldsymbol{\alpha}_i, \mathbf{D} \in \mathcal{C}} \sum_{i=1}^{n} \underbrace{\frac{1}{2} \|\mathbf{y}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2}_{\text{reconstruction}} + \underbrace{\lambda \psi(\boldsymbol{\alpha}_i)}_{\text{sparsity}},$$

  with $n > 100,000$

- Average the reconstruction of each patch.

http://di.ens.fr/w...2010_Julien.pdf

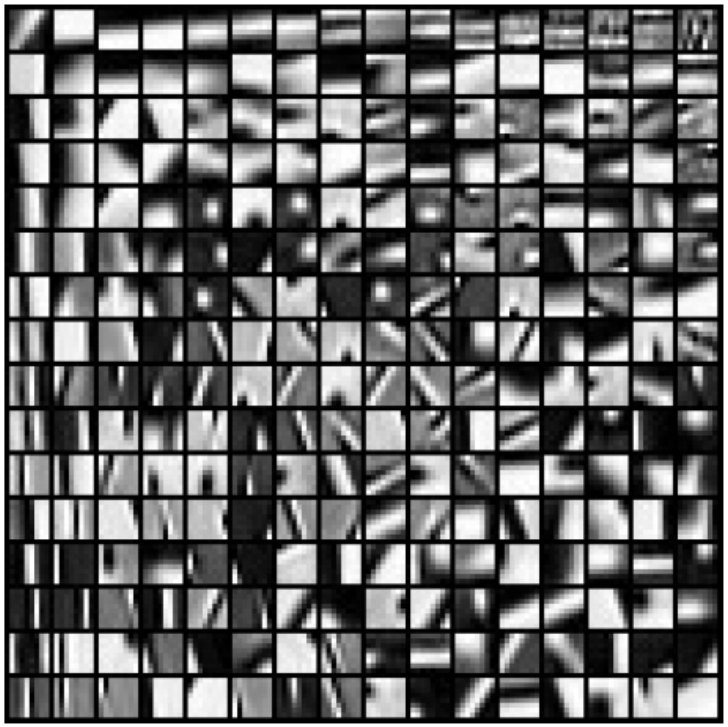
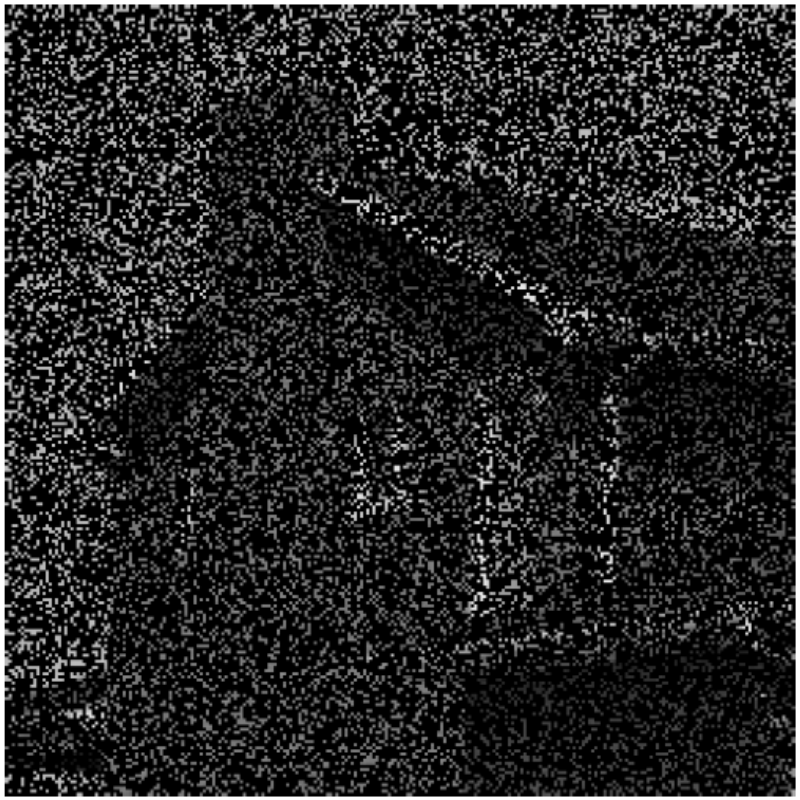


Figure: Dictionary trained on a noisy version of the image

http://di.ens.fr/w...2010_Julien.pdf

Image restoration



http://di.ens.fr/w...2010_Julien.pdf

Inverse half-toning (Original)

http://di.ens.fr/w...2010_Julien.pdf

Inverse half-toning (Reconstructed)

http://di.ens.fr/w...2010_Julien.pdf

Inverse half-toning (Original)

http://di.ens.fr/w...2010_Julien.pdf

Inverse half-toning (Reconstructed)

http://di.ens.fr/w...2010_Julien.pdf

Important notes

- Dictionary Learning adapts to the data you want to restore.
- Dictionary Learning is well adapted to data that admit sparse representation. **Sparsity is for sparse data only**.

http://di.ens.fr/w...2010_Julien.pdf

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

In a linear model, one assumes a relation $y \approx \mathbf{w}^\top \mathbf{x}$ (or $y \approx \text{sign}(\mathbf{w}^\top \mathbf{x})$), and solves

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

http://di.ens.fr/w...2010_Julien.pdf

Sparse Linear Models: the Lasso

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

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

- 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,... as long as one wishes to **select** features.

http://di.ens.fr/w...2010 Julien.pdf

Optimization for Dictionary Learning

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

$$\mathcal{C} \triangleq \{\mathbf{D} \in \mathbb{R}^{m\times p} \text{ s.t. } \forall j = 1,\ldots,p, \quad \|\mathbf{d}_j\|_2 \leq 1\}.$$

- Classical optimization alternates between $\mathbf{D}$ and $\boldsymbol{\alpha}$.
- Good results, but slow!

http://di.ens.fr/w...2010_Julien.pdf

Inpainting a photo (Original)

http://di.ens.fr/w...2010_Julien.pdf

Inpainting a photo (Reconstructed)

http://di.ens.fr/w...2010_Julien.pdf

Matrix Factorization Problems and Dictionary Learning

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

can be rewritten

$$\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^{p\times n}\\ \mathbf{D}\in\mathcal{C}}} \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\boldsymbol{\alpha}\|_F^2 + \lambda\|\boldsymbol{\alpha}\|_1,$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$ and $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]$.

http://di.ens.fr/w...2010_Julien.pdf

Matrix Factorization: Principal Components Analysis

$$\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^{p\times n}\\ \mathbf{D}\in\mathbb{R}^{m\times p}}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\boldsymbol{\alpha}\|_F^2 \quad \text{s.t.} \quad \mathbf{D}^\top\mathbf{D}=\mathbf{I} \text{ and } \boldsymbol{\alpha}\boldsymbol{\alpha}^\top \text{ is diagonal.}$$

$\mathbf{D}=[\mathbf{d}_1,\ldots,\mathbf{d}_p]$ are the principal components.

http://di.ens.fr/w...2010_Julien.pdf

Matrix Factorization: Hard Clustering

$$\min_{\substack{\boldsymbol{\alpha}\in\{0,1\}^{p\times n}\\ \mathbf{D}\in\mathbb{R}^{m\times p}}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\boldsymbol{\alpha}\|_F^2 \quad \text{s.t.} \quad \forall i\in\{1,\ldots,p\}, \ \sum_{j=1}^{p}\boldsymbol{\alpha}_i[j]=1.$$

$\mathbf{D}=[\mathbf{d}_1,\ldots,\mathbf{d}_p]$ are the centroids of the p clusters.

http://di.ens.fr/w...2010_Julien.pdf

Matrix Factorization: Soft Clustering

$$\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}_+^{p\times n}\\ \mathbf{D}\in\mathbb{R}^{m\times p}}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\boldsymbol{\alpha}\|_F^2, \quad \text{s.t.} \quad \forall i\in\{1,\ldots,p\}, \ \sum_{j=1}^{p}\boldsymbol{\alpha}_i[j]=1.$$

$\mathbf{D}=[\mathbf{d}_1,\ldots,\mathbf{d}_p]$ are the centroids of the p clusters.

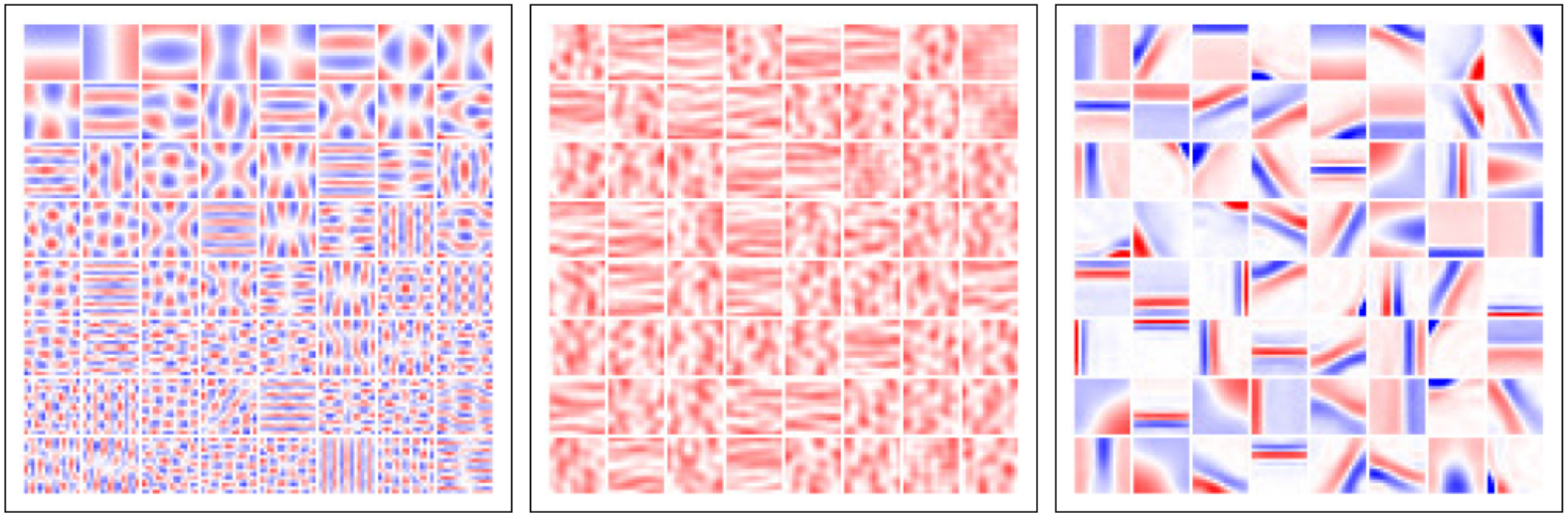
http://di.ens.fr/w...2010_Julien.pdf

Non-negative Matrix Factorization + sparsity

$$\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}_+^{p\times n}\\ \mathbf{D}\in\mathbb{R}_+^{m\times p}}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\boldsymbol{\alpha}\|_F^2+\lambda\|\boldsymbol{\alpha}\|_1.$$

http://di.ens.fr/w...2010_Julien.pdf

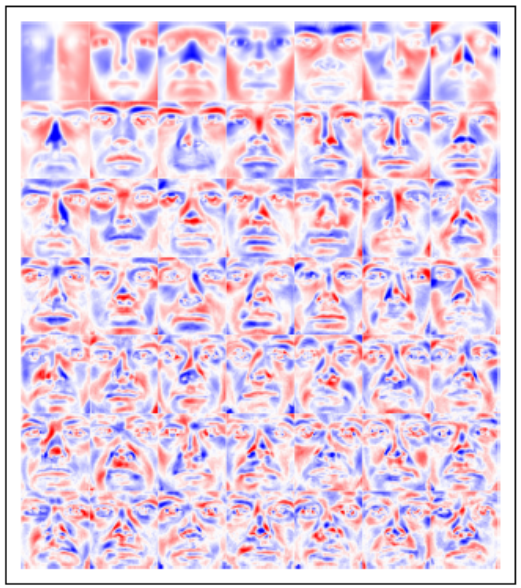
Natural patches



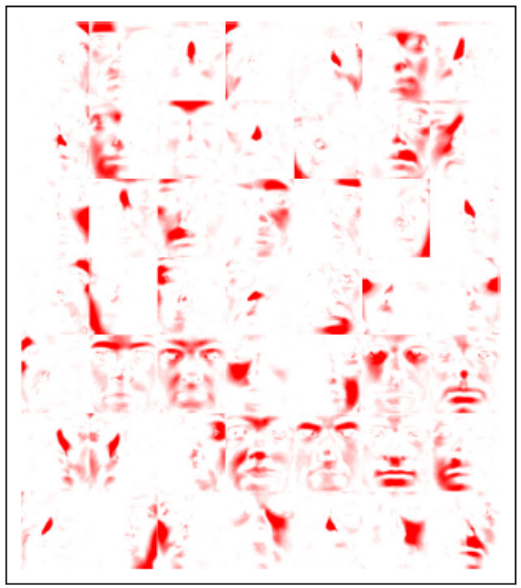
(a) PCA (b) NNMF (c) DL

http://di.ens.fr/w...2010_Julien.pdf

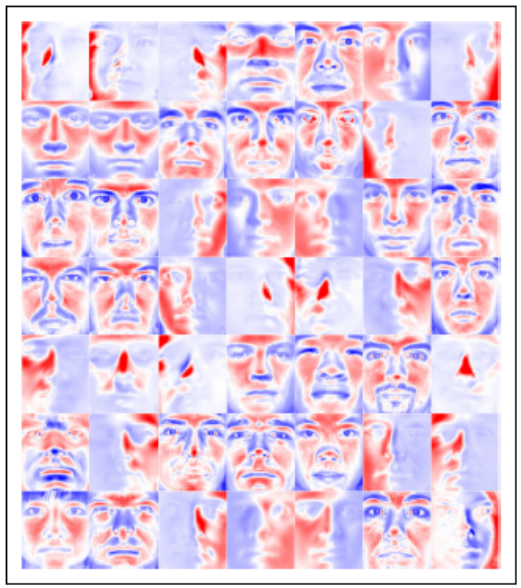
Faces



(d) PCA



(e) NNMF



(f) DL

http://di.ens.fr/w...2010_Julien.pdf

Learning Codebooks for Image Classification

Let an image be represented by a set of low-level descriptors $\mathbf{y}_i$ at N locations identified with their indices $i = 1, \ldots, N$.

- hard-quantization:

$$\mathbf{y}_i \approx \mathbf{D}\boldsymbol{\alpha}_i, \quad \boldsymbol{\alpha}_i \in \{0,1\}^p \text{ and } \sum_{j=1}^{p} \boldsymbol{\alpha}_i[j] = 1$$

- soft-quantization:

$$\boldsymbol{\alpha}_i[j] = \frac{e^{-\beta\|\mathbf{y}_i - \mathbf{d}_j\|_2^2}}{\sum_{k=1}^{p} e^{-\beta\|\mathbf{y}_i - \mathbf{d}_k\|_2^2}}$$

- sparse coding:

$$\mathbf{y}_i \approx \mathbf{D}\boldsymbol{\alpha}_i, \quad \boldsymbol{\alpha}_i = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y}_i - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

http://di.ens.fr/w...2010...Julien.pdf

Learning dictionaries with a discriminative cost function

Idea:

Let us consider 2 sets S_-, S_+ of signals representing 2 different classes. Each set should admit a dictionary best adapted to its reconstruction.

Classification procedure for a signal $\mathbf{y} \in \mathbb{R}^n$:

$$\min(\mathbf{R}^\star(\mathbf{y}, \mathbf{D}_-), \mathbf{R}^\star(\mathbf{y}, \mathbf{D}_+))$$

where

$$\mathbf{R}^\star(\mathbf{y}, \mathbf{D}) = \min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \text{ s.t. } \|\boldsymbol{\alpha}\|_0 \leq L.$$

"Reconstructive" training

$$\begin{cases} \min_{\mathbf{D}_-} \sum_{i \in S_-} \mathbf{R}^\star(\mathbf{y}_i, \mathbf{D}_-) \\ \min_{\mathbf{D}_+} \sum_{i \in S_+} \mathbf{R}^\star(\mathbf{y}_i, \mathbf{D}_+) \end{cases}$$

[Grosse et al., 2007], [Huang and Aviyente, 2006],
[Sprechmann et al., 2010b] for unsupervised clustering (CVPR '10)

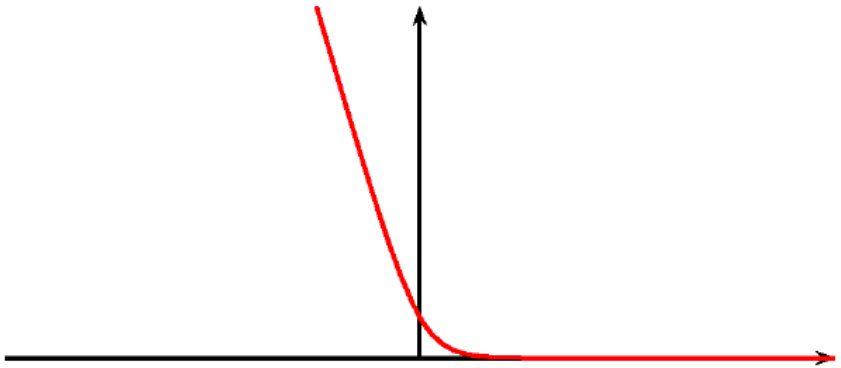
Learning dictionaries with a discriminative cost function

"Discriminative" training

[Mairal, Bach, Ponce, Sapiro, and Zisserman, 2008a]

$$\min_{\mathbf{D}_-,\mathbf{D}_+} \sum_i \mathcal{C}\Big(\lambda z_i \big(\mathbf{R}^\star(\mathbf{y}_i,\mathbf{D}_-) - \mathbf{R}^\star(\mathbf{y}_i,\mathbf{D}_+)\big)\Big),$$

where $z_i \in \{-1,+1\}$ is the label of $\mathbf{y}_i$.



Logistic regression function

http://di.ens.fr/w...2010_Julien.pdf

Examples of dictionaries

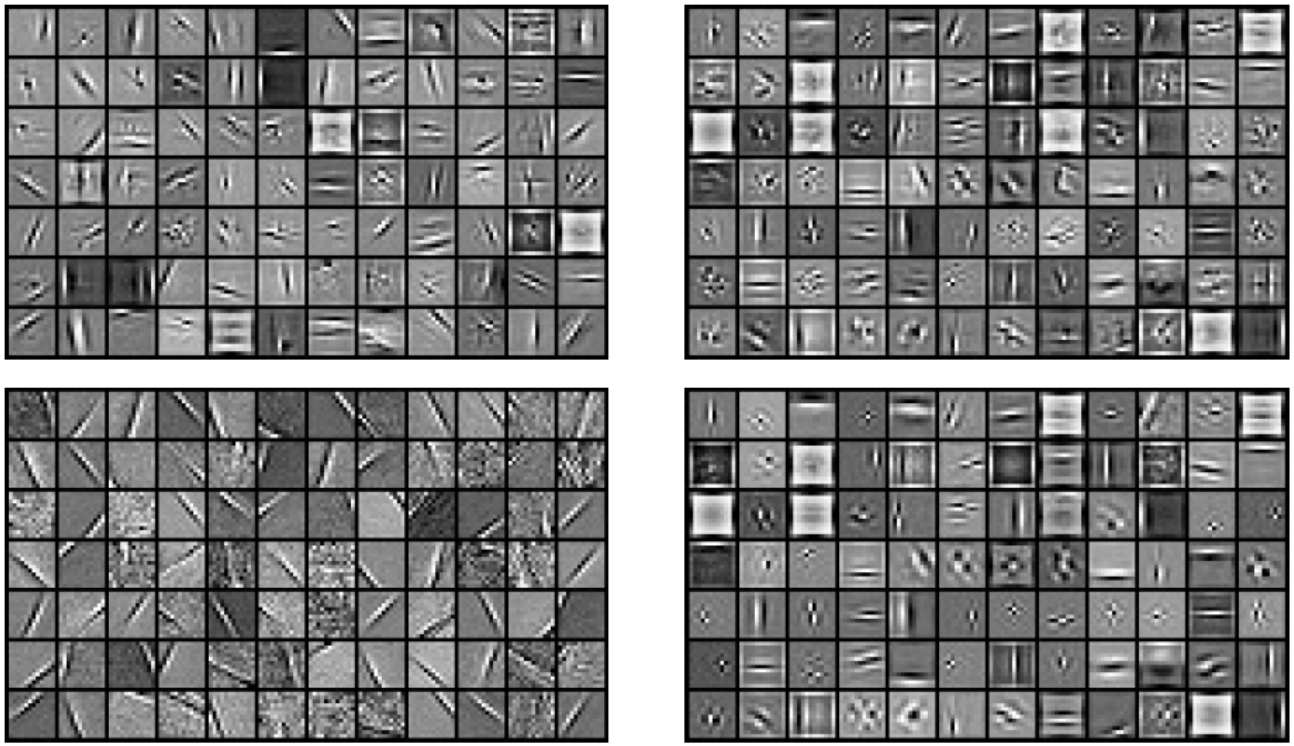
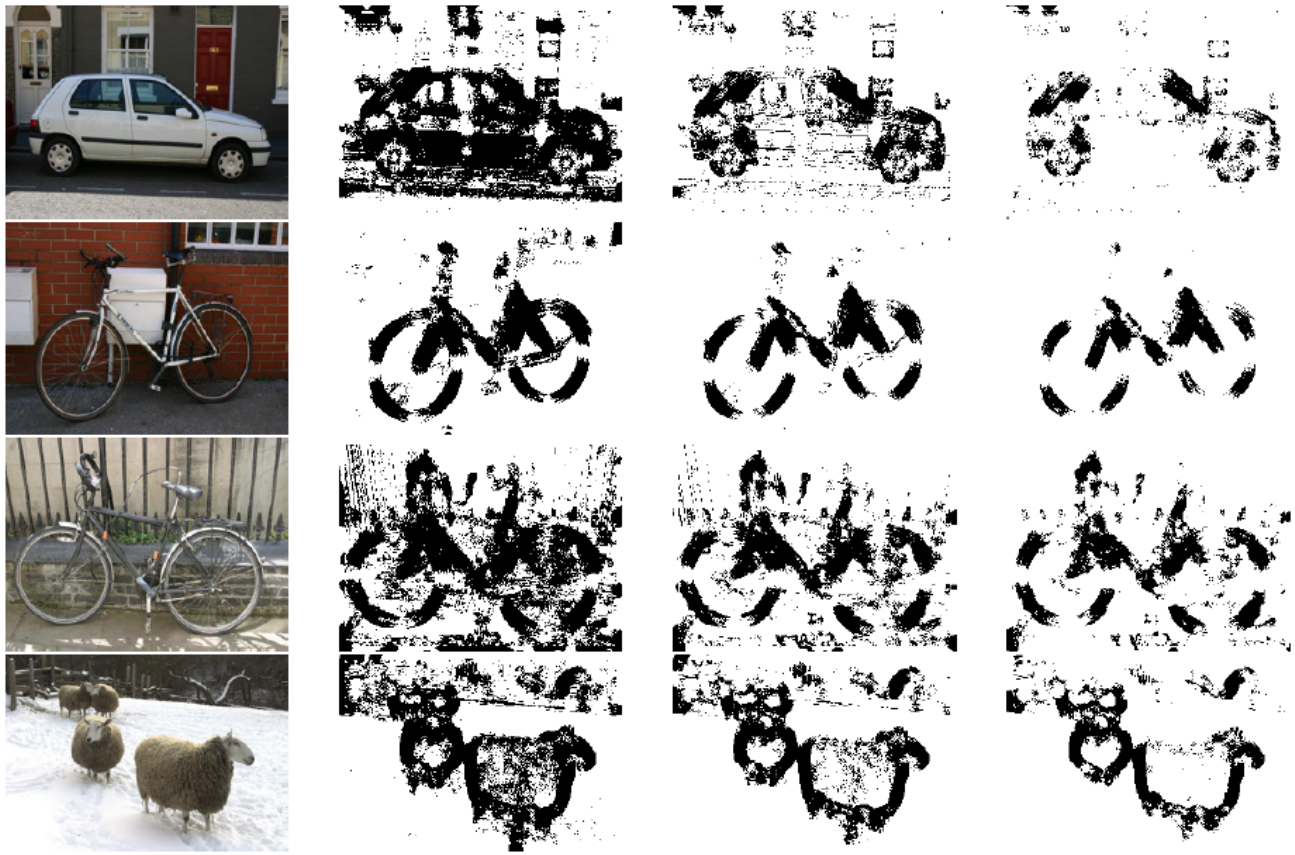


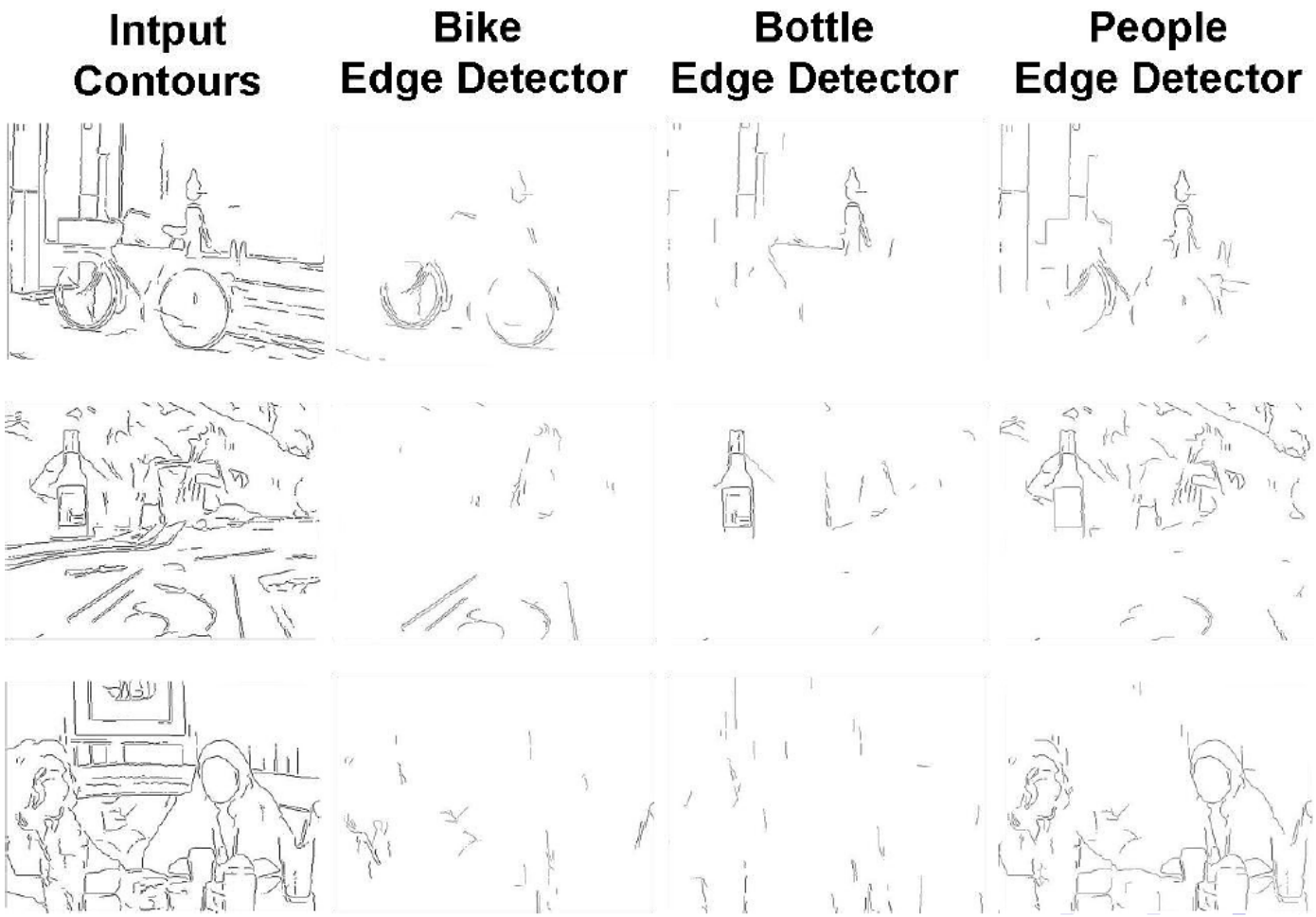
Figure: Top: reconstructive, Bottom: discriminative, Left: Bicycle, Right: Background

Pixelwise classification



http://di.ens.fr/w...2010_Julien.pdf

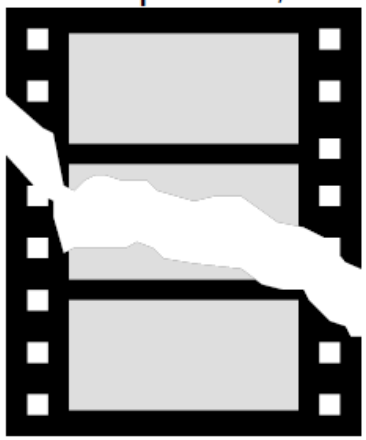
Application to edge detection and classification

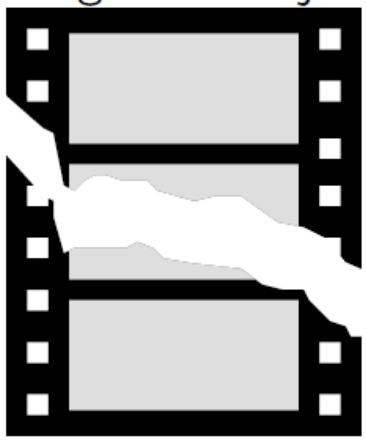


http://di.ens.fr/w...2010...Julien.pdf

Background Subtraction

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





http://di.ens.fr/w...2010_Julien.pdf

Background Subtraction

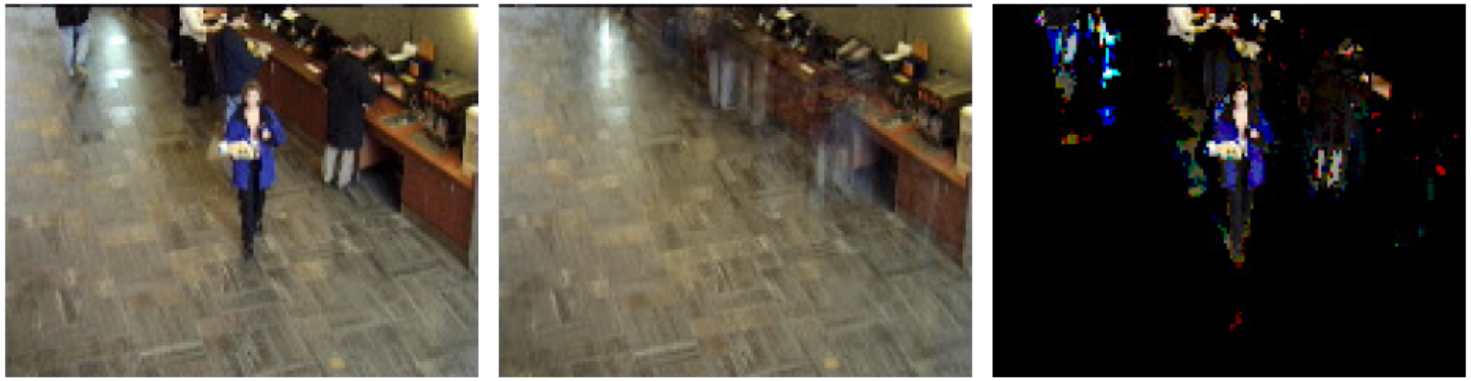
$$\underbrace{\mathbf{y}}_{\text{frame}} \approx \underbrace{\mathbf{D}\boldsymbol{\alpha}}_{\text{linear combination of background frames}} + \underbrace{\mathbf{e}}_{\text{error term}}.$$

Solved by

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p, \mathbf{e}\in\mathbb{R}^m} \frac{1}{2}\|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha} - \mathbf{e}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\| + \lambda_2\psi(\mathbf{e}).$$

http://di.ens.fr/w...2010_Julien.pdf

Background Subtraction



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

http://di.ens.fr/w...2010_Julien.pdf

Recall: The Sparse Decomposition Problem

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \underbrace{\frac{1}{2}\|\mathbf{y}-\mathbf{D}\boldsymbol{\alpha}\|_2^2}_{\text{data fitting term}} + \underbrace{\lambda\psi(\boldsymbol{\alpha})}_{\substack{\text{sparsity-inducing}\\ \text{regularization}}}$$

ψ induces sparsity in $\boldsymbol{\alpha}$. It can be

- the ℓ_0 "pseudo-norm". $\|\boldsymbol{\alpha}\|_0 \triangleq \#\{i \text{ s.t. } \boldsymbol{\alpha}[i] \neq 0\}$ (NP-hard)
- the ℓ_1 norm. $\|\boldsymbol{\alpha}\|_1 \triangleq \sum_{i=1}^p |\boldsymbol{\alpha}[i]|$ (convex)
- ...

This is a selection problem.

http://di.ens.fr/w...2010_Julien.pdf

The Sparse Decomposition Problem

The main class of methods are

- greedy procedures [Mallat and Zhang, 1993], [Weisberg, 1980]
- homotopy [Osborne et al., 2000], [Efron et al., 2004], [Markowitz, 1956]
- soft-thresholding based methods [Fu, 1998], [Daubechies et al., 2004], [Friedman et al., 2007], [Nesterov, 2007], [Beck and Teboulle, 2009], ...
- reweighted-ℓ_2 methods [Daubechies et al., 2009],...
- active-set methods [Roth and Fischer, 2008].
- ...

http://di.ens.fr/w...2010_Julien.pdf

Matching Pursuit

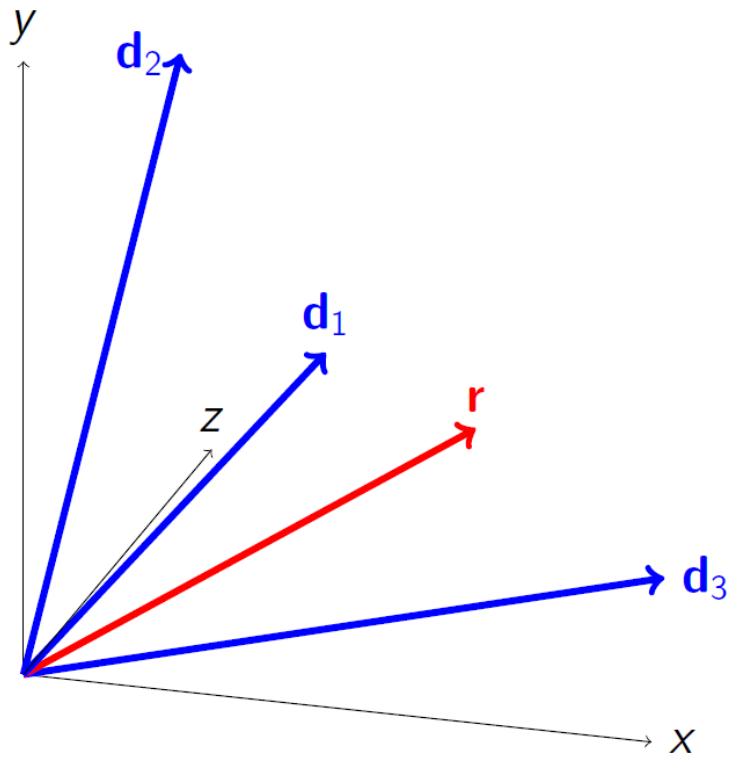


Figure: $\boldsymbol{\alpha} = (0, 0, 0)$

http://di.ens.fr/w...2010_Julien.pdf

Matching Pursuit

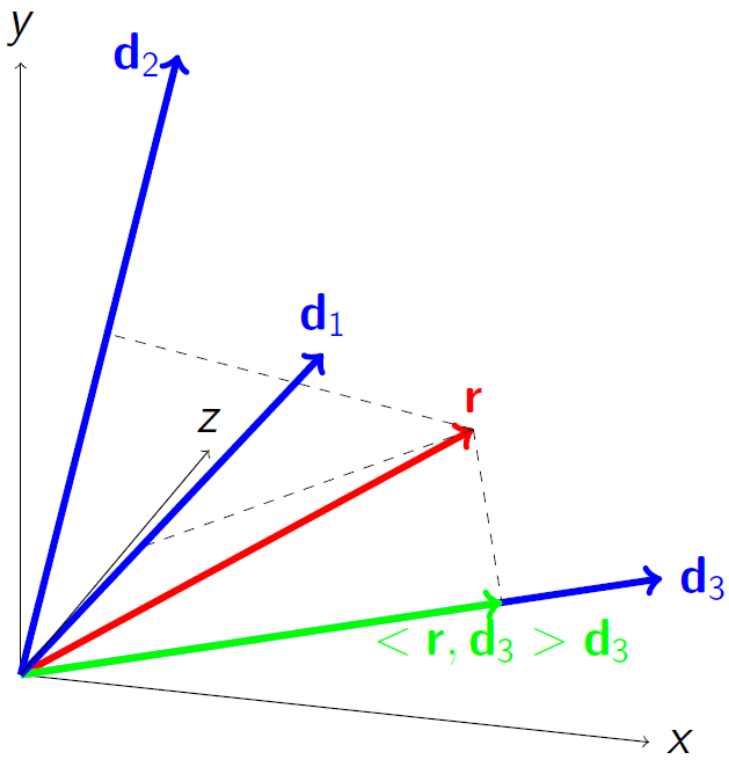


Figure: $\boldsymbol{\alpha} = (0, 0, 0)$

http://di.ens.fr/w...2010_Julien.pdf

Matching Pursuit

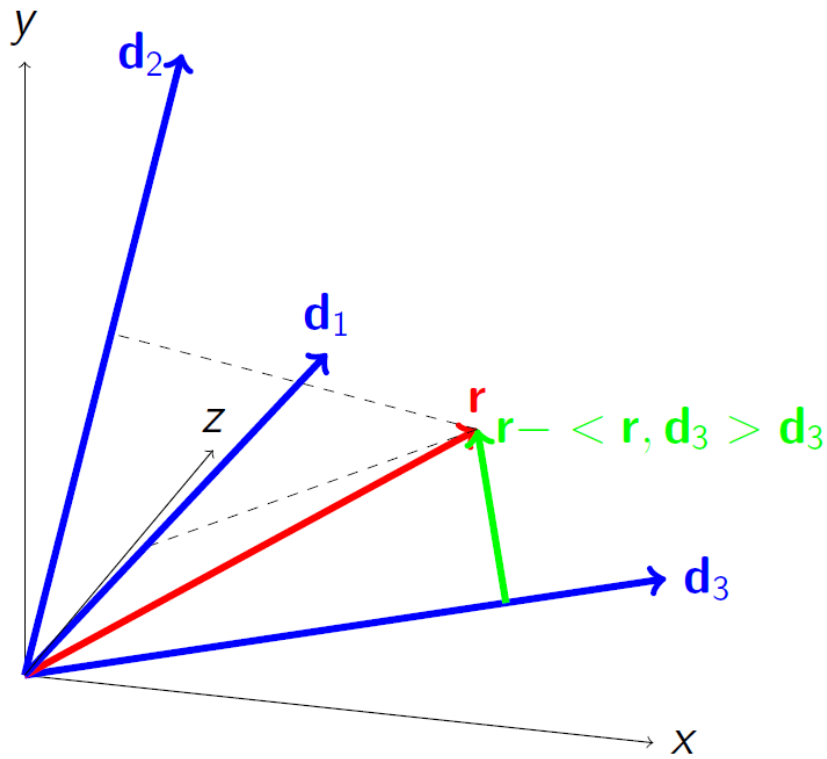


Figure: $\alpha = (0, 0, 0)$

http://di.ens.fr/w...2010_Julien.pdf

Matching Pursuit

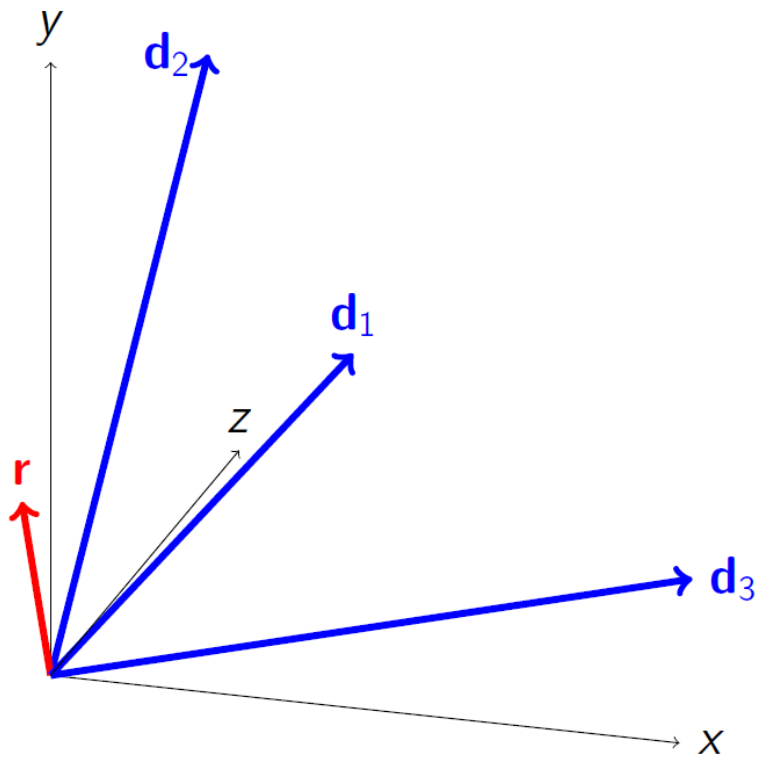


Figure: $\alpha = (0, 0, 0.75)$

http://di.ens.fr/w...2010_Julien.pdf

Matching Pursuit

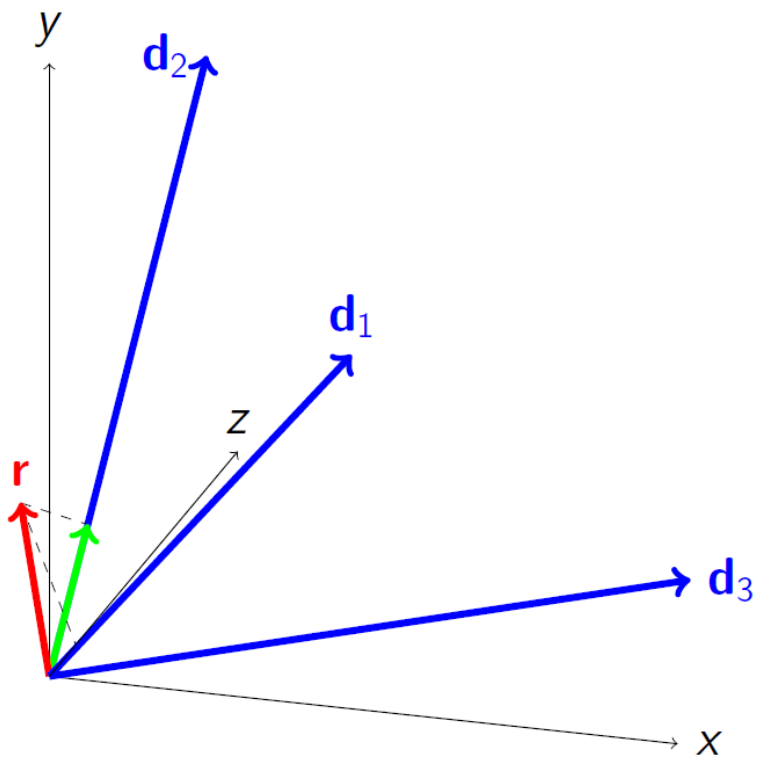


Figure: $\alpha = (0, 0, 0.75)$

http://di.ens.fr/w...2010_Julien.pdf

Matching Pursuit

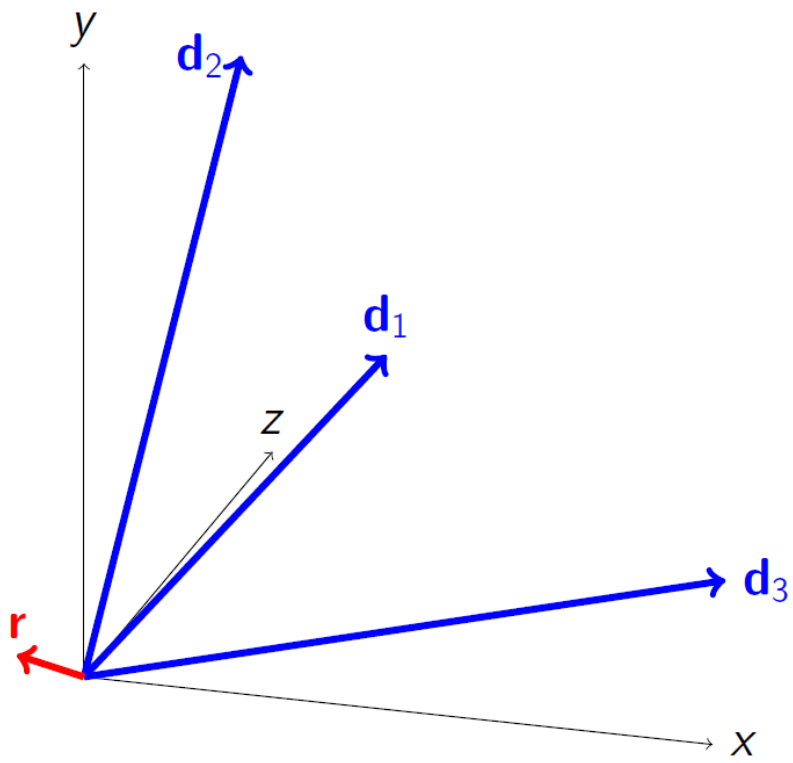


Figure: $\alpha = (0, 0.24, 0.75)$

http://di.ens.fr/w...2010_Julien.pdf

Matching Pursuit

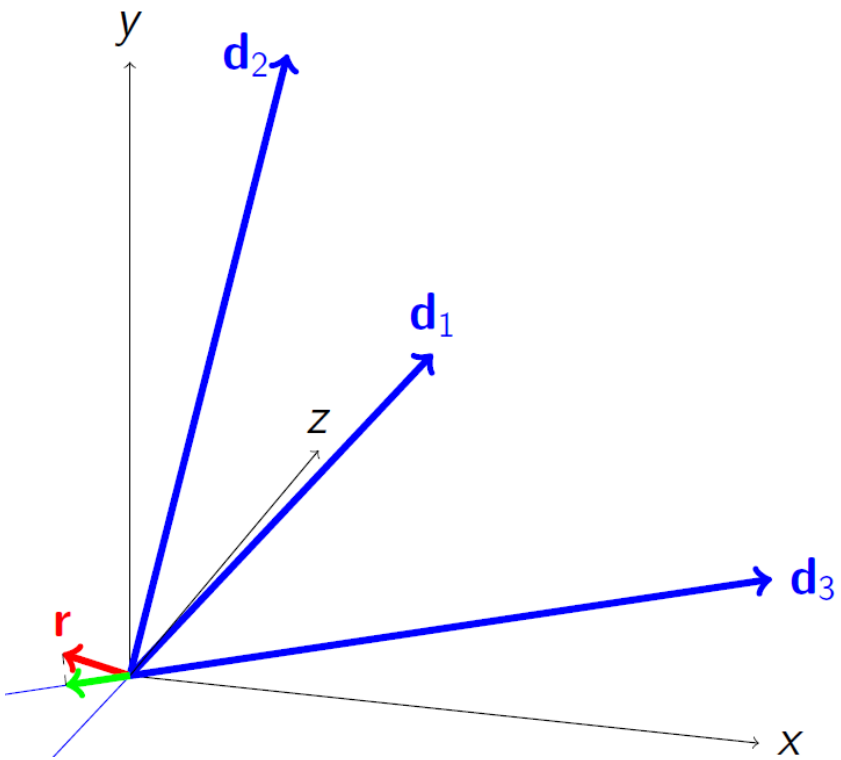


Figure: $\alpha = (0, 0.24, 0.75)$

http://di.ens.fr/w...2010_Julien.pdf

Matching Pursuit

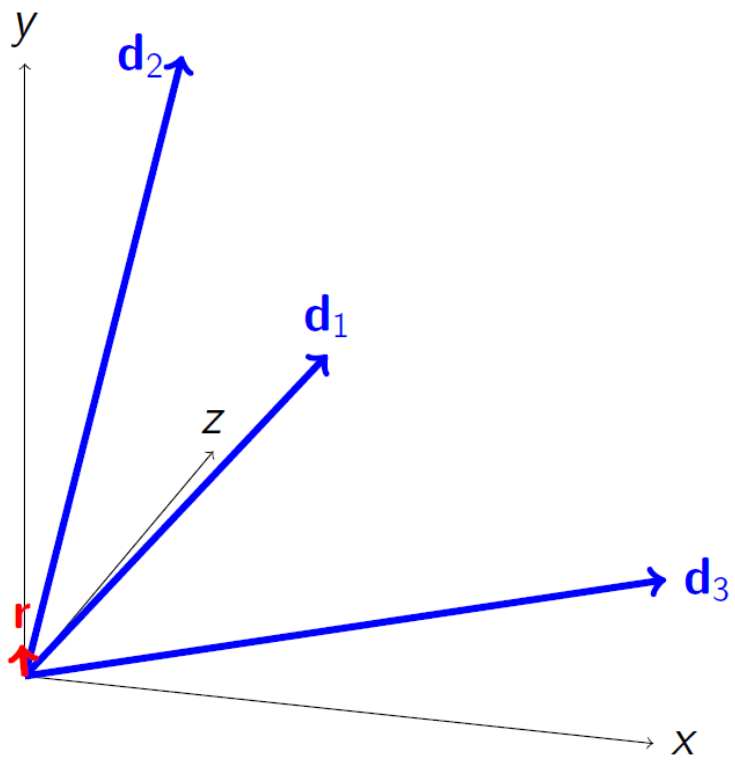


Figure: $\alpha = (0, 0.24, 0.65)$

http://di.ens.fr/w...2010_Julien.pdf

Matching Pursuit

$$\min_{\alpha \in \mathbb{R}^p} \| \underbrace{\mathbf{y} - \mathbf{D}\alpha}_{\mathbf{r}} \|_2^2 \quad \text{s.t.} \quad \|\alpha\|_0 \leq L$$

1: $\alpha \leftarrow 0$
2: $\mathbf{r} \leftarrow \mathbf{y}$ (residual).
3: **while** $\|\alpha\|_0 < L$ **do**
4: Select the atom with maximum correlation with the residual

$$\hat{i} \leftarrow \arg\max_{i=1,\ldots,p} |\mathbf{d}_i^T \mathbf{r}|$$

5: Update the residual and the coefficients

$$\begin{aligned} \alpha[\hat{i}] &\leftarrow \alpha[\hat{i}] + \mathbf{d}_{\hat{i}}^T \mathbf{r} \\ \mathbf{r} &\leftarrow \mathbf{r} - (\mathbf{d}_{\hat{i}}^T \mathbf{r}) \mathbf{d}_{\hat{i}} \end{aligned}$$

6: **end while**

Orthogonal Matching Pursuit

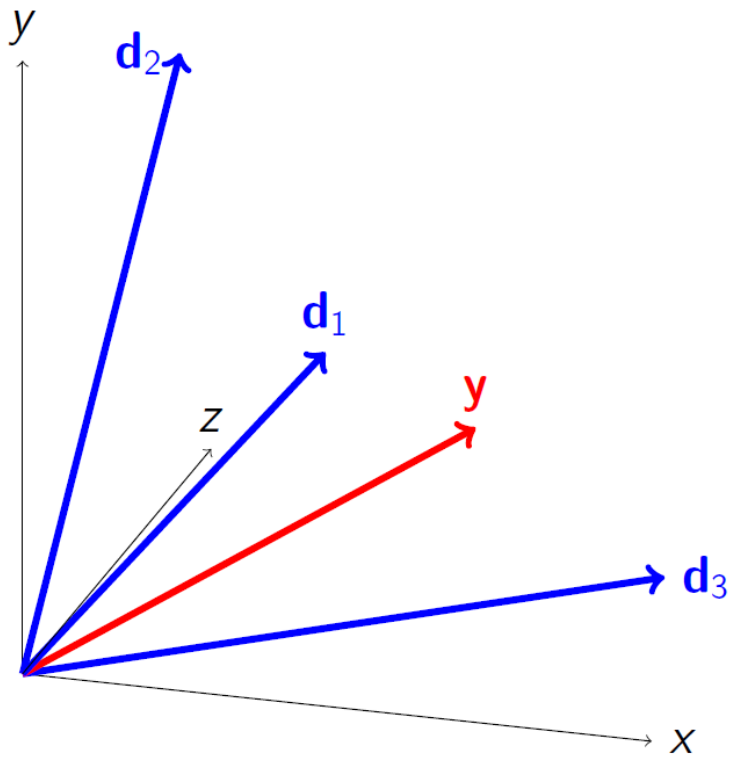


Figure: $\boldsymbol{\alpha} = (0, 0, 0)$; $\Gamma = \emptyset$

http://di.ens.fr/w...2010_Julien.pdf

Orthogonal Matching Pursuit

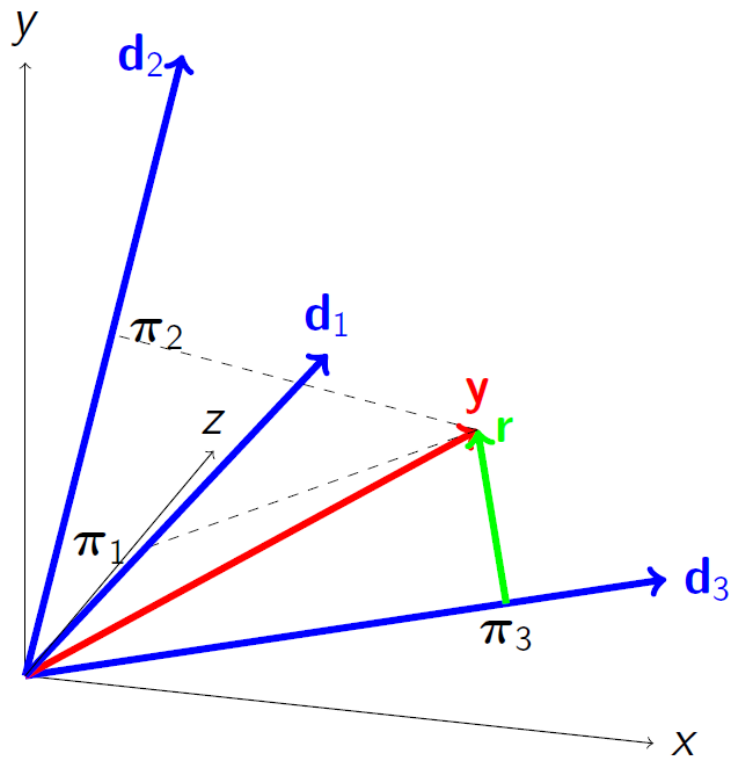


Figure: $\alpha = (0, 0, 0.75)$; $\Gamma = \{3\}$

http://di.ens.fr/w...2010_Julien.pdf

Orthogonal Matching Pursuit

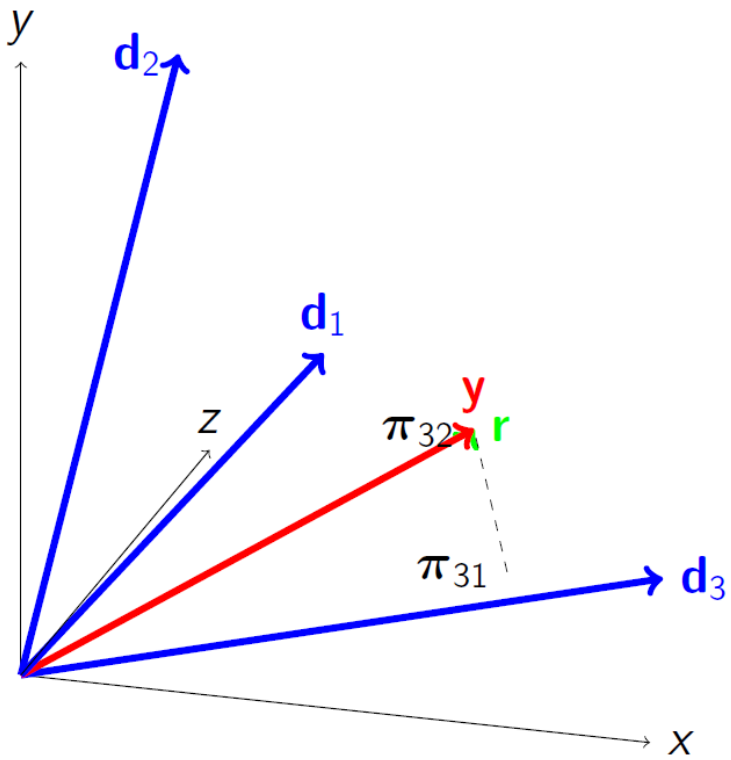


Figure: $\alpha = (0, 0.29, 0.63)$; $\Gamma = \{3, 2\}$

http://di.ens.fr/w...2010_Julien.pdf

Orthogonal Matching Pursuit

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

1: $\Gamma = \emptyset$.
2: **for** *iter* $= 1, \ldots, L$ **do**
3: Select the atom which most reduces the objective

$$\hat{\imath} \leftarrow \arg\min_{i \in \Gamma^c} \left\{ \min_{\alpha'} \|\mathbf{y} - \mathbf{D}_{\Gamma \cup \{i\}} \alpha'\|_2^2 \right\}$$

4: Update the active set: $\Gamma \leftarrow \Gamma \cup \{\hat{\imath}\}$.
5: Update the residual (orthogonal projection)

$$\mathbf{r} \leftarrow (\mathbf{I} - \mathbf{D}_\Gamma (\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1} \mathbf{D}_\Gamma^T)\mathbf{y}.$$

6: Update the coefficients

$$\alpha_\Gamma \leftarrow (\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1} \mathbf{D}_\Gamma^T \mathbf{y}.$$

7: **end for**

http://di.ens.fr/w...2010_Julien.pdf

Orthogonal Matching Pursuit

Contrary to MP, an atom can only be selected one time with OMP. It is, however, more difficult to implement efficiently. The keys for a good implementation in the case of a large number of signals are

- Precompute the Gram matrix $\mathbf{G} = \mathbf{D}^T\mathbf{D}$ once in for all,
- Maintain the computation of $\mathbf{D}^T\mathbf{r}$ for each signal,
- Maintain a Cholesky decomposition of $(\mathbf{D}_\Gamma^T\mathbf{D}_\Gamma)^{-1}$ for each signal.

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

$$\underbrace{O(p^2m)}_{\text{Gram matrix}} + \underbrace{O(nL^3)}_{\text{Cholesky}} + \underbrace{O(n(pm + pL^2))}_{\mathbf{D}^T\mathbf{r}} = O(np(m + L^2))$$

It is also possible to use the matrix inversion lemma instead of a Cholesky decomposition (same complexity, but less numerical stability)

http://di.ens.fr/w...2010_Julien.pdf

Optimality conditions of the Lasso

- **Directional derivative** in the direction $\mathbf{u}$ at $\boldsymbol{\alpha}$:

$$\nabla f(\boldsymbol{\alpha}, \mathbf{u}) = \lim_{t \to 0^+} \frac{f(\boldsymbol{\alpha} + t\mathbf{u}) - f(\boldsymbol{\alpha})}{t}$$

- Main idea: in non smooth situations, one may need to look at all directions $\mathbf{u}$ and not simply p independent ones!
- **Proposition 1:** if f is differentiable in $\boldsymbol{\alpha}$, $\nabla f(\boldsymbol{\alpha}, \mathbf{u}) = \nabla f(\boldsymbol{\alpha})^T \mathbf{u}$.
- **Proposition 2:** $\boldsymbol{\alpha}$ is optimal iff for all $\mathbf{u}$ in $\mathbb{R}^p$, $\nabla f(\boldsymbol{\alpha}, \mathbf{u}) \geq 0$.

http://di.ens.fr/w...2010_Julien.pdf

Optimality conditions of the Lasso

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

$\boldsymbol{\alpha}^\star$ is optimal iff for all $\mathbf{u}$ in $\mathbb{R}^p$, $\nabla f(\boldsymbol{\alpha},\mathbf{u})\geq 0$—that is,

$$-\mathbf{u}^T\mathbf{D}^T(\mathbf{y}-\mathbf{D}\boldsymbol{\alpha}^\star)+\lambda\sum_{i,\boldsymbol{\alpha}^\star[i]\neq 0}\operatorname{sign}(\boldsymbol{\alpha}^\star[i])\mathbf{u}[i]+\lambda\sum_{i,\boldsymbol{\alpha}^\star[i]=0}|\mathbf{u}_i|\geq 0,$$

which is equivalent to the following conditions:

$$\forall i=1,\ldots,p,\quad\begin{cases}|\mathbf{d}_i^T(\mathbf{y}-\mathbf{D}\boldsymbol{\alpha}^\star)| &\leq \lambda & \text{if } \boldsymbol{\alpha}^\star[i]=0\\ \mathbf{d}_i^T(\mathbf{y}-\mathbf{D}\boldsymbol{\alpha}^\star) &= \lambda\operatorname{sign}(\boldsymbol{\alpha}^\star[i]) & \text{if } \boldsymbol{\alpha}^\star[i]\neq 0\end{cases}$$

http://di.ens.fr/w...2010_Julien.pdf

Optimization for Dictionary Learning

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

$$\mathcal{C} \triangleq \{\mathbf{D} \in \mathbb{R}^{m\times p} \text{ s.t. } \forall j = 1,\ldots,p, \quad \|\mathbf{d}_j\|_2 \leq 1\}.$$

- Classical optimization alternates between $\mathbf{D}$ and $\boldsymbol{\alpha}$.
- Good results, but very slow!

http://di.ens.fr/w...2010_Julien.pdf

Optimization for Dictionary Learning

Classical formulation of dictionary learning

$$\min_{\mathbf{D}\in\mathcal{C}} f_n(\mathbf{D}) = \min_{\mathbf{D}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} l(\mathbf{y}_i, \mathbf{D}),$$

where

$$l(\mathbf{x}, \mathbf{D}) \triangleq \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1.$$

Which formulation are we interested in?

$$\min_{\mathbf{D}\in\mathcal{C}} \left\{ f(\mathbf{D}) = \mathbb{E}_{\mathbf{y}}[l(\mathbf{y}, \mathbf{D})] \approx \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} l(\mathbf{y}_i, \mathbf{D}) \right\}$$

[Bottou and Bousquet, 2008]: Online learning can

- handle potentially infinite or dynamic datasets,
- be dramatically faster than batch algorithms.

http://di.ens.fr/w...2010_Julien.pdf

Optimization for Dictionary Learning

Require: $\mathbf{D}_0 \in \mathbb{R}^{m \times p}$ (initial dictionary); $\lambda \in \mathbb{R}$

1: $\mathbf{A}_0 = 0$, $\mathbf{B}_0 = 0$.
2: **for** t=1,...,T **do**
3: Draw $\mathbf{y}_t$
4: Sparse Coding: $\alpha_t \leftarrow \arg\min_{\alpha \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{y}_t - \mathbf{D}_{t-1}\alpha\|_2^2 + \lambda\|\alpha\|_1$,
5: Aggregate sufficient statistics
$\mathbf{A}_t \leftarrow \mathbf{A}_{t-1} + \alpha_t \alpha_t^T$, $\mathbf{B}_t \leftarrow \mathbf{B}_{t-1} + \mathbf{y}_t \alpha_t^T$
6: Dictionary Update (block-coordinate descent)

$$\begin{aligned}\mathbf{D}_t &\leftarrow \arg\min_{\mathbf{D} \in \mathcal{C}} \frac{1}{t} \sum_{i=1}^{t} \left(\frac{1}{2}\|\mathbf{y}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1 \right). \\ &= \arg\min_{\mathbf{D} \in \mathcal{C}} \frac{1}{t} \left(\frac{1}{2} \mathrm{Tr}(\mathbf{D}^T \mathbf{D} \mathbf{A}_t) - \mathrm{Tr}(\mathbf{D}^T \mathbf{B}_t) \right).\end{aligned}$$

7: **end for**

http://di.ens.fr/w...2010_Julien.pdf

Optimization for Dictionary Learning

Which guarantees do we have?

Under a few reasonable assumptions,

- we build a surrogate function $\hat{f}_t$ of the expected cost f verifying

$$\lim_{t \to +\infty} \hat{f}_t(\mathbf{D}_t) - f(\mathbf{D}_t) = 0,$$

- $\mathbf{D}_t$ is asymptotically close to a stationary point.

Extensions (all implemented in SPAMS)

- non-negative matrix decompositions.
- sparse PCA (sparse dictionaries).
- fused-lasso regularizations (piecewise constant dictionaries)

---

http://di.ens.fr/w...2010_Julien.pdf

Thanks for the attention!
The end.