Matrix Sparsity - Structured Sparsity

Francis Bach - Guillaume Obozinski

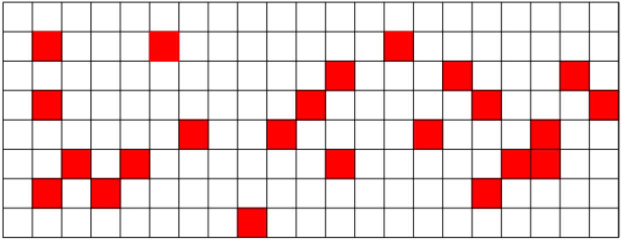
Willow group - INRIA - ENS - Paris

ECML 2010, Barcelona, September 20th

Outline

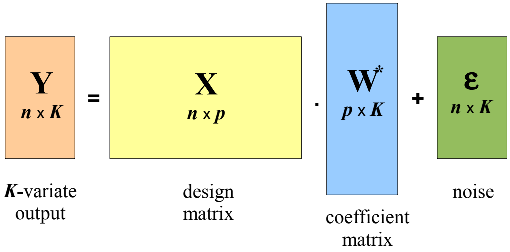
Learning on matrices - Collaborative Filtering (CF)

- Given $n_{\mathcal{X}}$ "movies" $\mathbf{x} \in \mathcal{X}$ and $n_{\mathcal{Y}}$ "customers" $\mathbf{y} \in \mathcal{Y}$,
- predict the "rating" $z(\mathbf{x}, \mathbf{y}) \in \mathcal{Z}$ of customer $\mathbf{y}$ for movie $\mathbf{x}$
- Training data: large $n_{\mathcal{X}} \times n_{\mathcal{Y}}$ incomplete matrix Z that describes the known ratings of some customers for some movies
- Goal: complete the matrix.



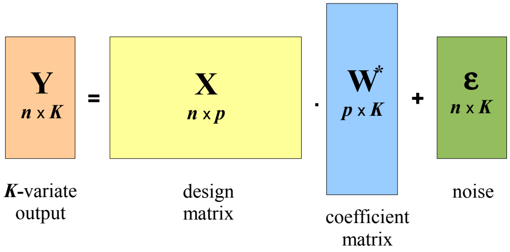
Learning on matrices - Multivariate problems

- Multivariate linear regression



Learning on matrices - Multivariate problems

- Multivariate linear regression



- Multiclass classification

$$\min_W \sum_{i=1}^{n} \frac{1}{n} \ell(w_1^\top x^{(i)}, \ldots, w_K^\top x^{(i)}, y^{(i)})$$

with

- $y^{(i)} \in \{0,1\}^K$
- One parameter vector $w_k \in \mathbb{R}^p$ per class
- ℓ is e.g. the multiclass logistic loss

Learning on matrices - Multi-task learning

- k prediction tasks on same covariates $x \in \mathbb{R}^p$
  - Each model parameterized by: $w^k \in \mathbb{R}^p, \quad 1 \leq k \leq K$

Learning on matrices - Multi-task learning

- k prediction tasks on same covariates $x \in \mathbb{R}^p$
  - Each model parameterized by: $w^k \in \mathbb{R}^p, \quad 1 \leq k \leq K$
  - Empirical risks: $L_k(w^k) = \frac{1}{n} \sum_{i=1}^{n} \ell_k({w^k}^\top x_i^k, y_i^k)$

Learning on matrices - Multi-task learning

- k prediction tasks on same covariates $x \in \mathbb{R}^p$
  - Each model parameterized by: $w^k \in \mathbb{R}^p, \quad 1 \leq k \leq K$
  - Empirical risks: $L_k(w^k) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \ell_k({w^k}^\top x_i^k, y_i^k)$
  - All parameters form a matrix:

$$W = [w^1, \ldots, w^K] = \begin{bmatrix} w_1^1 & \ldots & w_1^K \\ \vdots & w_j^k & \vdots \\ w_p^1 & \ldots & w_p^K \end{bmatrix} = \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix} \in \mathbb{R}^{p \times K}$$

Learning on matrices - Multi-task learning

- k prediction tasks on same covariates $x \in \mathbb{R}^p$
  - Each model parameterized by: $w^k \in \mathbb{R}^p, \quad 1 \leq k \leq K$
  - Empirical risks: $L_k(w^k) = \frac{1}{n} \sum_{i=1}^{n} \ell_k({w^k}^\top x_i^k, y_i^k)$
  - All parameters form a matrix:

$$W = [w^1, \ldots, w^K] = \begin{bmatrix} w_1^1 & \ldots & w_1^K \\ \vdots & w_j^k & \vdots \\ w_p^1 & \ldots & w_p^K \end{bmatrix} = \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix} \in \mathbb{R}^{p \times K}$$

- Many applications
  - Multi-category classification (one task per class) (Amit et al., 2007)
- Share parameters between various tasks
  - similar to fixed effect/random effect models (Raudenbush and Bryk, 2002)

Learning on matrices - Image denoising

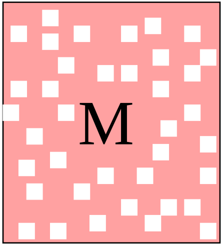
- Simultaneously denoise all patches of a given image
- Example from Mairal et al. (2009b)

Outline

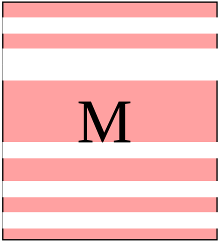
Two types of sparsity for matrices $M \in \mathbb{R}^{n \times p}$

I - Directly on the elements of M

Many zero elements: $M_{ij} = 0$



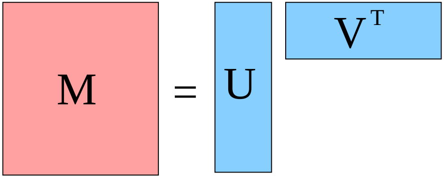
Many zero rows (or columns):
$(M_{i1}, \ldots, M_{ip}) = 0$



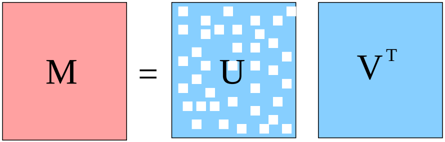
Two types of sparsity for matrices $M \in \mathbb{R}^{n \times p}$

II - Through a factorization of $M = UV^\top$

- $M = UV^\top$, $U \in \mathbb{R}^{n \times m}$ and $V \in \mathbb{R}^{n \times m}$
- **Low rank**: m small



- **Sparse decomposition**: U sparse



- Same as dictionary learning with notations $M = X, V = D$ and $A = U^\top$.

Outline

Joint variable selection (Obozinski et al., 2009)

$$\sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_k({w^k}^\top x_i^k, y_i^k) + \lambda \Omega(W)$$

- Joint matrix of predictors $W = (w_1, \ldots, w_k) \in \mathbb{R}^{p \times k}$:

$$W = [w^1, \ldots, w^K] = \begin{bmatrix} w_1^1 & \ldots & w_1^K \\ \vdots & w_j^k & \vdots \\ w_p^1 & \ldots & w_p^K \end{bmatrix} = \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix} \in \mathbb{R}^{p \times K} \longrightarrow$$

- Select all variables that are relevant to at least one task

$$\sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_k({w^k}^\top x_i^k, y_i^k) + \lambda \sum_{j=1}^{p} \|w_j\|_2$$

- Can improve performance over ℓ_1-regularization (Obozinski et al., 2008; Lounici et al., 2009)

Applications for simultaneous selection

Multi-class image classification (Quattoni et al., 2008)
$\longrightarrow$ algorithms for the regularization by a sum of ℓ_∞-norm (ℓ_1/ℓ_∞).
$\longrightarrow$ increase in performance

Multi-class tumor classification based on gene expression data (Obozinski et al., 2009)
$\longrightarrow$ smaller gene signatures

Source localization in M/EEG inverse problems from several experiments (Gramfort, 2010)

Outline

Rank constraints and sparsity of the spectrum

Rank

Given a matrix $M \in \mathbb{R}^{n \times p}$

- Singular value decomposition (SVD): $M = U \, \mathrm{Diag}(s) \, V^\top$
  where U, V orthogonal, $s \in \mathbb{R}_+^m$ are singular values
- $Rank(M) = \|s\|_0$
- Rank of M is the minimum size m of **all** factorizations of M into $M = UV^\top$, $U \in \mathbb{R}^{n \times m}$ and $V \in \mathbb{R}^{p \times m}$

Rank constraints and sparsity of the spectrum

Rank

Given a matrix $M \in \mathbb{R}^{n \times p}$

- Singular value decomposition (SVD): $M = U \,\mathrm{Diag}(s)\, V^\top$ where U, V orthogonal, $s \in \mathbb{R}_+^m$ are singular values
- $Rank(M) = \|s\|_0$
- Rank of M is the minimum size m of **all** factorizations of M into $M = UV^\top$, $U \in \mathbb{R}^{n \times m}$ and $V \in \mathbb{R}^{p \times m}$

Rank constrained Learning

$$\min_{W \in \mathbb{R}^{p \times k}} L(W) \quad \text{s.t.} \quad \mathrm{rank}(W) \leq m$$

Examples:

- Collaborative filtering
- Multi-task learning with task parameters assumed in a low dimensional subspace (Argyriou et al., 2009)

Low-rank via factorization

Reduced-rank multivariate regression

$$\min_{W} \|Y - XW\|_F^2 \quad \text{s.t. } \operatorname{rank}(W) \le k$$

- Well studied (Anderson, 1951; Izenman, 1975; Reinsel and Velu, 1998)
- Is solved directly using the SVD (by OLS + SVD + projection)

General factorization

$$\min_{U \in \mathbb{R}^{p \times m}, V \in \mathbb{R}^{k \times m}} L(UV^\top)$$

- Still non-convex but convex w.r.t. U and V separately
- Optimization by alternating procedures

Trace norm relaxation

With SVD $W = U\mathrm{Diag}(s)V^\top$, $\quad \mathrm{rank}(W) = \|s\|_0 \quad \xrightarrow{\text{Relax}} \|s\|_1$.

- $M \mapsto \|s\|_1$ is actually a *unitary invariant* norm: the trace norm, nuclear norm or unitary norm
- Write it $M \mapsto \|M\|_{\mathrm{tr}}$
- Dual norm to the spectral norm $\|M\|_2 = \|s\|_\infty$

Trace norm regularization

$$\min_{W \in \mathbb{R}^{p \times k}} L(W) + \lambda \|W\|_{\mathrm{tr}}$$

- Convex problem
- Algorithms:
  - Proximal methods
  - Iterated Reweighted Least-Square (Argyriou et al., 2009)
  - Common bottleneck: require iterative SVD

Trace norm and collaborative filtering

$$\min_{M \in \mathbb{R}^{p \times n}} \sum_{(i,j) \in S} \|M_{ij} - M^0_{ij}\|_2^2 + \lambda \|M\|_{tr}$$

- semi-definite program (Fazel et al., 2001)
- see also max-margin approaches to CF (Srebro et al., 2005)
- Statistical results:
  - High-dimensional inference for noisy matrix completion (Srebro et al., 2005; Candès and Plan, 2009)
  - May recover entire matrix from slightly more entries than the minimum of the two dimensions

Outline

Two different views of PCA

Given data matrix $X = (x_1^\top, \ldots, x_n^\top)^\top \in \mathbb{R}^{n \times p}$,

Two different views of PCA

Given data matrix $X = (x_1^\top, \ldots, x_n^\top)^\top \in \mathbb{R}^{n \times p}$,

Analysis view

Find projection $v \in \mathbb{R}^p$ maximizing variance:

$$\begin{array}{cc} \max_{v \in \mathbb{R}^p} & v^\top X^\top X\, v \\ \text{s.t.} & \|v\|_2 \leq 1 \end{array}$$

$\rightarrow$ deflate and iterate to obtain more components.

Two different views of PCA

Given data matrix $X = (x_1^\top, \ldots, x_n^\top)^\top \in \mathbb{R}^{n \times p}$,

Analysis view

Find projection $v \in \mathbb{R}^p$ maximizing variance:

$$\begin{array}{ll} \max_{v \in \mathbb{R}^p} & v^\top X^\top X\, v \\ \text{s.t.} & \|v\|_2 \leq 1 \end{array}$$

$\rightarrow$ deflate and iterate to obtain more components.

Synthesis view

Find $V = [v_1, \ldots, v_k]$ s.t. x_i have low reconstruction error on span(V):

$$\min_{u_i, v_i \in \mathbb{R}^p} \|X - \sum_{i=1}^{k} u_i v_i^\top\|_F^2$$

Two different views of PCA

Given data matrix $X = (x_1^\top, \ldots, x_n^\top)^\top \in \mathbb{R}^{n \times p}$,

Analysis view

Find projection $v \in \mathbb{R}^p$ maximizing variance:

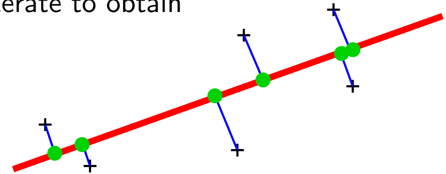
$$\max_{v \in \mathbb{R}^p} \quad v^\top X^\top X \, v \qquad \text{s.t.} \quad \|v\|_2 \leq 1$$

$\longrightarrow$ deflate and iterate to obtain more components.

Synthesis view

Find $V = [v_1, \ldots, v_k]$ s.t. x_i have low reconstruction error on span(V):

$$\min_{u_i, v_i \in \mathbb{R}^p} \|X - \sum_{i=1}^{k} u_i v_i^\top\|_F^2$$



Two different views of PCA

Given data matrix $X = (x_1^\top, \dots, x_n^\top)^\top \in \mathbb{R}^{n \times p}$,

Analysis view

Find projection $v \in \mathbb{R}^p$ maximizing variance:

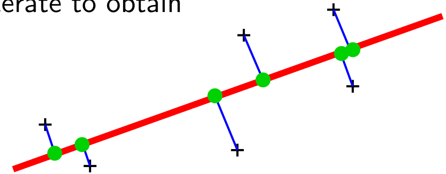
$$\max_{v \in \mathbb{R}^p} \quad v^\top X^\top X\, v \qquad \text{s.t.} \quad \|v\|_2 \le 1$$

$\rightarrow$ deflate and iterate to obtain more components.

Synthesis view

Find $V = [v_1, \dots, v_k]$ s.t. x_i have low reconstruction error on span(V):

$$\min_{u_i, v_i \in \mathbb{R}^p} \|X - \sum_{i=1}^{k} u_i v_i^\top\|_F^2$$



- For regular PCA, the two views are equivalent!
- Not true if constraints on u, v change

Sparse PCA - Analysis view

Add sparsity constraint:

$$\max_{\|v\|_2=1,\ \|v\|_0 \leqslant k} v^\top X^\top X\, v$$

Sparse PCA - Analysis view

Add sparsity constraint:

$$\max_{\|v\|_2=1,\ \|v\|_0 \leqslant k} v^\top X^\top X\, v$$

Convex relaxation **DSPCA** (d'Aspremont et al., 2007)

relaxed into $$\max_{\|v\|_2=1,\ \|v\|_1 \leqslant k^{1/2}} v^\top X^\top X\, v$$

then relaxed into $$\max_{M \succcurlyeq 0,\ \mathrm{tr}(M)=1,\ 1^\top |M| 1 \leqslant k} \mathrm{tr}(X^\top X M), \quad \text{using } M = vv^\top.$$

Sparse PCA - Analysis view

Add sparsity constraint:

$$\max_{\|v\|_2=1,\ \|v\|_0 \leqslant k} v^\top X^\top X\, v$$

Convex relaxation **DSPCA** (d'Aspremont et al., 2007)

relaxed into $$\max_{\|v\|_2=1,\ \|v\|_1 \leqslant k^{1/2}} v^\top X^\top X\, v$$

then relaxed into $$\max_{M \succcurlyeq 0,\ \mathrm{tr}(M)=1,\ 1^\top |M| 1 \leqslant k} \mathrm{tr}(X^\top X M), \quad \text{using } M = vv^\top.$$

- Requires deflation for multiple components (Mackey, 2009)
- More refined convex relaxation (d'Aspremont et al., 2008)
- Analysis of non-convex formulation (Moghaddam et al., 2006)

Sparse PCA - Synthesis view

Find $V = [v_1, \dots, v_m] \in \mathbb{R}^{p \times n}$ sparse and $U = [u_1, \dots, u_m] \in \mathbb{R}^{n \times n}$ s.t.

$$\sum_{i=1}^{n} \left\| x_i - \sum_{j=1}^{m} u_{ij} v_j \right\|_2^2 \text{ is small } \Leftrightarrow \quad \|X - UV^\top\|_F^2, \text{ is small}$$

Sparse matrix factorization (Witten et al., 2009; Bach et al., 2008)

- Penalize columns v_i of V by the ℓ_1-norm for sparsity
- Penalize columns u_i of U by the ℓ_2-norm to avoid trivial solutions

$$\min_{U,V} \|X - UV^\top\|_F^2 \quad + \quad \frac{\lambda}{2} \sum_{i=1}^{m} \left\{ \|u_i\|_2^2 + \|v_i\|_1^2 \right\}$$

$$\min_{U,V} \|X - UV^\top\|_F^2 \quad + \quad \lambda \sum_{i} \|u_i\|_2 \|v_i\|_1$$

$$\min_{U,V} \|X - UV^\top\|_F^2 \quad + \quad \lambda \sum_{i} \|v_i\|_1 \quad \text{s.t. } \|u_i\|_2 \leq 1$$

yield the same solutions for $u_j v_j^\top$ (Bach et al., 2008).

Efficient algorithms for sparse matrix factorization

Focus on previous formulation:

$$\min_{U,V} \|X - UV^\top\|_F^2 + \lambda \sum_i \|v_j\|_1 \quad \text{s.t. } \|u_j\|_2 \leq 1$$

- Problem is convex in U and V separately, but not jointly.
  $\rightarrow$ Alternating scheme: optimize U and V in turn.
- Even better: use simple column updates (Lee et al., 2007; Witten et al., 2009):

$$\text{With} \quad \widetilde{X} = X - \sum_{j' \neq j} u_j v_j^\top, \quad \text{we have}$$

$$\text{either} \quad u_j \leftarrow \frac{\widetilde{X} v_j}{\|\widetilde{X} v_j\|} \quad \text{or} \quad v_j \leftarrow \text{argmin}_v \|X^\top u_j - v\|_2^2 + \lambda \|v\|_1$$

  - requires no matrix inversion
  - $+$ can take advantage of efficient algorithms for Lasso
  - can use warm start $+$ active sets

Sparse PCA - Synthesis view II

"Sparse projector" (Zou et al., 2006)

Find $\widetilde{V} = [\tilde{v}_1, \ldots, \tilde{v}_m] \in \mathbb{R}^{p \times n}$ and $V = [v_1, \ldots, v_m] \in \mathbb{R}^{p \times n}$ such that

$$\min_{\widetilde{V}, V} \quad \sum_{i=1}^{n} \|x_i - \widetilde{V} V^\top x_i\|_2^2 + \lambda_1 \|V\|_1 + \lambda_2 \|V\|_F^2$$

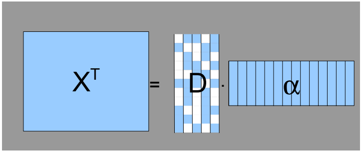
$$\text{such that} \quad \widetilde{V}^\top \widetilde{V} = I_p$$

- The data should be reconstructed from sparse projections
- Non-convex formulation $\rightarrow$ alternating minimization

Sparse PCA *vs* Dictionary Learning a.k.a. Sparse Coding

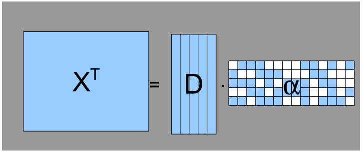
In signal processing $X^\top = \underbrace{V}_{\text{dictionary } D} \underbrace{U^\top}_{\text{decomposition coefficients } \alpha} = D\alpha$

Sparse PCA



- e.g. microarray data
- sparse dictionary
- (Witten et al., 2009; Bach et al., 2008)

Dictionary Learning



- e.g. overcomplete dictionaries for natural images
- sparse decomposition
- (Elad and Aharon, 2006)

Outline

Dictionary Learning

$$\min_{\substack{\mathbf{A}\in\mathbb{R}^{k\times n}\\ \mathbf{D}\in\mathbb{R}^{p\times k}}} \sum_{i=1}^{n} \left(\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1\right) \quad \text{s.t.} \quad \forall j,\ \|\mathbf{d}_j\|_2 \leq 1.$$

- As before not jointly convex but convex in each $\mathbf{d}_j$ and $\boldsymbol{\alpha}_j$
- Alternating scheme becomes slow for large signal databases ...

[$\rightarrow$] use *Stochastic Optimization / Online learning* (Mairal et al., 2009a)

- can handle potentially infinite datasets
- can adapt to dynamic training sets

Inpainting a 12-Mpixel photograph

Inpainting a 12-Mpixel photograph

Inpainting a 12-Mpixel photograph

Inpainting a 12-Mpixel photograph

Outline

Sparsity with Structure

Notion emerged very recently through the work of several authors:
Yuan and Lin (2006), Zhao et al. (2009), Baraniuk et al. (2008), Bach (2008), Jacob et al. (2009), Jenatton et al. (2009), Jenatton et al. (2010b), He and Carin (2009), Huang et al. (2009).

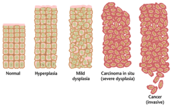
The support is sparse but we have prior information about its structure.

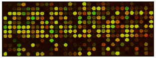
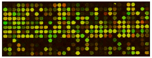
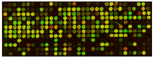
- The variables should be selected in groups.
- The variables lie in a hierarchy.
- The variables lie on a graph or network and the support should be localized or densely connected on the graph.
- The variables are pixels of an image and form rectangles or convex shapes.

Outline

Biological markers for cancer

Metastasis prognosis: Predict if a tumor will produce metastases.



Gene expression in tumor	Metastasis?
	✓
⋮	⋮
	✗
	?

- Can we predict metastasis and identify few predictive genes?

Biological pathways as relevant groups of genes

Predictive genes are naturally grouped in *biological pathways*

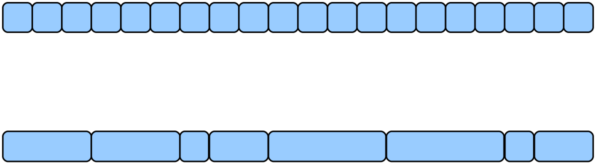
- Correspond to genes participating in same biological mechanisms
- Contain often very correlated genes
- The pathways form overlapping groups
- Ultimately relevant to the biologist

$\Rightarrow$ Instead of selecting genes individually, select entire pathways.

The support is a **union of overlapping groups**.

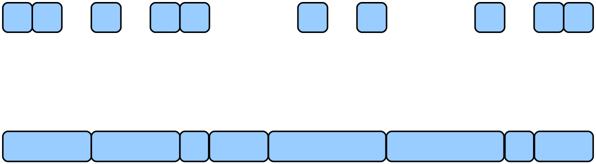
Sparsity patterns induced for $L(w) + \lambda\,\Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$



Sparsity patterns induced for $L(w) + \lambda\,\Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$



Sparsity patterns induced for $L(w) + \lambda\,\Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso (Yuan and Lin, 2006): $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

Sparsity patterns induced for $L(w) + \lambda\,\Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

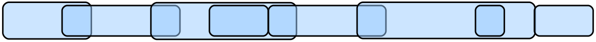
Group Lasso (Yuan and Lin, 2006): $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

Sparsity patterns induced for $L(w) + \lambda \Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso (Yuan and Lin, 2006): $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

Group Lasso when groups overlap: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

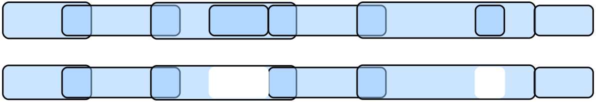


Sparsity patterns induced for $L(w) + \lambda \Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso (Yuan and Lin, 2006): $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

Group Lasso when groups overlap: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

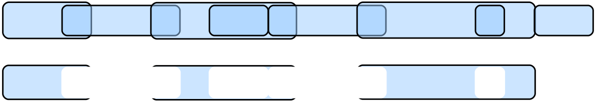


Sparsity patterns induced for $L(w) + \lambda\,\Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso (Yuan and Lin, 2006): $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

Group Lasso when groups overlap: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

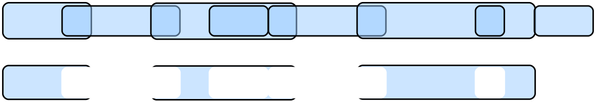


Sparsity patterns induced for $L(w) + \lambda \Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso (Yuan and Lin, 2006): $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

Group Lasso when groups overlap: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$



The support obtained is

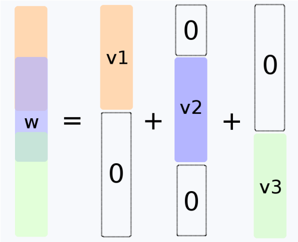
- An intersection of the complements of the groups set to 0 (cf. Jenatton et al. (2009))
- Not a union of groups

New formulation

(Jacob, Obozinski and Vert (2009))

Introducing latent variables v_g:

$$\begin{cases} \min_{w,v} L(w) + \lambda \sum_{g \in \mathcal{G}} \|v_g\|_2 \\ w = \sum_{g \in \mathcal{G}} v_g \\ \operatorname{supp}(v_g) \subseteq g. \end{cases}$$



Properties

- Resulting support is a *union* of groups in $\mathcal{G}$.
- Possible to select one variable without selecting all the groups containing it.

A new "overlap" norm

Equivalent reformulation

$$\begin{cases} \min_{w,v} L(w) + \lambda \sum_{g \in \mathcal{G}} \|v_g\|_2 \\ w = \sum_{g \in \mathcal{G}} v_g \\ \text{supp}\,(v_g) \subseteq g. \end{cases} = \min_{w} L(w) + \lambda \Omega_{overlap}(w)$$

with

$$\Omega_{overlap}(w) \triangleq \begin{cases} \min_{v} \sum_{g \in \mathcal{G}} \|v_g\|_2 \\ w = \sum_{g \in \mathcal{G}} v_g \\ \text{supp}\,(v_g) \subseteq g. \end{cases} \qquad (*)$$

A new "overlap" norm

Equivalent reformulation

$$\left\{\begin{array}{l}\min\limits_{w,v} L(w) + \lambda \sum\limits_{g \in \mathcal{G}} \|v_g\|_2 \\ w = \sum_{g \in \mathcal{G}} v_g \\ \operatorname{supp}(v_g) \subseteq g.\end{array}\right. = \min_{w} L(w) + \lambda \Omega_{overlap}(w)$$

with

$$\Omega_{overlap}(w) \triangleq \left\{\begin{array}{l}\min\limits_{v} \sum\limits_{g \in \mathcal{G}} \|v_g\|_2 \\ w = \sum_{g \in \mathcal{G}} v_g \\ \operatorname{supp}(v_g) \subseteq g.\end{array}\right. \qquad (*)$$

- $\Omega_{overlap}(w)$ is a norm of w.

A new "overlap" norm

Equivalent reformulation

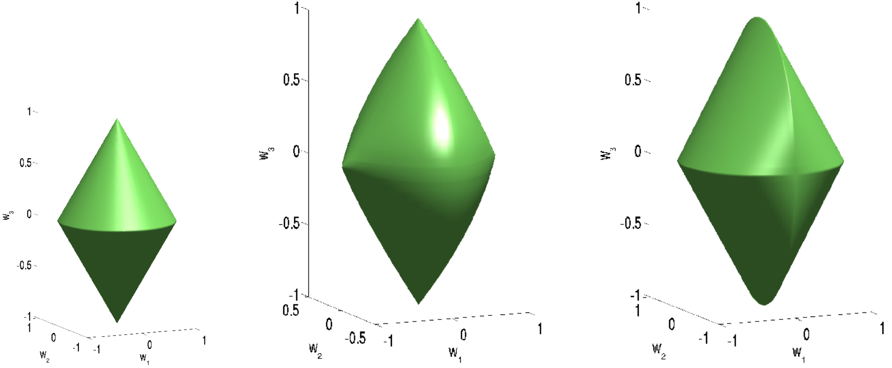
$$\begin{cases} \min\limits_{w,v} L(w) + \lambda \sum\limits_{g\in\mathcal{G}} \|v_g\|_2 \\ w = \sum_{g\in\mathcal{G}} v_g \\ \operatorname{supp}(v_g) \subseteq g. \end{cases} = \min_w L(w) + \lambda\Omega_{overlap}(w)$$

with

$$\Omega_{overlap}(w) \triangleq \begin{cases} \min\limits_{v} \sum\limits_{g\in\mathcal{G}} \|v_g\|_2 \\ w = \sum_{g\in\mathcal{G}} v_g \\ \operatorname{supp}(v_g) \subseteq g. \end{cases} \qquad (*)$$

- $\Omega_{overlap}(w)$ is a norm of w.

Overlap and group unity balls



Balls for $\Omega^{\mathcal{G}}_{\text{group}}(\cdot)$ (middle) and $\Omega^{\mathcal{G}}_{\text{overlap}}(\cdot)$ (right) for the groups $\mathcal{G} = \{\{1,2\},\{2,3\}\}$ where w_2 is represented as the vertical coordinate. Left: group-lasso ($\mathcal{G} = \{\{1,2\},\{3\}\}$), for comparison.

Results

Breast cancer data

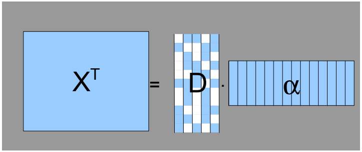
- Gene expression data for $8,141$ genes in 295 breast cancer tumors.
- Canonical pathways from MSigDB containing 639 groups of genes, 637 of which involve genes from our study.

Method	ℓ_1	$\Omega^{\mathcal{G}}_{\text{overlap}}(.)$
Misclassification error	0.38 ± 0.04	0.36 ± 0.03
Number of pathways involved	$148, 58, 183$	$6, 5, 78$

Outline

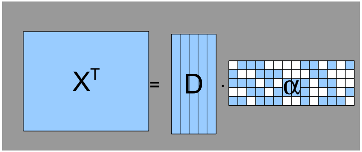
Sparse PCA / Dictionary Learning

Sparse PCA



- e.g. microarray data
- sparse dictionary
- (Witten et al., 2009; Bach et al., 2008)

Dictionary Learning



- e.g. overcomplete dictionaries for natural images
- sparse decomposition
- (Elad and Aharon, 2006)

Other constraints

Structured matrix factorizations - Many instances

- $M = UV^\top$, $U \in \mathbb{R}^{n\times m}$ and $V \in \mathbb{R}^{p\times m}$
- **Structure on U and/or V**
  - Low-rank: U and V have few columns
  - Dictionary learning / sparse PCA: U or V has many zeros
  - Clustering (k-means): $U \in \{0,1\}^{n\times m}$, $U1 = 1$
  - Pointwise positivity: non negative matrix factorization (NMF)
  - Specific patterns of zeros
  - etc.
- **Many applications**
  - e.g., source separation (Févotte et al., 2009), exploratory data analysis

From SPCA to SSPCA

Sparse PCA:

$$\min_{\substack{\mathbf{A}\in\mathbb{R}^{k\times n}\\ \mathbf{D}\in\mathbb{R}^{p\times k}}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^{k} \|\mathbf{d}_j\|_1 \quad \text{s.t.} \quad \forall j,\ \|\boldsymbol{\alpha}_j\|_2 \leq 1.$$

Sparse structured PCA

$$\min_{\substack{\mathbf{A}\in\mathbb{R}^{k\times n}\\ \mathbf{D}\in\mathbb{R}^{p\times k}}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^{k} \Omega(\mathbf{d}_j) \quad \text{s.t.} \quad \forall j,\ \|\boldsymbol{\alpha}_j\|_2 \leq 1.$$

- No orthogonality
- Not jointly convex but convex in each $\mathbf{d}_j$ and $\boldsymbol{\alpha}_j$
- $\Rightarrow$ efficient block-coordinate descent algorithms

Faces

Faces

- A basis to decompose faces?
- Eigenfaces
- Find parts?
- Localized components
- NMF (Lee and Seung, 1999)

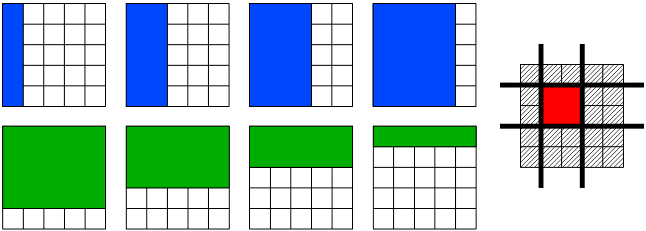
Faces

Faces

NMF

Rectangular supports

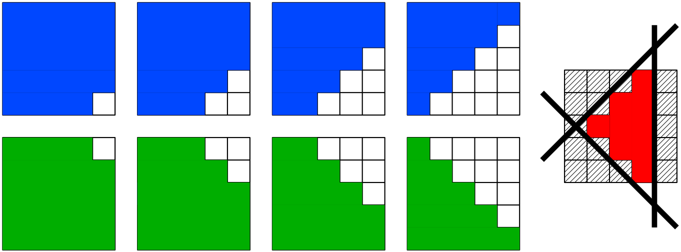
- $\Omega(\mathbf{d}) = \sum_{g \in \mathcal{G}} \|\mathbf{d}_g\|_2$: Selection of rectangles on the 2D-grid.



- $\mathcal{G}$ 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

General "convex" supports

- $\Omega(\mathbf{d}) = \sum_{g \in \mathcal{G}} \|\mathbf{d}_g\|_2$: Selection of "convex" patterns on a 2-D grids.



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

Sparse Structured PCA (Jenatton, Obozinski and Bach (2009))

- Learning **sparse and structured dictionary elements**:

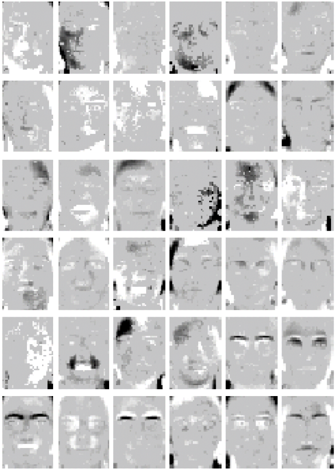
$$\min_{\substack{\mathbf{A}\in\mathbb{R}^{k\times n}\\ \mathbf{D}\in\mathbb{R}^{p\times k}}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^{p} \Omega(\mathbf{d}_j) \text{ s.t. } \forall i,\ \|\boldsymbol{\alpha}_i\|_2 \leq 1$$

- Structure of the dictionary elements determined by the choice of $\mathcal{G}$ (and thus Ω)
- Efficient learning procedures through *variational formulation*.
  - Reweighted ℓ_2: $\displaystyle\sum_{g\in\mathcal{G}} \|\mathbf{y}_g\|_2 = \min_{\eta_g\geqslant 0, g\in\mathcal{G}} \frac{1}{2} \sum_{g\in\mathcal{G}} \left\{ \frac{\|\mathbf{y}_g\|_2^2}{\eta_g} + \eta_g \right\}$

Faces

- AR Face database
- 100 individuals (50 W/50 M)
- For each
  - 14 non-occluded
  - 12 occluded
  - lateral illuminations
  - reduced resolution to 38×27 pixels

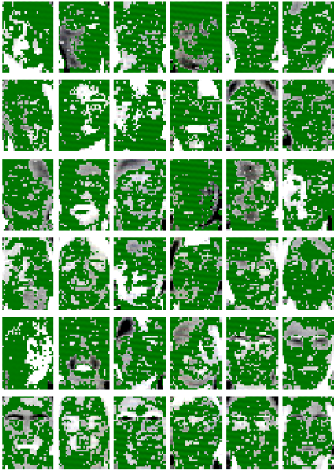
Decomposition of faces



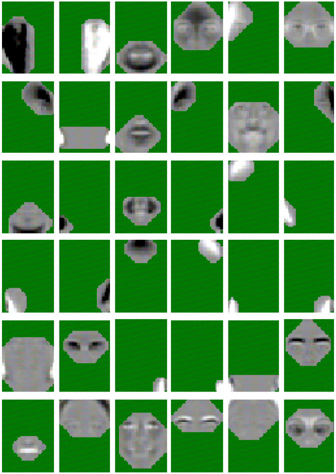
SPCA

SSPCA

Decomposition of faces II

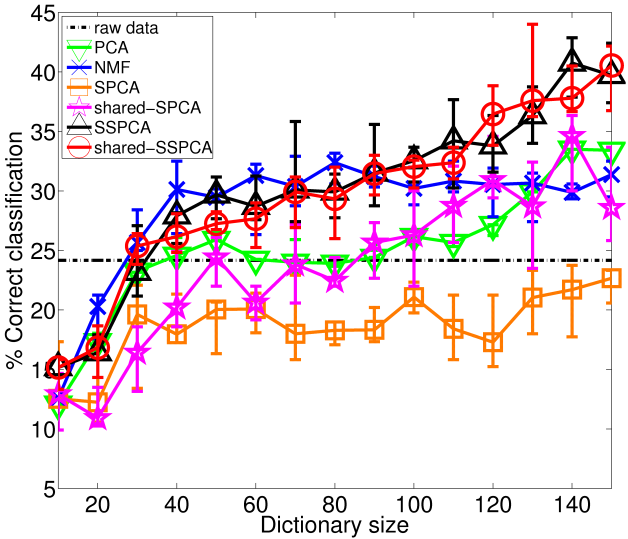


SPCA



SSPCA

k-NN classification based on decompositions



Outline

Hierarchical Topic Models for text corpora

Flat Topic Model

Each document x_j is modeled through word counts:
x_{ij} = nb of occurrences of word i in document j, $\qquad x_j^\top \mathbf{1} = n_j$,
θ=topic proportions, $\qquad D$=topic word frequencies

$$\text{Model} \quad x_j \quad \text{as.} \quad x_j \sim \mathcal{M}(D\theta, n_j)$$

- Low-rank matrix factorization of word-document matrix
- Multinomial PCA (Buntine and Perttu, 2003)
- Bayesian approach: Latent Dirichlet Allocation (Blei et al., 2003)

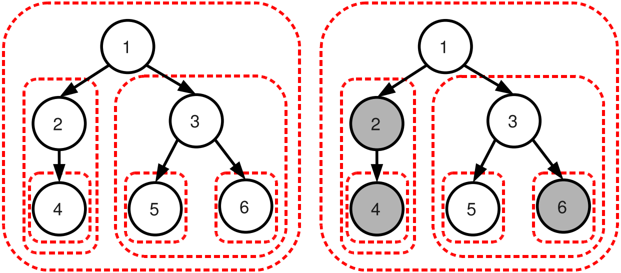
Hierarchical Model: Organise the topics in a tree ?

- Previous approaches: non-parametric Bayesian methods (Hierarchical Chinese Restaurant Process and nested Dirichlet Process): Blei et al. (2004)
- Can we obtain a similar model with **structured** matrix factorization?

Hierarchical Norm

(Jenatton, Mairal, Obozinski and Bach, 2010)

- Structure on codes $\boldsymbol{\alpha}$ (not on dictionary $\mathbf{D}$)
- Hierarchical penalization: $\Omega(\boldsymbol{\alpha}) = \sum_{g \in \mathcal{G}} \|\boldsymbol{\alpha}_g\|_2$ where groups g in $\mathcal{G}$ are equal to set of descendants of some nodes in a tree



- Variable selected after its ancestors (Zhao et al., 2009; Bach, 2008)

Hierarchical Dictionary Learning

Efficient Optimization

$$\min_{\substack{\mathbf{A}\in\mathbb{R}^{k\times n}\\ \mathbf{D}\in\mathbb{R}^{p\times k}}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda\Omega(\boldsymbol{\alpha}_i) \text{ s.t. } \forall j,\ \|\mathbf{d}_j\|_2 \leq 1.$$

- Proximal methods
- Requires solving $\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \boldsymbol{\alpha}\|_2^2 + \lambda\Omega(\boldsymbol{\alpha})$
- Can we do this for tree-structured norms?

Tree-structured groups

Proposition (Jenatton et al., 2010a)

- If $\mathcal{G}$ is a *tree-structured* set of groups, i.e.,

$$g \cap g' \neq \emptyset \Rightarrow g \subset g' \text{ or } g' \subset g,$$

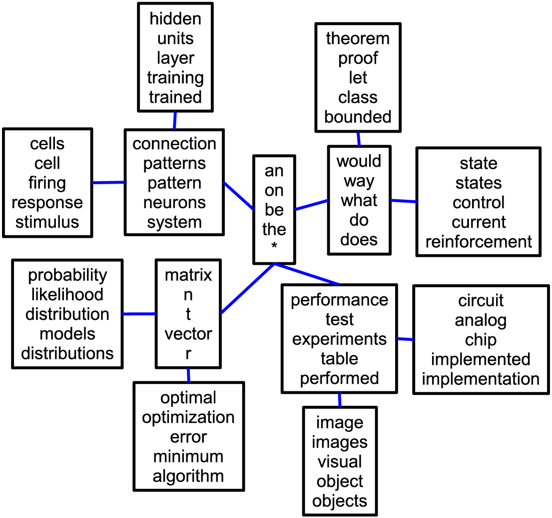
- If the groups are sorted from the leaves to the root,
- If Π_g is
  - the proximal operator $w_g \mapsto \mathrm{Prox}_{\mu\|\cdot\|_q}(w_g)$ on the subspace corresponding to group g and
  - the identity on the orthogonal

Then the proximal operator for Ω is the composition of all operators from the leaves to the root.

$$\mathrm{Prox}_{\mu\Omega} = \Pi_{g_m} \circ \ldots \circ \Pi_{g_1}. \quad (1)$$

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

Tree of Topics



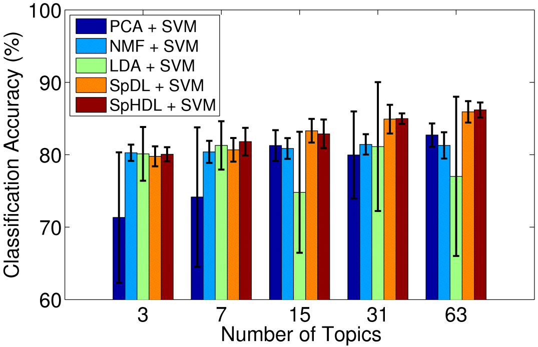
NIPS abstracts

- 1714 documents
- 8274 words

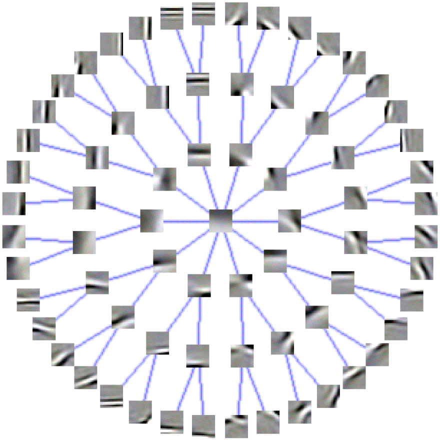
Classification based on topics

Comparison on predicting newsgroup article subjects

- 20 newsgroup articles (1425 documents, 13312 words)



Hierarchical dictionary for image patches



Summary

Sparse linear estimation with ℓ_1-regularization

- Convex optimization and algorithms
- Theoretical results

Group sparsity

- Block norm
- Multiple Kernel Learning

Matrix Sparsity

- Row sparsity for Multivariate Learning
- Low rank, SPCA and Dictionary Learning

Structured Sparsity

- Overlapping groups and supports stable by union or intersection
- SSPCA and Hierarchical Dictionary Learning

Conclusions

Sparse methods are not limited to regression

High-dimension

- Sparse methods performs well with very many predictors:
- Can algorithms tackle $\log(p) = o(n)$ for $n > 100$?

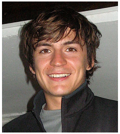
Performance

- Inducing sparsity does not always improve predictive performance
- Sparsity is a prior
- "Problems are sparse if you look at them the right way"

Capture structure

- Structured sparsity enhances interpretability
- Norm design: make the right norm for your problem

Acknowledgements



Rodolphe Jenatton

Julien Mairal

Laurent Jacob

Jean Ponce

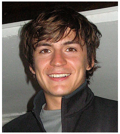
Jean-Philippe Vert

Ben Taskar

Martin Wainwright

Michael Jordan

Acknowledgements



Rodolphe Jenatton

Julien Mairal

Laurent Jacob

Jean Ponce

Jean-Philippe Vert

Ben Taskar

Martin Wainwright

Michael Jordan

Thank you!

References I

Amit, Y., Fink, M., Srebro, N., and Ullman, S. (2007). Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine Learning (ICML)*.

Anderson, T. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22(3):327–351.

Argyriou, A., Micchelli, C., and Pontil, M. (2009). On spectral learning. *Journal of Machine Learning Research*. To appear.

Bach, F. (2008). Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*.

Bach, F., Mairal, J., and Ponce, J. (2008). Convex sparse matrix factorizations. Technical Report 0812.1869, ArXiv.

Baraniuk, R. G., Cevher, V., Duarte, M. F., and Hegde, C. (2008). Model-based compressive sensing. Technical report, arXiv:0808.3572.

Blei, D., Griffiths, T., Jordan, M., and Tenenbaum, J. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, 16:106.

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

Buntine, W. and Perttu, S. (2003). Is multinomial PCA multi-faceted clustering or dimensionality reduction. In *International Workshop on Artificial Intelligence and Statistics (AISTATS)*.

References II

Candès, E. and Plan, Y. (2009). Matrix completion with noise. Submitted.

d'Aspremont, A., Bach, F., and El Ghaoui, L. (2008). Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9:1269–1294.

d'Aspremont, A., Ghaoui, E. L., Jordan, M. I., and Lanckriet, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–48.

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

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

Févotte, C., Bertin, N., and Durrieu, J.-L. (2009). Nonnegative matrix factorization with the itakura-saito divergence. with application to music analysis. *Neural Computation*, 21(3).

Gramfort, A. (2010). Multi-condition M/EEG inverse modeling with sparsity assumptions: how to estimate what is common and what is specific in multiple experimental conditions. In *17th International Conference on Biomagnetism Advances in Biomagnetism–Biomag2010*, pages 124–127. Springer.

He, L. and Carin, L. (2009). Exploiting structure in wavelet-based Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 57:3488–3497.

Huang, J., Zhang, T., and Metaxas, D. (2009). Learning with structured sparsity. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*.

References III

Izenman, A. (1975). Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlaps and graph lasso. In Bottou, L. and Littman, M., editors, *Proceedings of the 26th International Conference on Machine Learning*, pages 433–440, Montreal. Omnipress.

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

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. (2010a). Proximal methods for sparse hierarchical dictionary learning. In Fürnkranz, J. and Joachims, T., editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 487–494, Haifa, Israel. Omnipress.

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

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

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

Lounici, K., Tsybakov, A., Pontil, M., and van de Geer, S. (2009). Taking advantage of sparsity in multi-task learning. In *Conference on Computational Learning Theory (COLT)*.

Mackey, L. (2009). Deflation methods for sparse PCA. *Advances in Neural Information Processing Systems (NIPS)*, 21.

References IV

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2009a). Online dictionary learning for sparse coding. In *International Conference on Machine Learning (ICML)*.

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

Moghaddam, B., Weiss, Y., and Avidan, S. (2006). Spectral bounds for sparse PCA: Exact and greedy algorithms. In *Advances in Neural Information Processing Systems*, volume 18.

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

Obozinski, G., Wainwright, M., and Jordan, M. (2008). High-dimensional union support recovery in multivariate regression. In *Advances in Neural Information Processing Systems (NIPS)*.

Quattoni, A., Collins, M., and Darrell, T. (2008). Transfer learning for image classification with sparse prototype representations. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*.

Raudenbush, S. and Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage Pub.

Reinsel, G. and Velu, R. (1998). *Multivariate reduced-rank regression*. Springer New York.

Srebro, N., Rennie, J. D. M., and Jaakkola, T. S. (2005). Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*.

References V

Witten, D., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67.

Zhao, P., Rocha, G., and Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.*, 15:265–286.