Sparse methods for machine learning

Francis Bach

Willow project, INRIA - Ecole Normale Supérieure

CVPR Tutorial - June 2010

Special thanks to R. Jenatton, G. Obozinski

Sparse methods for machine learning
Outline

- **Sparse linear estimation with the ℓ_1-norm**
  - Lasso
  - Important theoretical results
- **Structured sparse methods on vectors**
  - Groups of features / Multiple kernel learning
- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)

Supervised learning and regularization

- Data: $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $i = 1, \dots, n$
- Minimize with respect to function $f : \mathcal{X} \to \mathcal{Y}$:

$$\sum_{i=1}^{n} \ell(y_i, f(x_i)) \quad + \quad \frac{\lambda}{2}\|f\|^2$$

Error on data + Regularization

Loss & function space ? Norm ?

- Two theoretical/algorithmic issues:

1. Loss
2. **Function space / norm**

Regularizations

- **Main goal: avoid overfitting**

- **Two main lines of work**:

  1. Euclidean and Hilbertian norms (i.e., ℓ_2-norms)
     - Possibility of non linear predictors
     - Non parametric supervised learning and kernel methods
     - Well developped theory and algorithms (see, e.g., Wahba, 1990; Schölkopf and Smola, 2001; Shawe-Taylor and Cristianini, 2004)

Regularizations

- **Main goal: avoid overfitting**

- **Two main lines of work**:

1. Euclidean and Hilbertian norms (i.e., ℓ_2-norms)
   - Possibility of non linear predictors
   - Non parametric supervised learning and kernel methods
   - Well developped theory and algorithms (see, e.g., Wahba, 1990; Schölkopf and Smola, 2001; Shawe-Taylor and Cristianini, 2004)
2. Sparsity-inducing norms
   - Usually restricted to linear predictors on vectors $f(x) = w^\top x$
   - Main example: ℓ_1-norm $\|w\|_1 = \sum_{i=1}^{p} |w_i|$
   - Perform model selection as well as regularization
   - **Theory and algorithms "in the making"**

ℓ_2-norm vs. ℓ_1-norm

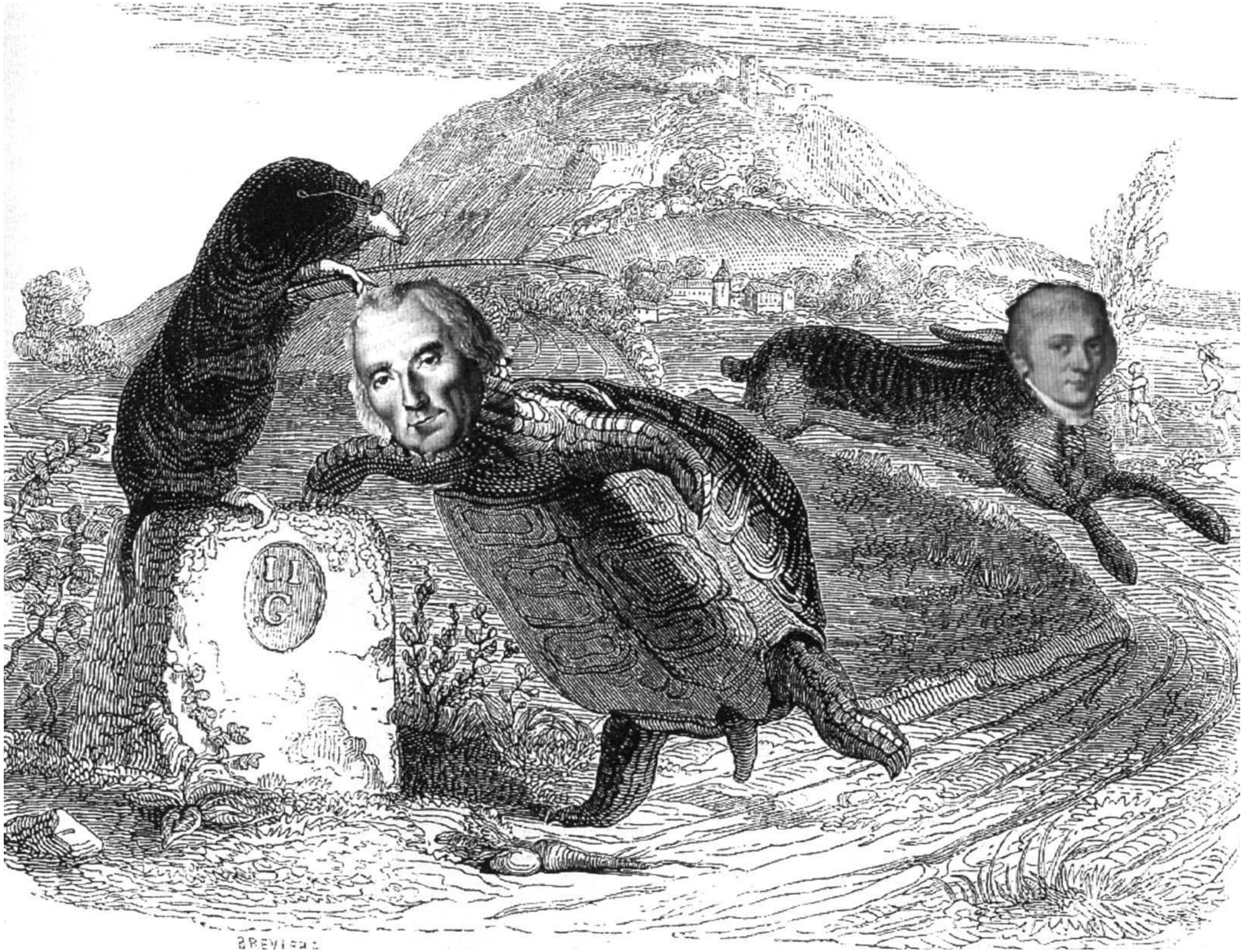
- ℓ_1-norms lead to interpretable models

- ℓ_2-norms can be run implicitly with very large feature spaces (e.g., kernel trick)

- **Algorithms**:
  - Smooth convex optimization vs. nonsmooth convex optimization

- **Theory**:
  - better predictive performance?

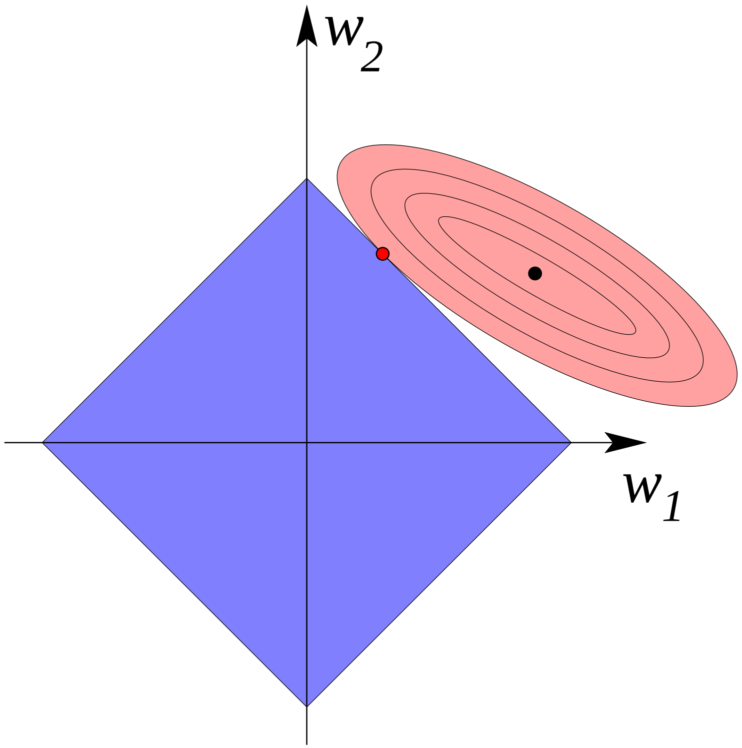
ℓ_2 vs. ℓ_1 - Gaussian hare vs. Laplacian tortoise

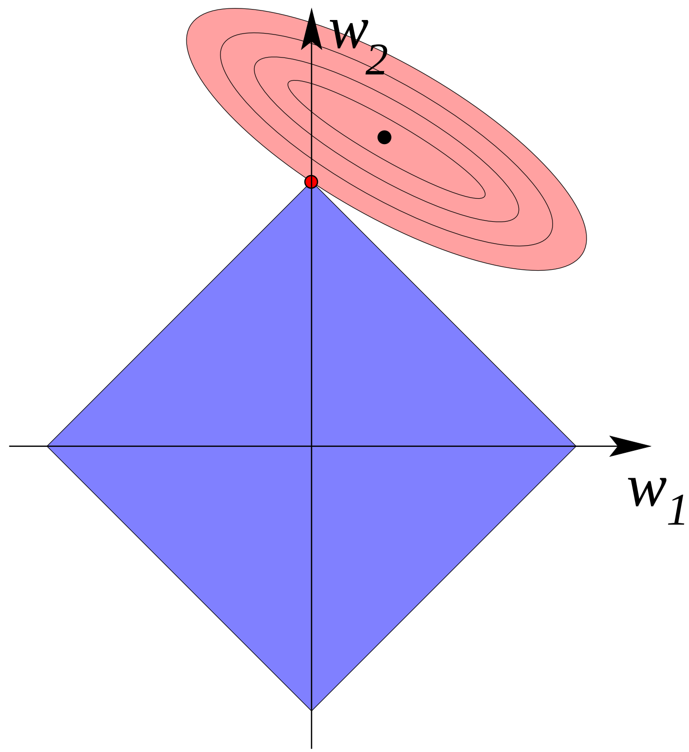


- First-order methods (Fu, 1998; Beck and Teboulle, 2009)
- Homotopy methods (Markowitz, 1956; Efron et al., 2004)

Why ℓ_1-norm constraints leads to sparsity?

- Example: minimize quadratic function $Q(w)$ subject to $\|w\|_1 \leqslant T$.
  - coupled soft thresholding
- Geometric interpretation
  - NB : penalizing is "equivalent" to constraining





ℓ_1-norm regularization (linear setting)

- Data: covariates $x_i \in \mathbb{R}^p$, responses $y_i \in \mathcal{Y}$, $i = 1, \dots, n$

- Minimize with respect to loadings/weights $w \in \mathbb{R}^p$:

$$J(w) = \sum_{i=1}^{n} \ell(y_i, w^\top x_i) \quad + \quad \lambda \|w\|_1$$

Error on data + Regularization

- Including a constant term b? Penalizing or constraining?

- square loss $\Rightarrow$ basis pursuit in signal processing (Chen et al., 2001), Lasso in statistics/machine learning (Tibshirani, 1996)

Lasso - Two main recent theoretical results

1. **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if

$$\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\text{sign}(\mathbf{w}_{\mathbf{J}})\|_\infty \leqslant 1,$$

where $\mathbf{Q} = \lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{p\times p}$ and $\mathbf{J} = \text{Supp}(\mathbf{w})$

Lasso - Two main recent theoretical results

1. **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if

$$\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\text{sign}(\mathbf{w}_{\mathbf{J}})\|_\infty \leqslant 1,$$

where $\mathbf{Q} = \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{p\times p}$ and $\mathbf{J} = \text{Supp}(\mathbf{w})$

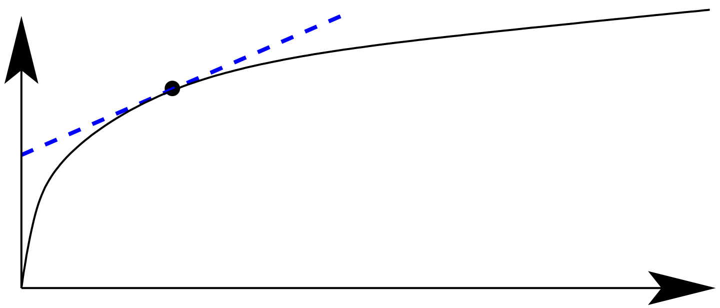
- **The Lasso is usually not model-consistent**
  - Selects more variables than necessary (see, e.g., Lv and Fan, 2009)
  - **Fixing the Lasso**: adaptive Lasso (Zou, 2006), relaxed Lasso (Meinshausen, 2008), thresholding (Lounici, 2008), Bolasso (Bach, 2008a), stability selection (Meinshausen and Bühlmann, 2008)

Adaptive Lasso and concave penalization

- **Adaptive Lasso** (Zou, 2006; Huang et al., 2008)
  - Weighted ℓ_1-norm: $\min_{w \in \mathbb{R}^p} L(w) + \lambda \sum_{j=1}^{p} \frac{|w_j|}{|\hat{w}_j|^\alpha}$
  - $\hat{w}$ estimator obtained from ℓ_2 or ℓ_1 regularization

- **Reformulation in terms of concave penalization**

$$\min_{w \in \mathbb{R}^p} L(w) + \sum_{j=1}^{p} g(|w_j|)$$



  - Example: $g(|w_j|) = |w_j|^{1/2}$ or $\log |w_j|$. Closer to the ℓ_0 penalty
  - Concave-convex procedure: replace $g(|w_j|)$ by affine upper bound
  - Better sparsity-inducing properties (Fan and Li, 2001; Zou and Li, 2008; Zhang, 2008b)

Lasso - Two main recent theoretical results

1. **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if
$$\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\mathrm{sign}(\mathbf{w}_\mathbf{J})\|_\infty \leqslant 1,$$
where $\mathbf{Q} = \lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{p\times p}$ and $\mathbf{J} = \mathrm{Supp}(\mathbf{w})$

Lasso - Two main recent theoretical results

1. **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if

$$\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\text{sign}(\mathbf{w}_{\mathbf{J}})\|_\infty \leqslant 1,$$

where $\mathbf{Q} = \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{p\times p}$ and $\mathbf{J} = \text{Supp}(\mathbf{w})$

2. **Exponentially many irrelevant variables** (Zhao and Yu, 2006; Wainwright, 2009; Bickel et al., 2009; Lounici, 2008; Meinshausen and Yu, 2008): under appropriate assumptions, consistency is possible as long as

$$\log p = O(n)$$

Alternative sparse methods
Greedy methods

- Forward selection
- Forward-backward selection
- Non-convex method
  - Harder to analyze
  - Simpler to implement
  - Problems of stability
- Positive theoretical results (Zhang, 2009, 2008a)
  - Similar sufficient conditions than for the Lasso
- **Bayesian methods** : see Seeger (2008)

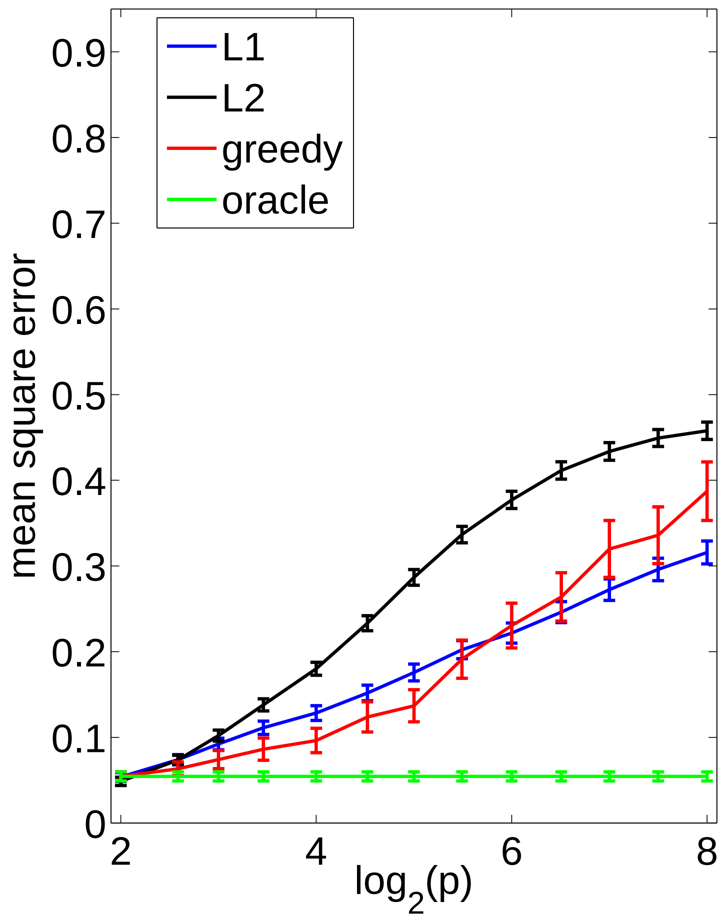
Comparing Lasso and other strategies for linear regression

- Compared methods to reach the least-square solution
  - Ridge regression: $\displaystyle \min_{w\in\mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$
  - Lasso: $\displaystyle \min_{w\in\mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_1$
  - Forward greedy:
    * Initialization with empty set
    * Sequentially add the variable that best reduces the square loss

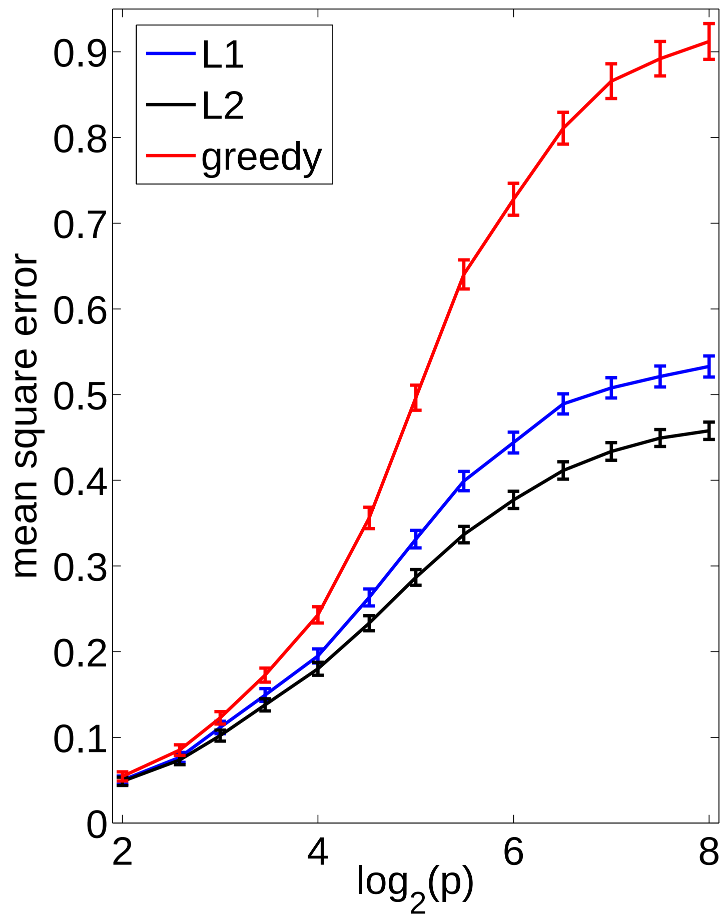
- Each method builds a path of solutions from 0 to ordinary least-squares solution

Simulation results

- i.i.d. Gaussian design matrix, $k = 4$, $n = 64$, $p \in [2, 256]$, SNR $= 1$
- Note stability to non-sparsity and variability



Sparse



Rotated (non sparse)

Extensions - Going beyond the Lasso

- ℓ_1-norm for **linear** feature selection in **high dimensions**
  - Lasso usually not applicable directly

Extensions - Going beyond the Lasso

- ℓ_1-norm for **linear** feature selection in **high dimensions**
  - Lasso usually not applicable directly
- **Sparse methods are not limited to the square loss**
  - logistic loss: algorithms (Beck and Teboulle, 2009) and theory (Van De Geer, 2008; Bach, 2009)
- **Sparse methods are not limited to supervised learning**
  - Learning the structure of Gaussian graphical models (Meinshausen and Bühlmann, 2006; Banerjee et al., 2008)
  - Sparsity on matrices (last part of this session)
- **Sparse methods are not limited to linear variable selection**
  - Multiple kernel learning (next part of this session)

Sparse methods for machine learning
Outline

- **Sparse linear estimation with the ℓ_1-norm**
  - Lasso
  - Important theoretical results

- **Structured sparse methods on vectors**
  - Groups of features / Multiple kernel learning

- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)

Penalization with grouped variables (Yuan and Lin, 2006)

- Assume that $\{1,\ldots,p\}$ is **partitioned** into m groups $G_1,\ldots,G_m$
- Penalization by $\sum_{i=1}^{m} \|w_{G_i}\|_2$, often called ℓ_1-ℓ_2 norm
- Induces **group sparsity**
  - Some groups entirely set to zero
  - no zeros within groups
- In this tutorial:
  - Groups may have infinite size $\Rightarrow$ **MKL**
  - Groups may overlap $\Rightarrow$ **structured sparsity**

Linear vs. non-linear methods

- All methods in this tutorial are **linear in the parameters**
- By replacing x by features $\Phi(x)$, they can be made **non linear in the data**
- **Implicit vs. explicit features**
  - ℓ_1-norm: explicit features
  - ℓ_2-norm: representer theorem allows to consider implicit features if their dot products can be computed easily (kernel methods)

Kernel methods: regularization by ℓ_2-norm

- Data: $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $i = 1, \dots, n$, with **features** $\Phi(x) \in \mathcal{F} = \mathbb{R}^p$
  - Predictor $f(x) = w^\top \Phi(x)$ linear in the features

- Optimization problem: $$\min_{w \in \mathbb{R}^p} \sum_{i=1}^n \ell(y_i, w^\top \Phi(x_i)) + \frac{\lambda}{2} \|w\|_{\mathbf{2}}^2$$

Kernel methods: regularization by ℓ_2-norm

- Data: $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $i = 1, \dots, n$, with **features** $\Phi(x) \in \mathcal{F} = \mathbb{R}^p$
  - Predictor $f(x) = w^\top \Phi(x)$ linear in the features
- Optimization problem:

$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^n \ell(y_i, w^\top \Phi(x_i)) + \frac{\lambda}{2} \|w\|_{\mathbf{2}}^2$$

- **Representer theorem** (Kimeldorf and Wahba, 1971): solution must be of the form $w = \sum_{i=1}^n \alpha_i \Phi(x_i)$
  - Equivalent to solving:

$$\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^\top K \alpha$$

  - Kernel matrix $K_{ij} = k(x_i, x_j) = \Phi(x_i)^\top \Phi(x_j)$

Multiple kernel learning (MKL) (Lanckriet et al., 2004b; Bach et al., 2004a)

- **Sparsity with non-linearities**
  - replace $f(x) = \sum_{j=1}^{p} w_j^\top x_j$ with $x \in \mathbb{R}^p$ and $w_j \in \mathbb{R}$
  - by $f(x) = \sum_{j=1}^{p} w_j^\top \Phi_j(x)$ with $x \in \mathcal{X}$, $\Phi_j(x) \in \mathcal{F}_j$ an $w_j \in \mathcal{F}_j$
- Replace the ℓ_1-norm $\sum_{j=1}^{p} |w_j|$ by "block" ℓ_1-norm $\sum_{j=1}^{p} \|w_j\|_2$
- Multiple feature maps / kernels on $x \in \mathcal{X}$:
  - p "feature maps" $\Phi_j : \mathcal{X} \mapsto \mathcal{F}_j$, $j = 1, \dots, p$.
  - Predictor: $f(x) = {w_1}^\top \Phi_1(x) + \cdots + {w_p}^\top \Phi_p(x)$
  - Generalized additive models (Hastie and Tibshirani, 1990)

Regularization for multiple features

$$
\begin{array}{ccccccc}
 & & \Phi_1(x)^\top & w_1 & & & \\
 & \nearrow & \vdots & \vdots & \searrow & & \\
x & \longrightarrow & \Phi_j(x)^\top & w_j & \longrightarrow & w_1^\top \Phi_1(x) + \cdots + w_p^\top \Phi_p(x) \\
 & \searrow & \vdots & \vdots & \nearrow & & \\
 & & \Phi_p(x)^\top & w_p & & &
\end{array}
$$

- Regularization by $\sum_{j=1}^{p} \|w_j\|_2^2$ is equivalent to using $K = \sum_{j=1}^{p} K_j$
  - Summing kernels is equivalent to concatenating feature spaces

Regularization for multiple features

$$
\begin{array}{ccccc}
 & & \Phi_1(x)^\top \quad w_1 & & \\
 & \nearrow & \vdots \qquad \vdots & \searrow & \\
x & \longrightarrow & \Phi_j(x)^\top \quad w_j & \longrightarrow & w_1^\top \Phi_1(x) + \cdots + w_p^\top \Phi_p(x) \\
 & \searrow & \vdots \qquad \vdots & \nearrow & \\
 & & \Phi_p(x)^\top \quad w_p & &
\end{array}
$$

- Regularization by $\sum_{j=1}^{p} \|w_j\|_2^2$ is equivalent to using $K = \sum_{j=1}^{p} K_j$
- Regularization by $\sum_{j=1}^{p} \|w_j\|_2$ imposes sparsity at the group level
- **Main questions when regularizing by block ℓ_1-norm**:
  1. Algorithms (Bach et al., 2004a; Rakotomamonjy et al., 2008)
  2. Analysis of sparsity inducing properties (Bach, 2008b)
  3. Equivalent to learning a **sparse** combination $\sum_{j=1}^{p} \eta_j K_j$

Applications of multiple kernel learning

- **Selection of hyperparameters for kernel methods**
- **Fusion from heterogeneous data sources** (Lanckriet et al., 2004a)
- **Two regularizations on the same function space**:
  - Uniform combination $\Leftrightarrow$ ℓ_2-norm
  - Sparse combination $\Leftrightarrow$ ℓ_1-norm
  - MKL always leads to more interpretable models
  - MKL does not always lead to better predictive performance
    - In particular, with few well-designed kernels
    - Be careful with normalization of kernels (Bach et al., 2004b)

Applications of multiple kernel learning

- **Selection of hyperparameters for kernel methods**

- **Fusion from heterogeneous data sources** (Lanckriet et al., 2004a)

- **Two regularizations on the same function space**:
  - Uniform combination $\Leftrightarrow$ ℓ_2-norm
  - Sparse combination $\Leftrightarrow$ ℓ_1-norm
  - MKL always leads to more interpretable models
  - MKL does not always lead to better predictive performance
    - In particular, with few well-designed kernels
    - Be careful with normalization of kernels (Bach et al., 2004b)

- **Sparse methods**: new possibilities and new features

Sparse methods for machine learning
Outline

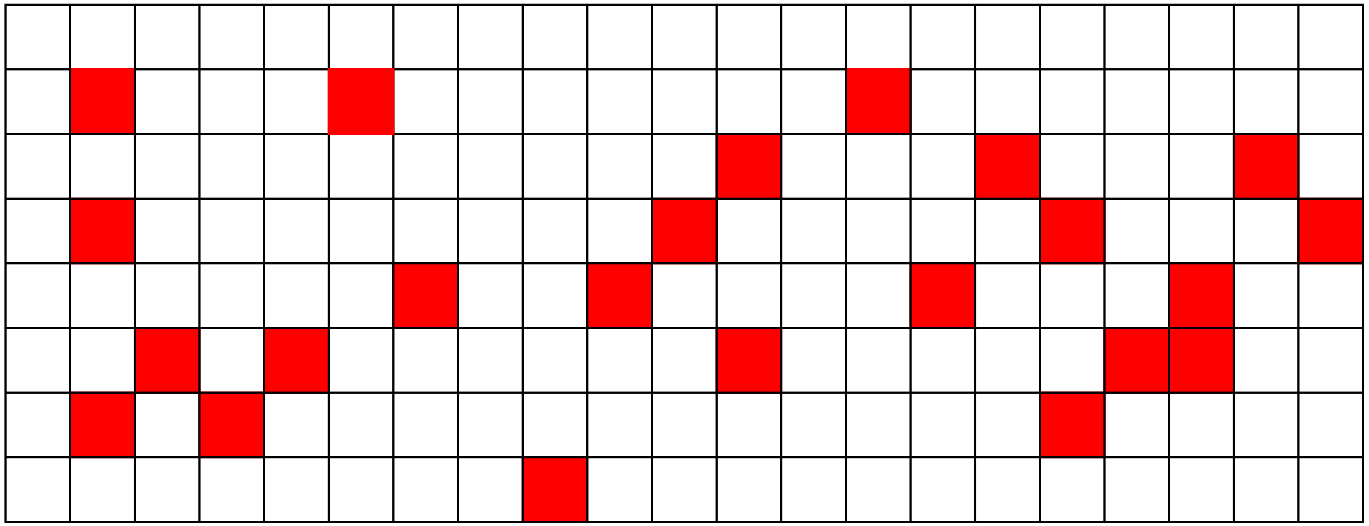
- **Sparse linear estimation with the ℓ_1-norm**
  - Lasso
  - Important theoretical results
- **Structured sparse methods on vectors**
  - Groups of features / Multiple kernel learning
- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)

Learning on matrices - Image denoising

- Simultaneously denoise all patches of a given image
- Example from Mairal, Bach, Ponce, Sapiro, and Zisserman (2009b)

Learning on matrices - Collaborative filtering

- 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 $\mathbf{Z}$ that describes the known ratings of some customers for some movies
- **Goal**: complete the matrix.



Learning on matrices - Multi-task learning

- k linear prediction tasks on same covariates $\mathbf{x} \in \mathbb{R}^p$
  - k weight vectors $\mathbf{w}_j \in \mathbb{R}^p$
  - Joint matrix of predictors $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_k) \in \mathbb{R}^{p \times k}$

- Classical application
  - Multi-category classification (one task per class) (Amit et al., 2007)

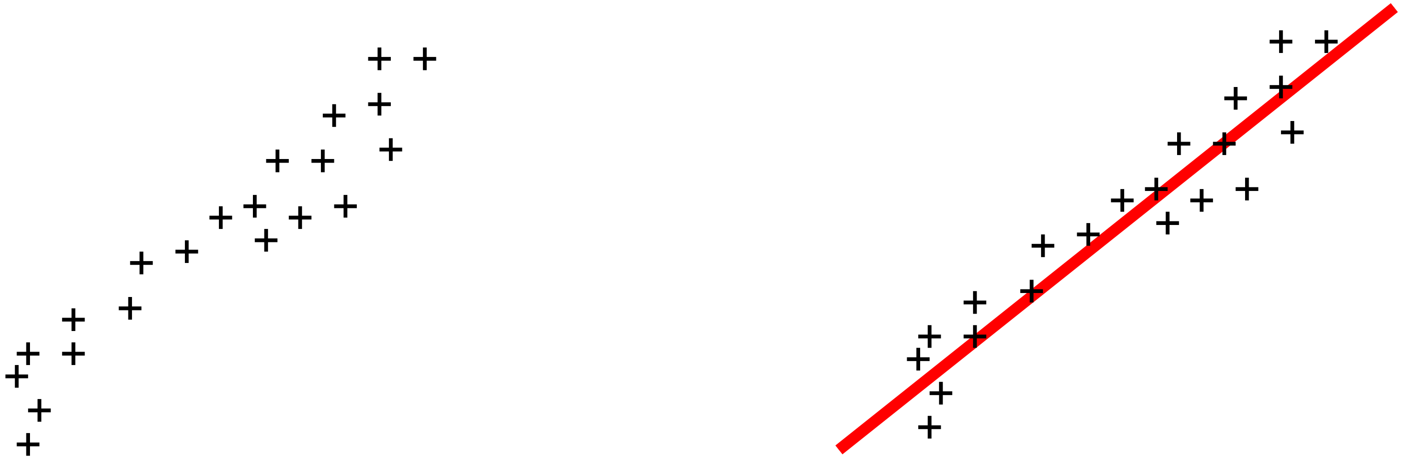
- **Share parameters between tasks**

- **Joint variable selection** (Obozinski et al., 2009)
  - Select variables which are predictive for all tasks

- **Joint feature selection** (Pontil et al., 2007)
  - Construct linear features common to all tasks

Matrix factorization - Dimension reduction

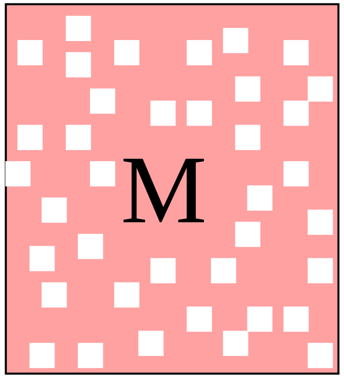
- Given data matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{p \times n}$
  - **Principal component analysis**: $\boxed{\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i \Rightarrow \mathbf{X} = \mathbf{D}\mathbf{A}}$



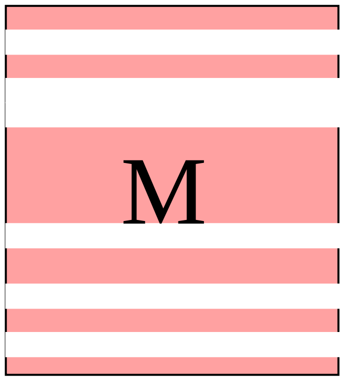
  - **K-means**: $\boxed{\mathbf{x}_i \approx \mathbf{d}_k \Rightarrow \mathbf{X} = \mathbf{D}\mathbf{A}}$

Two types of sparsity for matrices $\mathbf{M} \in \mathbb{R}^{n \times p}$
I - Directly on the elements of $\mathbf{M}$

- Many zero elements: $\mathbf{M}_{ij} = 0$

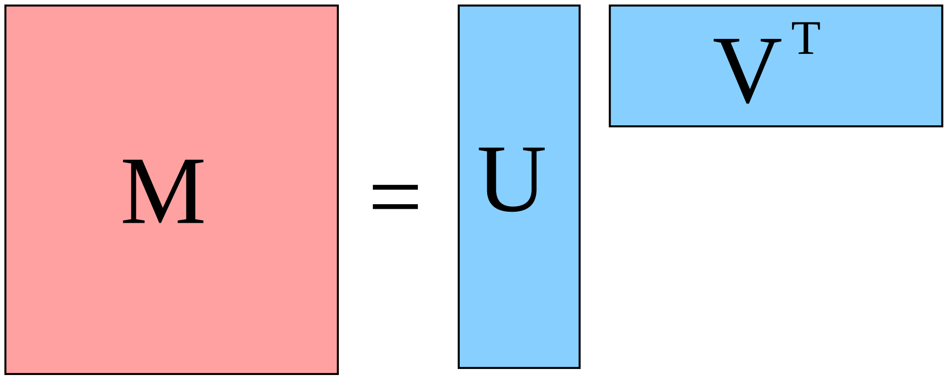


- Many zero rows (or columns): $(\mathbf{M}_{i1}, \ldots, \mathbf{M}_{ip}) = 0$

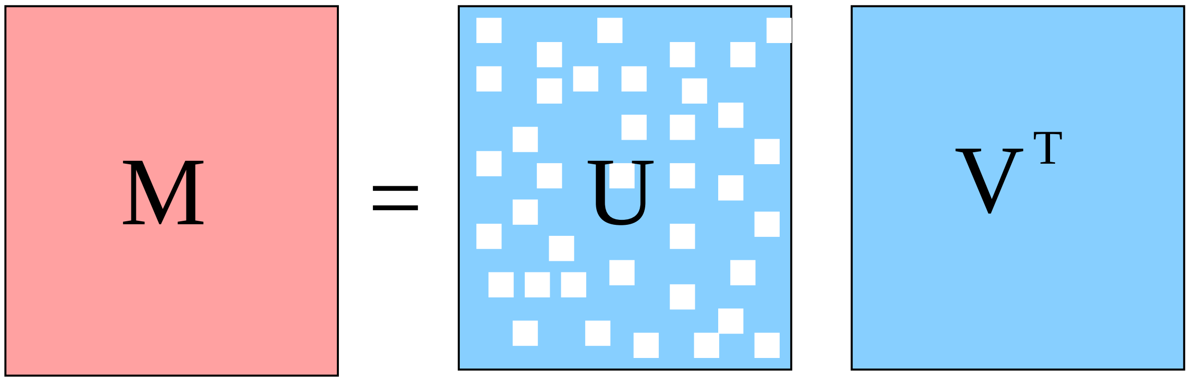


Two types of sparsity for matrices $\mathbf{M} \in \mathbb{R}^{n \times p}$
II - Through a factorization of $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$

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



- **Sparse decomposition**: $\mathbf{U}$ sparse



Structured sparse matrix factorizations

- Matrix $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times k}$ and $\mathbf{V} \in \mathbb{R}^{p \times k}$

- **Structure on $\mathbf{U}$ and/or $\mathbf{V}$**
  - Low-rank: $\mathbf{U}$ and $\mathbf{V}$ have few columns
  - Dictionary learning / sparse PCA: $\mathbf{U}$ has many zeros
  - Clustering (k-means): $\mathbf{U} \in \{0,1\}^{n \times m}$, $\mathbf{U}\mathbf{1} = \mathbf{1}$
  - Pointwise positivity: non negative matrix factorization (NMF)
  - Specific patterns of zeros (Jenatton et al., 2010)
  - Low-rank + sparse (Candès et al., 2009)
  - etc.

- **Many applications**

- **Many open questions** (Algorithms, identifiability, etc.)

Low-rank matrix factorizations
Trace norm

- Given a matrix $\mathbf{M} \in \mathbb{R}^{n \times p}$
  - Rank of $\mathbf{M}$ is the minimum size m of **all** factorizations of $\mathbf{M}$ into $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times m}$ and $\mathbf{V} \in \mathbb{R}^{p \times m}$
  - Singular value decomposition: $\mathbf{M} = \mathbf{U}\,\mathrm{Diag}(\mathbf{s})\mathbf{V}^\top$ where $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns and $\mathbf{s} \in \mathbb{R}_+^m$ are singular values

- Rank of $\mathbf{M}$ equal to the number of non-zero singular values

Low-rank matrix factorizations
Trace norm

- Given a matrix $\mathbf{M} \in \mathbb{R}^{n \times p}$
  - Rank of $\mathbf{M}$ is the minimum size m of **all** factorizations of $\mathbf{M}$ into $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times m}$ and $\mathbf{V} \in \mathbb{R}^{p \times m}$
  - Singular value decomposition: $\mathbf{M} = \mathbf{U}\,\mathrm{Diag}(\mathbf{s})\mathbf{V}^\top$ where $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns and $\mathbf{s} \in \mathbb{R}_+^m$ are singular values

- Rank of $\mathbf{M}$ equal to the number of non-zero singular values

- **Trace-norm (a.k.a. nuclear norm)** $=$ sum of singular values

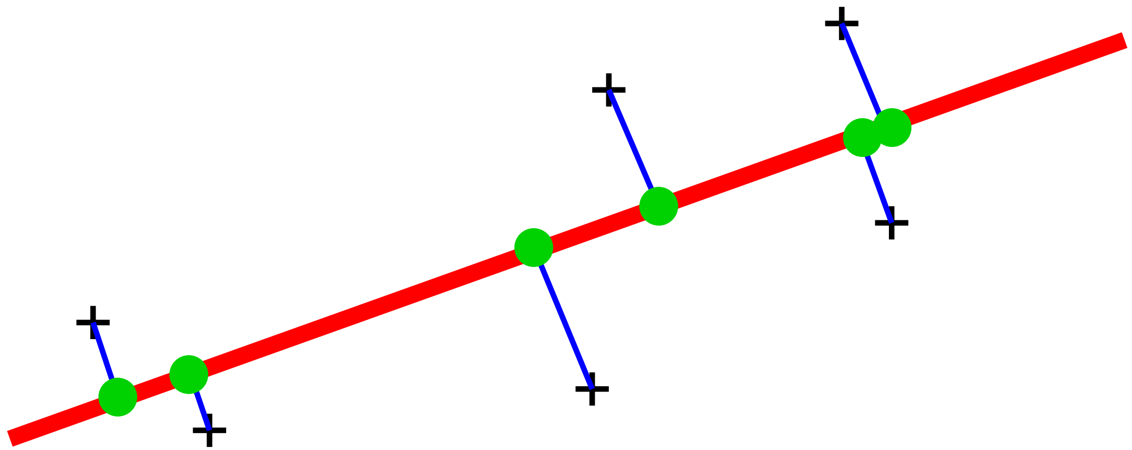
- Convex function, leads to a semi-definite program (Fazel et al., 2001)

- First used for collaborative filtering (Srebro et al., 2005)

Sparse principal component analysis

- Given data $\mathbf{X} = (\mathbf{x}_1^\top, \dots, \mathbf{x}_n^\top) \in \mathbb{R}^{p \times n}$, two views of PCA:
  - **Analysis view**: find the projection $\mathbf{d} \in \mathbb{R}^p$ of maximum variance (with deflation to obtain more components)
  - **Synthesis view**: find the basis $\mathbf{d}_1, \dots, \mathbf{d}_k$ such that all $\mathbf{x}_i$ have low reconstruction error when decomposed on this basis

- For regular PCA, the two views are equivalent



Sparse principal component analysis

- Given data $\mathbf{X} = (\mathbf{x}_1^\top, \dots, \mathbf{x}_n^\top) \in \mathbb{R}^{p \times n}$, two views of PCA:
  - **Analysis view**: find the projection $\mathbf{d} \in \mathbb{R}^p$ of maximum variance (with deflation to obtain more components)
  - **Synthesis view**: find the basis $\mathbf{d}_1, \dots, \mathbf{d}_k$ such that all $\mathbf{x}_i$ have low reconstruction error when decomposed on this basis

- For regular PCA, the two views are equivalent

- **Sparse extensions**
  - Interpretability
  - High-dimensional inference
  - Two views are differents
    - For analysis view, see d'Aspremont, Bach, and El Ghaoui (2008)

Sparse principal component analysis
Synthesis view

- Find $\mathbf{d}_1, \dots, \mathbf{d}_k \in \mathbb{R}^p$ **sparse** so that

$$\sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \sum_{j=1}^{k} (\boldsymbol{\alpha}_i)_j \mathbf{d}_j \right\|_2^2 = \sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i \right\|_2^2 \text{ is small}$$

  – Look for $\mathbf{A} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n) \in \mathbb{R}^{k \times n}$ and $\mathbf{D} = (\mathbf{d}_1, \dots, \mathbf{d}_k) \in \mathbb{R}^{p \times k}$ such that $\mathbf{D}$ is sparse and $\|\mathbf{X} - \mathbf{DA}\|_F^2$ is small

Sparse principal component analysis
Synthesis view

- Find $\mathbf{d}_1, \dots, \mathbf{d}_k \in \mathbb{R}^p$ **sparse** so that

$$\sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \sum_{j=1}^{k} (\boldsymbol{\alpha}_i)_j \mathbf{d}_j \right\|_2^2 = \sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 \text{ is small}$$

  – Look for $\mathbf{A} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n) \in \mathbb{R}^{k \times n}$ and $\mathbf{D} = (\mathbf{d}_1, \dots, \mathbf{d}_k) \in \mathbb{R}^{p \times k}$ such that $\mathbf{D}$ is sparse and $\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2$ is small

- Sparse formulation (Witten et al., 2009; Bach et al., 2008)
  – Penalize/constrain $\mathbf{d}_j$ by the ℓ_1-norm for sparsity
  – Penalize/constrain $\boldsymbol{\alpha}_i$ by the ℓ_2-norm to avoid trivial solutions

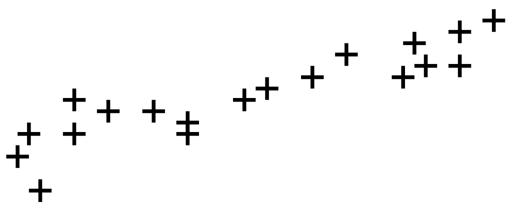
$$\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^{k} \|\mathbf{d}_j\|_1 \text{ s.t. } \forall i, \|\boldsymbol{\alpha}_i\|_2 \leqslant 1$$

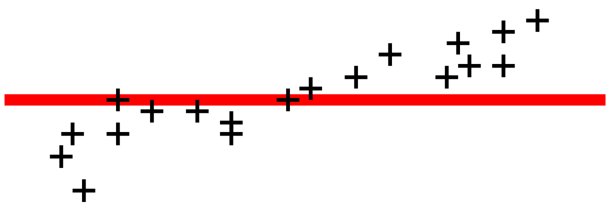
Sparse PCA vs. dictionary learning

- **Sparse PCA**: $\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i$, $\mathbf{D}$ sparse

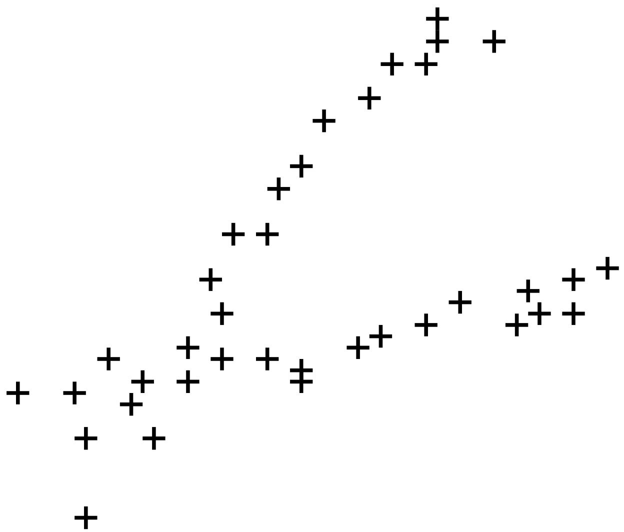
Sparse PCA vs. dictionary learning

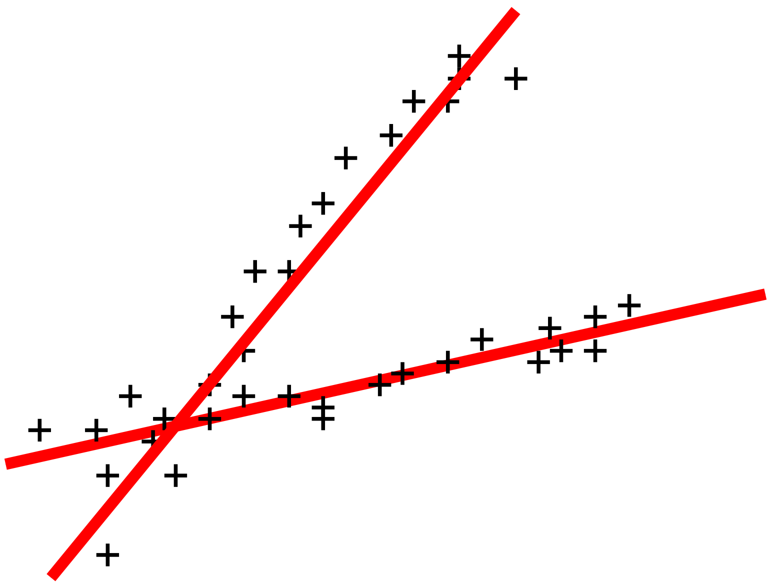
- **Sparse PCA**: $\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i$, $\mathbf{D}$ sparse





- **Dictionary learning**: $\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i$, $\boldsymbol{\alpha}_i$ sparse





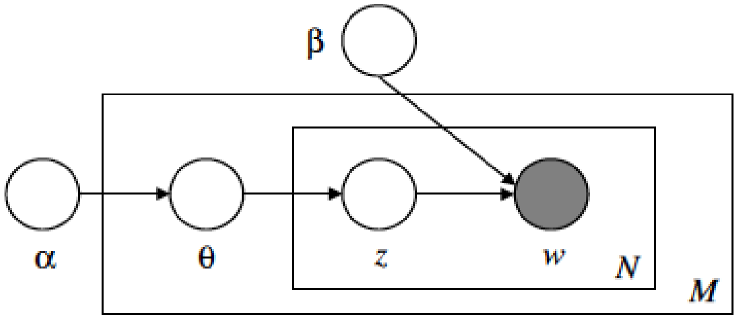
Structured matrix factorizations (Bach et al., 2008)

$$\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^{k} \|\mathbf{d}_j\|_\star \text{ s.t. } \forall i, \|\boldsymbol{\alpha}_i\|_\bullet \leqslant 1$$

$$\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{i=1}^{n} \|\boldsymbol{\alpha}_i\|_\bullet \text{ s.t. } \forall j, \|\mathbf{d}_j\|_\star \leqslant 1$$

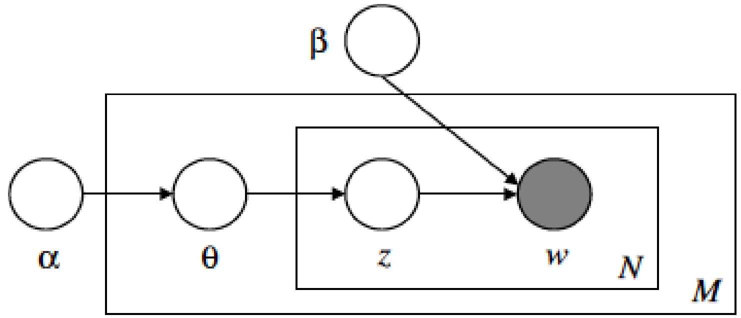
- Optimization by alternating minimization (non-convex)
- $\boldsymbol{\alpha}_i$ decomposition coefficients (or "code"), $\mathbf{d}_j$ dictionary elements
- Two related/equivalent problems:
  - **Sparse PCA = sparse dictionary** (ℓ_1-norm on $\mathbf{d}_j$)
  - **Dictionary learning = sparse decompositions** (ℓ_1-norm on $\boldsymbol{\alpha}_i$) (Olshausen and Field, 1997; Elad and Aharon, 2006; Lee et al., 2007)

Probabilistic topic models and matrix factorization



- **Latent Dirichlet allocation** (Blei et al., 2003)
  - For a document, sample $\theta \in \mathbb{R}^k$ from a Dirichlet(α)
  - For the n-th word of the same document,
    * sample a topic z_n from a multinomial with parameter θ
    * sample a word w_n from a multinomial with parameter $\beta(z_n, :)$

Probabilistic topic models and matrix factorization



- **Latent Dirichlet allocation** (Blei et al., 2003)
  - For a document, sample $\theta \in \mathbb{R}^k$ from a Dirichlet(α)
  - For the n-th word of the same document,
    * sample a topic z_n from a multinomial with parameter θ
    * sample a word w_n from a multinomial with parameter $\beta(z_n, :)$

- **Interpretation as multinomial PCA** (Buntine and Perttu, 2003)
  - Marginalizing over topic z_n, given θ, each word w_n is selected from a multinomial with parameter $\sum_{z=1}^{k} \theta_k \beta(z, :) = \beta^\top \theta$
  - Row of β = dictionary elements, θ code for a document

Probabilistic topic models and matrix factorization

- **Two different views on the same problem**
  - Interesting parallels to be made
  - Common problems to be solved
- **Structure on dictionary/decomposition coefficients** with adapted priors (Blei et al., 2004; Jenatton et al., 2010)
- **Identifiability and interpretation/evaluation of results**
- **Discriminative tasks** (Blei and McAuliffe, 2008; Lacoste-Julien et al., 2008; Mairal et al., 2009a)
- **Optimization and local minima**
  - Online learning (Mairal et al., 2009c)

Sparse methods for machine learning
Why use sparse methods?

- **Sparsity as a proxy to interpretability**
  - Structured sparsity
- **Sparsity for high-dimensional inference**
  - Influence on feature design
- **Sparse methods are not limited to least-squares regression**
- **Faster training/testing**
- **Better predictive performance?**
  - Problems are sparse if you look at them the right way

Conclusion - Interesting questions/issues

- **Exponentially many features**
  - Can we algorithmically achieve $\log p = O(n)$?
  - Use structure among features (Bach, 2008c)

- **Norm design**
  - What type of behavior may be obtained with sparsity-inducing norms?

- **Overfitting convexity**
  - Do we actually need convexity for matrix factorization problems?
  - Convexity used in inner loops
  - Joint convexity requires reformulation (Bach et al., 2008)

References

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

F. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML)*, 2008a.

F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008b.

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

F. Bach. Self-concordant analysis for logistic regression. Technical Report 0910.4627, ArXiv, 2009.

F. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2004a.

F. Bach, R. Thibaux, and M. I. Jordan. Computing regularization paths for learning multiple kernels. In *Advances in Neural Information Processing Systems 17*, 2004b.

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

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, 9: 485–516, 2008.

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.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

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

D.M. Blei and J. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, 2008.

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

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

E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Arxiv preprint arXiv:0912.3599*, 2009.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

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

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of statistics*, 32 (2):407–451, 2004.

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

J. Fan and R. Li. Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1361, 2001.

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

W. Fu. Penalized regressions: the bridge vs. the Lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998).

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

J. Huang, S. Ma, and C.H. Zhang. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618, 2008.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proc. ICML*, 2010.

G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applicat.*, 33:82–95, 1971.

S. Lacoste-Julien, F. Sha, and M.I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. *Advances in Neural Information Processing Systems (NIPS) 21*, 2008.

G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20:2626–2635, 2004a.

G. R. G. Lanckriet, N. Cristianini, L. El Ghaoui, P. Bartlett, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004b.

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

K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

J. Lv and Y. Fan. A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics*, 37(6A):3498–3528, 2009.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. *Advances in Neural Information Processing Systems (NIPS)*, 21, 2009a.

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

Julien Mairal, F. Bach, J. Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. 2009c.

H. M. Markowitz. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly*, 3:111–133, 1956.

N. Meinshausen. Relaxed Lasso. *Computational Statistics and Data Analysis*, 52(1):374–393, 2008.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of statistics*, 34(3):1436, 2006.

N. Meinshausen and P. Bühlmann. Stability selection. Technical report, arXiv: 0809.2932, 2008.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2008.

G. Obozinski, B. Taskar, and M.I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, pages 1–22, 2009.

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.

M. Pontil, A. Argyriou, and T. Evgeniou. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, 2007.

A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2001.

M.W. Seeger. Bayesian inference and optimal design for the sparse linear model. *The Journal of Machine Learning Research*, 9:759–813, 2008.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

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

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

S. A. Van De Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36 (2):614, 2008.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming. *IEEE transactions on information theory*, 55(5):2183, 2009.

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

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

M. Yuan and Y. Lin. On the non-negative garrotte estimator. *Journal of The Royal Statistical Society Series B*, 69(2):143–161, 2007.

T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. *Advances in Neural Information Processing Systems*, 22, 2008a.

T. Zhang. Multi-stage convex relaxation for learning with sparse regularization. *Advances in Neural Information Processing Systems*, 22, 2008b.

T. Zhang. On the consistency of feature selection using greedy least squares regression. *The Journal of Machine Learning Research*, 10:555–568, 2009.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533, 2008.