

Dictionary learning: another approach to building topic models



Guillaume Obozinski

Sierra group - INRIA - ENS - Paris



RussIR 2012
Yaroslavl, August 6-10th 2012

Dealing with the large number of parameters in topic models

Alternative approaches

- Frequentist approach: regularize + optimize \rightarrow *Dictionary Learning*

$$\min_{\theta_i} -\log p(\mathbf{x}_i|\theta_i) + \lambda\Omega(\theta_i)$$

- Bayesian approach: prior + integrate \rightarrow Latent Dirichlet Allocation

$$p(\theta_i|\mathbf{x}_i, \alpha) \propto p(\mathbf{x}_i|\theta_i) p(\theta_i|\alpha)$$

- “Frequentist + Bayesian” \rightarrow integrate + optimize

$$\max_{\alpha} \prod_{i=1}^M \int p(\mathbf{x}_i|\theta_i) p(\theta_i|\alpha) d\theta$$

... called *Empirical Bayes* approach or **Type II Maximum Likelihood**

From regularized pLSI (multinomial PCA) to ... “dictionary learning”

$$\min_{\theta_i} -\log p(\mathbf{x}_i|\theta_i) + \lambda\Omega(\theta_i)$$

From regularized pLSI (multinomial PCA) to ... “dictionary learning”

$$\min_{\boldsymbol{\theta}_i} -\log p(\mathbf{x}_i|\boldsymbol{\theta}_i) + \lambda\Omega(\boldsymbol{\theta}_i)$$

$$\min_{\boldsymbol{\theta}_i} \frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i) \quad \text{s.t.} \quad \forall (i, k) \quad \theta_{ik} \geq 0,$$

From regularized pLSI (multinomial PCA) to ... “dictionary learning”

$$\min_{\boldsymbol{\theta}_i} -\log p(\mathbf{x}_i|\boldsymbol{\theta}_i) + \lambda\Omega(\boldsymbol{\theta}_i)$$

$$\min_{\boldsymbol{\theta}_i} \frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i) \quad \text{s.t.} \quad \forall (i, k) \quad \theta_{ik} \geq 0,$$

How to find the best \mathbf{B} in this formulation?

From regularized pLSI (multinomial PCA) to ... “dictionary learning”

$$\min_{\boldsymbol{\theta}_i} -\log p(\mathbf{x}_i|\boldsymbol{\theta}_i) + \lambda\Omega(\boldsymbol{\theta}_i)$$

$$\min_{\boldsymbol{\theta}_i} \frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i) \quad \text{s.t.} \quad \forall (i, k) \quad \theta_{ik} \geq 0,$$

How to find the best \mathbf{B} in this formulation?

$$\sum_{i=1}^M \min_{\boldsymbol{\theta}_i} \left[\frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i) \right]$$
$$\text{s.t.} \quad \forall (i, k), \quad \theta_{ik} \geq 0,$$

From regularized pLSI (multinomial PCA) to ... “dictionary learning”

$$\min_{\boldsymbol{\theta}_i} -\log p(\mathbf{x}_i|\boldsymbol{\theta}_i) + \lambda\Omega(\boldsymbol{\theta}_i)$$

$$\min_{\boldsymbol{\theta}_i} \frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i) \quad \text{s.t.} \quad \forall (i, k) \quad \theta_{ik} \geq 0,$$

How to find the best \mathbf{B} in this formulation?

$$\begin{aligned} \min_{\mathbf{B}} \quad & \sum_{i=1}^M \min_{\boldsymbol{\theta}_i} \left[\frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i) \right] \\ \text{s.t.} \quad & \forall (i, k), \quad \theta_{ik} \geq 0, \end{aligned}$$

From regularized pLSI (multinomial PCA) to ... “dictionary learning”

$$\min_{\boldsymbol{\theta}_i} -\log p(\mathbf{x}_i|\boldsymbol{\theta}_i) + \lambda\Omega(\boldsymbol{\theta}_i)$$

$$\min_{\boldsymbol{\theta}_i} \frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i) \quad \text{s.t.} \quad \forall (i, k) \quad \theta_{ik} \geq 0,$$

How to find the best \mathbf{B} in this formulation?

$$\begin{aligned} \min_{\mathbf{B}} \quad & \sum_{i=1}^M \min_{\boldsymbol{\theta}_i} \left[\frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i) \right] \\ \text{s.t.} \quad & \forall (i, k), \quad \theta_{ik} \geq 0, \\ & \forall (i, j), \quad \mathbf{B}_{ji} \geq 0 \\ & \forall i, \quad \sum_{j=1}^d \mathbf{B}_{ji} = 1 \end{aligned}$$

A link to LSI?...

$$\begin{aligned} \min_{\mathbf{B}} \quad & \sum_{i=1}^M \min_{\boldsymbol{\theta}_i} \left[\frac{1}{2} \|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda \Omega(\boldsymbol{\theta}_i) \right] \\ \text{s.t.} \quad & \forall(i, k), \quad \theta_{ik} \geq 0, \quad \forall(i, j), \quad \mathbf{B}_{ji} \geq 0, \quad \forall i, \quad \sum_{j=1}^d \mathbf{B}_{ji} = 1 \end{aligned}$$

A link to LSI?...

$$\begin{aligned} \min_{\mathbf{B}} \min_{\boldsymbol{\Theta}=\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_M} \sum_{i=1}^M & \left[\frac{1}{2} \|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda \Omega(\boldsymbol{\theta}_i) \right] \\ \text{s.t.} \quad & \forall(i, k), \quad \theta_{ik} \geq 0, \quad \forall(i, j), \quad \mathbf{B}_{ji} \geq 0, \quad \forall i, \quad \sum_{j=1}^d \mathbf{B}_{ji} = 1 \end{aligned}$$

Rewriting as matrix factorization problem:

A link to LSI?...

$$\begin{aligned} \min_{\mathbf{B}} \min_{\Theta=\theta_1, \dots, \theta_M} \sum_{i=1}^M & \left[\frac{1}{2} \|\mathbf{x}_i - \mathbf{B}\theta_i\|_2^2 + \lambda \Omega(\theta_i) \right] \\ \text{s.t.} \quad & \forall(i, k), \quad \theta_{ik} \geq 0, \quad \forall(i, j), \quad \mathbf{B}_{ji} \geq 0, \quad \forall i, \quad \sum_{j=1}^d \mathbf{B}_{ji} = 1 \end{aligned}$$

Rewriting as matrix factorization problem:

$$\begin{aligned} \min_{\mathbf{B}} \min_{\Theta=\theta_1, \dots, \theta_M} \frac{1}{2} \|\mathbf{X} - \mathbf{B}\Theta\|_F^2 + \lambda \sum_{i=1}^M \Omega(\theta_i) \\ \text{s.t.} \quad & \forall(i, k), \quad \theta_{ik} \geq 0, \quad \forall(i, j), \quad \mathbf{B}_{ji} \geq 0, \quad \forall i, \quad \sum_{j=1}^d \mathbf{B}_{ji} = 1 \end{aligned}$$

What happens if we remove the constraints and regularization?

A link to LSI?...

$$\begin{aligned} \min_{\mathbf{B}} \min_{\Theta=\theta_1,\dots,\theta_M} \sum_{i=1}^M & \left[\frac{1}{2} \|\mathbf{x}_i - \mathbf{B}\theta_i\|_2^2 + \lambda \Omega(\theta_i) \right] \\ \text{s.t.} \quad & \forall(i, k), \quad \theta_{ik} \geq 0, \quad \forall(i, j), \quad \mathbf{B}_{ji} \geq 0, \quad \forall i, \quad \sum_{j=1}^d \mathbf{B}_{ji} = 1 \end{aligned}$$

Rewriting as matrix factorization problem:

$$\min_{\mathbf{B}} \min_{\Theta=\theta_1,\dots,\theta_M} \frac{1}{2} \|\mathbf{X} - \mathbf{B}\Theta\|_F^2$$

What happens if we remove the constraints and regularization?

A link to LSI?...

$$\begin{aligned} \min_{\mathbf{B}} \min_{\Theta=\theta_1, \dots, \theta_M} \sum_{i=1}^M \left[\frac{1}{2} \|\mathbf{x}_i - \mathbf{B}\theta_i\|_2^2 + \lambda \Omega(\theta_i) \right] \\ \text{s.t.} \quad \forall(i, k), \quad \theta_{ik} \geq 0, \quad \forall(i, j), \quad \mathbf{B}_{ji} \geq 0, \quad \forall i, \quad \sum_{j=1}^d \mathbf{B}_{ji} = 1 \end{aligned}$$

Rewriting as matrix factorization problem:

$$\min_{\mathbf{B}} \min_{\Theta=\theta_1, \dots, \theta_M} \frac{1}{2} \|\mathbf{X} - \mathbf{B}\Theta\|_F^2$$

What happens if we remove the constraints and regularization?

We get back **LSI**: $B = U_K$ and $\theta_i = \tilde{x}_i$

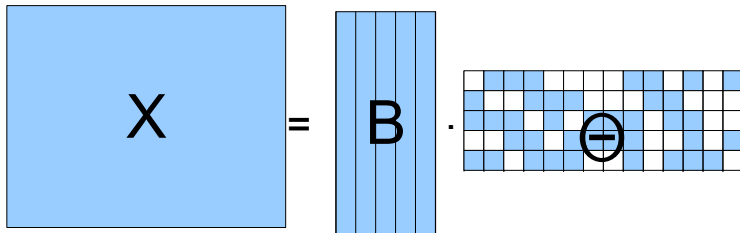
Topic models and matrix factorization

- $\mathbf{X} \in \mathbb{R}^{d \times M}$ with columns \mathbf{x}_i corresponding to documents
- \mathbf{B} the matrix whose columns correspond to different topics
- Θ the matrix of decomposition coefficients with columns θ_i associated each to one document and which encodes its “topic content”.

The diagram illustrates the matrix factorization equation $\mathbf{X} = \mathbf{B} \cdot \Theta$. On the left is a large light blue square labeled \mathbf{X} . To its right is an equals sign. Next is a tall light blue rectangle labeled \mathbf{B} , which is divided into five vertical stripes. To the right of \mathbf{B} is a dot, followed by a grid representing matrix Θ . The grid is 5 rows by 10 columns, with blue squares on a white background. A circled minus sign is placed over the grid, indicating that the matrix is sparse.

Topic models and matrix factorization

- $\mathbf{X} \in \mathbb{R}^{d \times M}$ with columns \mathbf{x}_i corresponding to documents
- \mathbf{B} the matrix whose columns correspond to different topics
- Θ the matrix of decomposition coefficients with columns θ_i associated each to one document and which encodes its “topic content”.



How about sparsity in topics?...

Ridge, penalization and sparsity

$$\min_{\boldsymbol{\theta}_i} \frac{1}{2} \|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda \Omega(\boldsymbol{\theta}_i)$$

A standard choice: $\Omega(\boldsymbol{\theta}) = \frac{1}{2} \|\boldsymbol{\theta}\|_2^2$

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|\mathbf{x} - \mathbf{B}\boldsymbol{\theta}\|_2^2 + \lambda \frac{1}{2} \|\boldsymbol{\theta}\|_2^2$$

This called Ridge regression, the most standard form of regression for a linear regression.

Ridge, penalization and sparsity

$$\min_{\boldsymbol{\theta}_i} \frac{1}{2} \|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda \Omega(\boldsymbol{\theta}_i)$$

A standard choice: $\Omega(\boldsymbol{\theta}) = \frac{1}{2} \|\boldsymbol{\theta}\|_2^2$

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|\mathbf{x} - \mathbf{B}\boldsymbol{\theta}\|_2^2 + \lambda \frac{1}{2} \|\boldsymbol{\theta}\|_2^2$$

This called Ridge regression, the most standard form of regression for a linear regression.

Can we choose Ω to obtain a sparse decomposition?

Define the pseudo ℓ_0 -norm $\|\boldsymbol{\theta}\|_0 = |\{k \mid \theta_k \neq 0\}|$

Ridge, penalization and sparsity

$$\min_{\boldsymbol{\theta}_i} \frac{1}{2} \|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda \Omega(\boldsymbol{\theta}_i)$$

A standard choice: $\Omega(\boldsymbol{\theta}) = \frac{1}{2} \|\boldsymbol{\theta}\|_2^2$

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|\mathbf{x} - \mathbf{B}\boldsymbol{\theta}\|_2^2 + \lambda \frac{1}{2} \|\boldsymbol{\theta}\|_2^2$$

This called Ridge regression, the most standard form of regression for a linear regression.

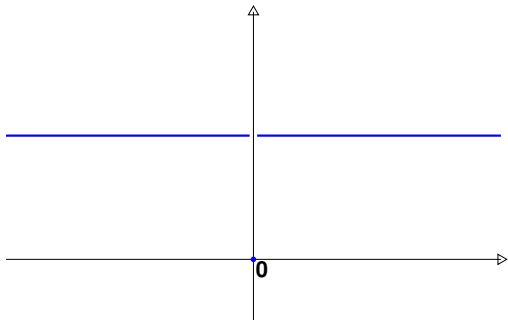
Can we choose Ω to obtain a sparse decomposition?

Define the pseudo ℓ_0 -norm $\|\boldsymbol{\theta}\|_0 = |\{k \mid \theta_k \neq 0\}|$

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|\mathbf{x} - \mathbf{B}\boldsymbol{\theta}\|_2^2 + \lambda \frac{1}{2} \|\boldsymbol{\theta}\|_2^2$$

Relaxing the ℓ_0 penalization

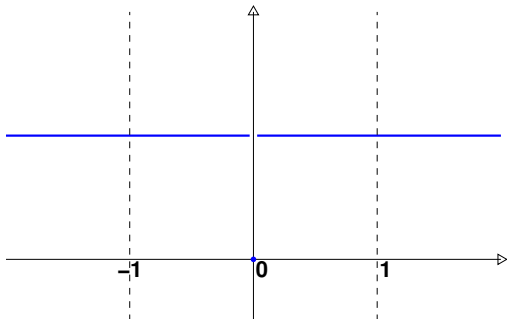
$$\|\theta\|_0 = \sum_{k=1}^K \mathbf{1}_{\{\theta_k \neq 0\}}$$



Relaxing the ℓ_0 penalization

$$\|\theta\|_0 = \sum_{k=1}^K \mathbf{1}_{\{\theta_k \neq 0\}}$$

Assume $\theta_k \in [-1, 1]$

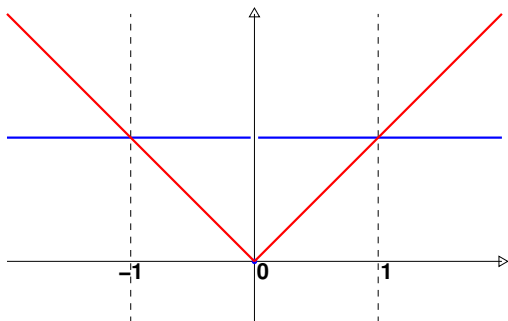


Relaxing the ℓ_0 penalization

$$\|\theta\|_0 = \sum_{k=1}^K \mathbf{1}_{\{\theta_k \neq 0\}}$$

Assume $\theta_k \in [-1, 1]$

Relax



Relaxing the ℓ_0 penalization

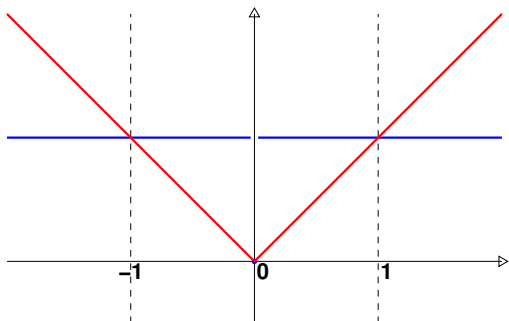
$$\|\theta\|_0 = \sum_{k=1}^K \mathbf{1}_{\{\theta_k \neq 0\}}$$

Assume $\theta_k \in [-1, 1]$

Relax

We obtain the ℓ_1 -norm:

$$\|\theta\|_1 = \sum_{k=1}^K |\theta_k|$$



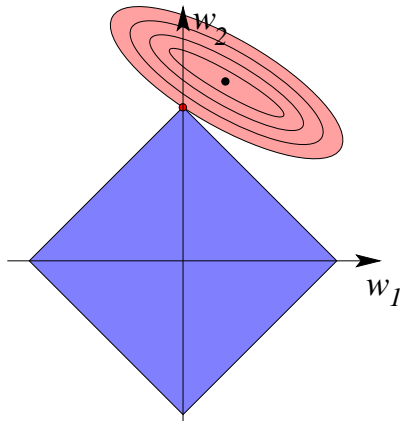
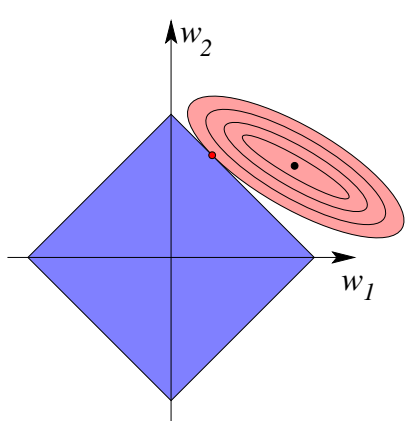
The LASSO (Tibshirani, 1996)

LASSO: Least Absolute Shrinkage and Selection operator

$$\min_{\boldsymbol{\theta}} \quad \frac{1}{2} \|\mathbf{x} - \mathbf{B}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1$$

Why ℓ_1 -norm constraints leads to sparsity?

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



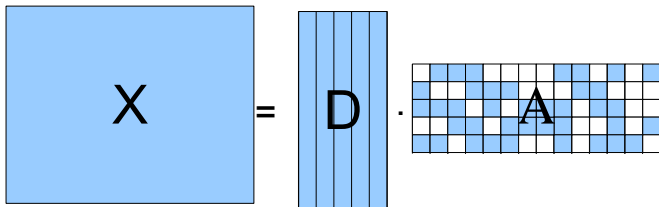
Decomposition of signals on a dictionary

The diagram shows the equation $X = D \cdot A$. Matrix X is represented by a solid light blue square. Matrix D is represented by a light blue rectangle with vertical lines, indicating its columns are the dictionary elements. Matrix A is represented by a light blue rectangle with vertical lines, indicating its columns are the coefficient vectors. The matrices are arranged horizontally with an equals sign and a dot operator between them.

- dictionary $\mathbf{D} = (\mathbf{d}^{(1)}, \dots, \mathbf{d}^{(K)})$ with $\mathbf{d}^{(k)}$ a dictionary element.
- matrix \mathbf{A} of loadings or decomposition coefficients vectors

Dictionary Learning

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



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

Structured matrix factorizations - Many instances

- $\mathbf{X} = \mathbf{DA}$, $\mathbf{D} \in \mathbb{R}^{p \times K}$ and $\mathbf{A} \in \mathbb{R}^{K \times M}$
- **Structure on \mathbf{D} and/or α**
 - Low-rank: \mathbf{D} and \mathbf{A}^\top have few columns
 - Dictionary learning / sparse PCA: \mathbf{D} or \mathbf{A} has many zeros
 - Clustering (k -means): $\mathbf{A} \in \{0, 1\}^{K \times M}$, $\mathbf{A}\mathbf{1} = \mathbf{1}$
 - 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

Inpainting a 12-Mpixel photograph

THE SALINAS VALLEY is in Northern California. It is a long narrow shale between two ranges of mountains, and the Salinas River winds and twists up the center until it falls at last into Monterey Bay.

I remember my childhood games for grasses and secret flowers. I remember where a road may live and what time the birds awoke in the summer and what trees and seasons smelled like how people looked and walked and smelled even. The memory of adults is very rich.

I remember that the Gabilan Mountains to the east of the valley were light gay mountains full of sun and loveliness and a kind of invitation, so that you wanted to climb into their warm forests almost as you want to climb into the lap of a beloved mother. They were beckoning mountains with a brown grass love. The Santa Lucias stood up against the sky to the west and kept the valley from the open sea, and they were dark and brooding unfriendly and dangerous. I always found in myself a dread of west and a love of east. Where I ever got such an idea I cannot say, unless it could be that the morning came over the peaks of the Gabilans and the night drifted back from the ridges of the Santa Lucias. It may be that the birth and death of the day had some part in my feeling about the two ranges of mountains.

From both sides of the valley little streams slipped out of the high canyons and fell into the bed of the Salinas River. In the winter of wet years the streams ran full-freshet, and they swelled the river until sometimes it raged and boiled, bank full, and then it was a destroyer. The river tore the edges of the farm lands and washed whole acres down; it toppled barns and houses into itself, to go floating and bobbing away. It trapped cows and pigs and sheep and drowned them in its muddy brown water and carried them to the sea. Then when the late spring came, the river drew off from its edges and the sand banks appeared. And in the summer the river didn't run at all above ground. Some pools would be left in the deep swirl places under a high bank. The tules and grasses grew back, and willows straightened up with the flood debris in their upper branches. The Salinas was only a part-time river. The summer sun drove it underground. It was not a fine river at all, but it was the only one we had and so we boasted about it how dangerous it was in a wet winter and how dry it was in a dry summer. You can boast about anything if it's all you have. Maybe the less you have, the more you are required to boast.

The floor of the Salinas Valley, between the ranges and below the foothills, is level because this valley used to be the bottom of a hundred-mile inlet from the sea. The river mouth at Moss Landing was centuries ago the entrance to this long inland water. Once, fifty miles down the valley, my father bored a well. The drill came up first with topsoil and then with gravel and then with white sea sand full of shells and even pi...

Inpainting a 12-Mpixel photograph



Inpainting a 12-Mpixel photograph



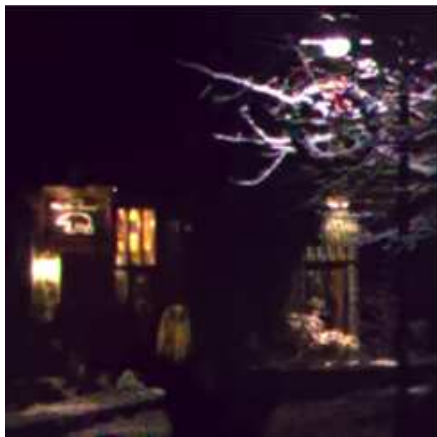
Inpainting a 12-Mpixel photograph



Denoising result (Mairal et al., 2009b)



Denoising result (Mairal et al., 2009b)



Variant of Dictionary Learning for topic models

$$\begin{aligned} \min_{\mathbf{D}, \mathbf{A}} \quad & \sum_{i=1}^M \|\mathbf{x}^{(i)} - \mathbf{D}\boldsymbol{\alpha}^{(i)}\|_2^2 + \lambda \sum_{i=1}^M \|\boldsymbol{\alpha}^{(i)}\|_1. \\ \text{s.t.} \quad & \boldsymbol{\alpha}^{(i)} \in \mathbb{R}_+^K, \\ & \mathbf{d}^{(k)} \in \mathbb{R}_+^p, \quad \mathbf{d}^\top \mathbf{1} = 1. \end{aligned}$$

Algorithms for sparse matrix factorization (Mairal et al., 2009a)

Focus on previous formulation:

$$\min_{\mathbf{D}, \mathbf{A}} \|\mathbf{X} - \mathbf{DA}\|_F^2 + \lambda \sum_{k=1}^K \|\alpha_k\|_1 \quad \text{s.t.} \quad \|\mathbf{d}^{(k)}\|_2 \leq 1$$

- Problem is convex in \mathbf{D} and \mathbf{A} separately, but not jointly.

Algorithms for sparse matrix factorization (Mairal et al., 2009a)

Focus on previous formulation:

$$\min_{\mathbf{D}, \mathbf{A}} \|\mathbf{X} - \mathbf{DA}\|_F^2 + \lambda \sum_{k=1}^K \|\alpha_k\|_1 \quad \text{s.t.} \quad \|\mathbf{d}^{(k)}\|_2 \leq 1$$

- Problem is convex in \mathbf{D} and \mathbf{A} separately, but not jointly.
→ Alternating scheme: optimize \mathbf{D} and \mathbf{A} in turn.

Algorithms for sparse matrix factorization (Mairal et al., 2009a)

Focus on previous formulation:

$$\min_{\mathbf{D}, \mathbf{A}} \|\mathbf{X} - \mathbf{DA}\|_F^2 + \lambda \sum_{k=1}^K \|\alpha_k\|_1 \quad \text{s.t.} \quad \|\mathbf{d}^{(k)}\|_2 \leq 1$$

- Problem is convex in \mathbf{D} and \mathbf{A} separately, but not jointly.
→ Alternating scheme: optimize \mathbf{D} and \mathbf{A} in turn.
- Even better: use simple column updates (Lee et al., 2007; Witten et al., 2009):

Algorithms for sparse matrix factorization (Mairal et al., 2009a)

Focus on previous formulation:

$$\min_{\mathbf{D}, \mathbf{A}} \|\mathbf{X} - \mathbf{DA}\|_F^2 + \lambda \sum_{k=1}^K \|\alpha_k\|_1 \quad \text{s.t. } \|\mathbf{d}^{(k)}\|_2 \leq 1$$

- Problem is convex in \mathbf{D} and \mathbf{A} separately, but not jointly.
→ Alternating scheme: optimize \mathbf{D} and \mathbf{A} in turn.
- Even better: use simple column updates (Lee et al., 2007; Witten et al., 2009):

$$\text{With } \tilde{\mathbf{X}} = \mathbf{X} - \sum_{j' \neq j} \mathbf{d}^{(k)} \alpha_k,$$

Algorithms for sparse matrix factorization (Mairal et al., 2009a)

Focus on previous formulation:

$$\min_{\mathbf{D}, \mathbf{A}} \|\mathbf{X} - \mathbf{DA}\|_F^2 + \lambda \sum_{k=1}^K \|\alpha_k\|_1 \quad \text{s.t.} \quad \|\mathbf{d}^{(k)}\|_2 \leq 1$$

- Problem is convex in \mathbf{D} and \mathbf{A} separately, but not jointly.
→ Alternating scheme: optimize \mathbf{D} and \mathbf{A} in turn.
- Even better: use simple column updates (Lee et al., 2007; Witten et al., 2009):

$$\text{With } \tilde{\mathbf{X}} = \mathbf{X} - \sum_{j' \neq j} \mathbf{d}^{(k)} \alpha_{j'}, \quad \text{we have } \mathbf{d}^{(k)} \leftarrow \frac{\tilde{\mathbf{X}} \alpha_k^\top}{\|\tilde{\mathbf{X}} \alpha_k^\top\|}$$

Algorithms for sparse matrix factorization (Mairal et al., 2009a)

Focus on previous formulation:

$$\min_{\mathbf{D}, \mathbf{A}} \|\mathbf{X} - \mathbf{DA}\|_F^2 + \lambda \sum_{k=1}^K \|\alpha_k\|_1 \quad \text{s.t.} \quad \|\mathbf{d}^{(k)}\|_2 \leq 1$$

- Problem is convex in \mathbf{D} and \mathbf{A} separately, but not jointly.
→ Alternating scheme: optimize \mathbf{D} and \mathbf{A} in turn.
- Even better: use simple column updates (Lee et al., 2007; Witten et al., 2009):

$$\text{With } \tilde{\mathbf{X}} = \mathbf{X} - \sum_{j' \neq j} \mathbf{d}^{(j')} \alpha_{j'}, \quad \text{we have } \mathbf{d}^{(k)} \leftarrow \frac{\tilde{\mathbf{X}} \alpha_k^\top}{\|\tilde{\mathbf{X}} \alpha_k^\top\|}$$

and

$$\alpha_k^\top \leftarrow \operatorname{argmin}_{\alpha \in \mathbb{R}^M} \|\mathbf{X}^\top \mathbf{d}^{(k)} - \alpha\|_2^2 + \lambda \|\alpha\|_1$$

Algorithms for sparse matrix factorization (Mairal et al., 2009a)

Focus on previous formulation:

$$\min_{\mathbf{D}, \mathbf{A}} \|\mathbf{X} - \mathbf{DA}\|_F^2 + \lambda \sum_{k=1}^K \|\alpha_k\|_1 \quad \text{s.t.} \quad \|\mathbf{d}^{(k)}\|_2 \leq 1$$

- Problem is convex in \mathbf{D} and \mathbf{A} separately, but not jointly.
→ Alternating scheme: optimize \mathbf{D} and \mathbf{A} in turn.
- Even better: use simple column updates (Lee et al., 2007; Witten et al., 2009):

$$\text{With } \tilde{\mathbf{X}} = \mathbf{X} - \sum_{j' \neq j} \mathbf{d}^{(j')} \alpha_{j'}, \quad \text{we have } \mathbf{d}^{(k)} \leftarrow \frac{\tilde{\mathbf{X}} \alpha_k^\top}{\|\tilde{\mathbf{X}} \alpha_k^\top\|}$$

and

$$\alpha_k^\top \leftarrow \operatorname{argmin}_{\alpha \in \mathbb{R}^M} \|\mathbf{X}^\top \mathbf{d}^{(k)} - \alpha\|_2^2 + \lambda \|\alpha\|_1$$

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

Algorithms for large databases

For large database it is significantly more efficient to use **online** algorithms and not batch algorithms.

For online algorithms for dictionary learning see: Mairal et al. (2009a)

For an online algorithm for variational Latent Dirichlet allocation: see Hoffman et al. (2010)

Structured Dictionary Learning and Structured Topic Models

Sparsity inducing norms

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

The most common choice for Ω :

- The ℓ_1 norm, $\|\mathbf{w}\|_1 = \sum_{j=1}^p |\mathbf{w}_j|$.
- Only **cardinality** is controlled!

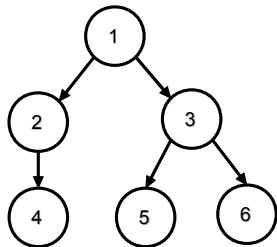
Another common choice for Ω :

- The ℓ_1 - ℓ_q norm (Yuan and Lin, 2007), with $q = 2$ or $q = \infty$

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

- The ℓ_1 - ℓ_q norm sets to zero **groups of variables**

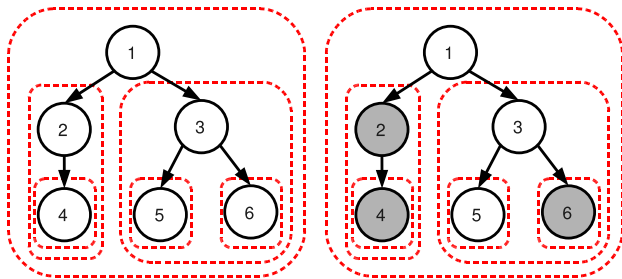
Hierarchical Norms (Zhao et al., 2009; Bach, 2008)



(Jenatton, Mairal, Obozinski and Bach, 2010a)

- Dictionary element selected only after its ancestors
- Structure on codes α (not on individual dictionary elements \mathbf{d}_i)

Hierarchical Norms (Zhao et al., 2009; Bach, 2008)



(Jenatton, Mairal, Obozinski and Bach, 2010a)

- Dictionary element selected only after its ancestors
- Structure on codes α (not on individual dictionary elements \mathbf{d}_i)
- Hierarchical penalization: $\Omega(\alpha) = \sum_{g \in \mathcal{G}} \|\alpha_g\|_2$ where groups g in \mathcal{G} are equal to **set of descendants** of some nodes in a tree

Hierarchical Dictionary Learning

Efficient Optimization

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

$$\begin{aligned} \min_{\mathbf{D}, \mathbf{A}} \quad & \sum_{i=1}^M \|\mathbf{x}^{(i)} - \mathbf{D}\boldsymbol{\alpha}^{(i)}\|_2^2 + \lambda \sum_{i=1}^M \Omega(\boldsymbol{\alpha}^{(i)}) \\ \text{s.t.} \quad & \boldsymbol{\alpha}^{(i)} \in \mathbb{R}_+^K, \\ & \mathbf{d}^{(k)} \in \mathbb{R}_+^p, \quad \mathbf{d}^\top \mathbf{1} = 1. \end{aligned}$$

- Can we solve these efficiently?

Hierarchical Dictionary Learning

Efficient Optimization

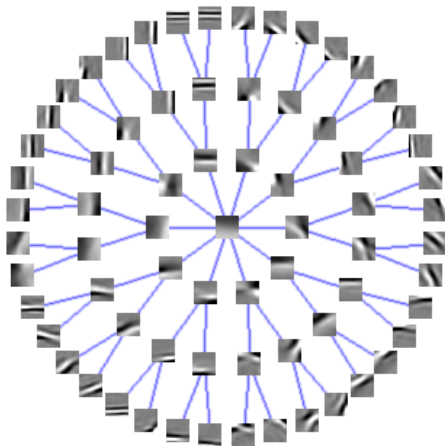
$$\min_{\substack{\mathbf{A} \in \mathbb{R}^{K \times M} \\ \mathbf{D} \in \mathbb{R}^{p \times K}}} \sum_{i=1}^M \|\mathbf{x}^{(i)} - \mathbf{D}\boldsymbol{\alpha}^{(i)}\|_2^2 + \lambda \Omega(\boldsymbol{\alpha}^{(i)}) \quad \text{s.t.} \quad \forall k, \|\mathbf{d}^{(k)}\|_2 \leq 1.$$

$$\begin{aligned} \min_{\mathbf{D}, \mathbf{A}} \quad & \sum_{i=1}^M \|\mathbf{x}^{(i)} - \mathbf{D}\boldsymbol{\alpha}^{(i)}\|_2^2 + \lambda \sum_{i=1}^M \Omega(\boldsymbol{\alpha}^{(i)}) \\ \text{s.t.} \quad & \boldsymbol{\alpha}^{(i)} \in \mathbb{R}_+^K, \\ & \mathbf{d}^{(k)} \in \mathbb{R}_+^p, \quad \mathbf{d}^\top \mathbf{1} = 1. \end{aligned}$$

- Can we solve these efficiently?

→ Proximal methods

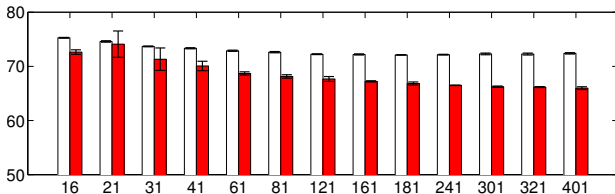
Hierarchical dictionary for image patches



Application to inpainting

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

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



Hierarchical Topic Models for text corpora

Flat Topic Model

Each document $\mathbf{x}^{(i)}$ is modeled through word counts:

x_{ij} = nb of occurrences of word j in document i , $\mathbf{1}^\top \mathbf{x}^{(i)} = N_i$,

θ =topic proportions, \mathbf{B} =topic word frequencies

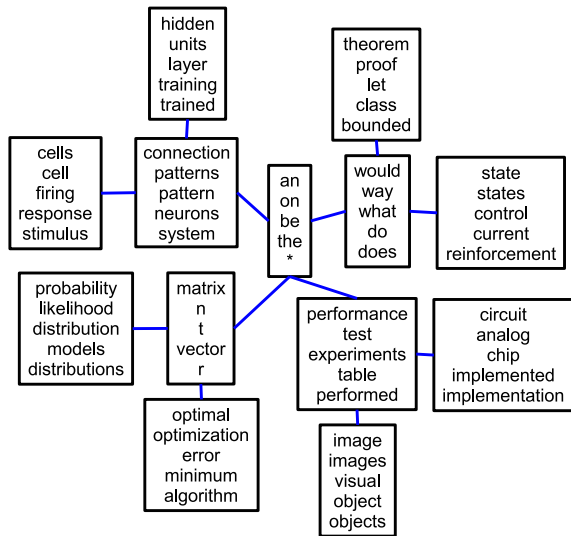
Model x_i as. $x_i \sim \mathcal{M}(\mathbf{B}\theta, N_i)$

- 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?

Tree of Topics



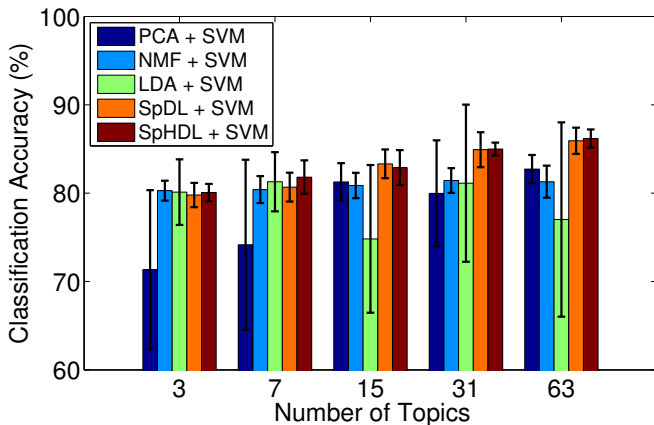
NIPS abstracts

- 1714 documents
- 8274 words

Classification based on topics

Comparison on predicting newsgroup article subjects

- 20 newsgroup articles (1425 documents, 13312 words)



First-order/proximal methods

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

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

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

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

First-order/proximal methods

- They require solving efficiently the proximal operator

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

- For the ℓ_1 -norm, this reduces to *soft-thresholding*:

$$\mathbf{w}_i^* = (\mathbf{u}_i - \lambda)_+ \text{sign}(\mathbf{u}_i).$$

- For the ℓ_1/ℓ_2 with **disjoint** groups, this reduces to *group-soft-thresholding*

$$\mathbf{w}_g^* = (\|\mathbf{u}_g\| - \lambda)_+ \frac{\mathbf{u}_g}{\|\mathbf{u}_g\|_2}$$

- There exist accelerated versions based on Nesterov optimal first-order method (gradient method with “extrapolation”) (Beck and Teboulle, 2009; Nesterov, 2007)
- suited for large-scale experiments.

Tree-structured groups

Proposition (Jenatton et al., 2011)

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

$$g \cap h = \emptyset \quad \text{or} \quad g \subset h \quad \text{or} \quad h \subset g$$

- For $q = 2$ or $q = \infty$, we define Prox_g and Prox_Ω as

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

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

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

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

→ *Tree-structured regularization* : Efficient **linear time** algorithm.

SPAMS: SPArse Modeling Software

SPAMS (SPArse Modeling Software) is an optimization toolbox for solving various sparse estimation problems.

- **Dictionary learning** and **matrix factorization**
- Solving **sparse decomposition problems**
- Solving **structured sparse decomposition problems**

<http://www.di.ens.fr/willow/SPAMS/>

Conclusions: Theory of Graphical Models

- Graphical models provide a nice and precise framework to construct and think about models of data.
- Can be used with frequentists **estimation** techniques
 - Maximum Likelihood Techniques
 - Expectation-Maximization algorithm
- Can be used with Bayesian **estimation** techniques
 - Computing posterior distribution over parameters, or computing posterior expectations
- In both cases, one needs to compute expectations (unless the data is completely observed). This is called the **inference problem**.
- Many **inference** algorithms:
 - Exact algorithms
 - Sum-product/ Belief propagation
 - Junction tree algorithm
 - Approximate algorithms
 - Gibbs sampling
 - Variational Inference (Mean field, loopy belief propagation)

Conclusions: PGM for IR...

- Some nice models (UM, pLSI, LDA)
- Still need more understanding
- Parallel approaches with matrix factorization and dictionary learning
- Still many structures in IR that could be modelled with PGMs and ML...

References I

- Bach, F. (2008). Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*.
- Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.
- 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)*.
- 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.
- 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).
- Hoffman, M., Blei, D., and Bach, F. (2010). Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 23:856–864.
- Jenatton, R., Mairal, J., Obozinski, G., Bach, F., et al. (2011). Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334.

References II

- Lee, H., Battle, A., Raina, R., and Ng, A. (2007). Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems (NIPS)*.
- Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2009a). Online learning for matrix factorization and sparse coding. Technical report, arXiv:0908.0050.
- 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)*.
- Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. *Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Tech. Rep*, 76.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of The Royal Statistical Society Series B*, 58(1):267–288.
- 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. (2007). On the non-negative garrotte estimator. *Journal of The Royal Statistical Society Series B*, 69(2):143–161.
- Zhao, P., Rocha, G., and Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497.