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 → *Dictionary Learning*

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

- Bayesian approach: prior + integrate → Latent Dirichlet Allocation

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

- "Frequentist + Bayesian" → integrate + optimize

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

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

From regularized pLSI (multinomial PCA) to ... "dictionary learning"

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

From regularized pLSI (multinomial PCA) to ... "dictionary learning"

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

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

From regularized pLSI (multinomial PCA) to ... "dictionary learning"

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

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

How to find the best **B** in this formulation?

From regularized pLSI (multinomial PCA) to ... "dictionary learning"

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

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

How to find the best **B** in this formulation?

$$\sum_{i=1}^{M} \min_{\boldsymbol{\theta}_i} \left[\frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i)\right]$$

$$\text{s.t.} \qquad \forall(i,k), \qquad \theta_{ik} \geq 0,$$

From regularized pLSI (multinomial PCA) to ... "dictionary learning"

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

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

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

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

From regularized pLSI (multinomial PCA) to ... "dictionary learning"

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

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

How to find the best **B** in this formulation?

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

A link to LSI?...

$$\min_{\mathbf{B}} \quad \sum_{i=1}^{M} \min_{\boldsymbol{\theta}_i} \Big[\frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i) \Big]$$

$$\text{s.t.} \quad \forall(i,k), \quad \theta_{ik} \geq 0, \quad \forall(i,j),\ \mathbf{B}_{ji} \geq 0, \quad \forall i,\ \sum_{j=1}^{d} \mathbf{B}_{ji} = 1$$

A link to LSI?...

$$\min_{\mathbf{B}} \min_{\boldsymbol{\Theta}=\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M} \sum_{i=1}^{M} \left[\frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i) \right]$$

$$\text{s.t.} \quad \forall (i,k), \quad \theta_{ik} \geq 0, \quad \forall (i,j),\ \mathbf{B}_{ji} \geq 0, \quad \forall i,\ \sum_{j=1}^{d} \mathbf{B}_{ji} = 1$$

Rewriting as matrix factorization problem:

A link to LSI?...

$$\min_{\mathbf{B}} \min_{\boldsymbol{\Theta}=\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M} \sum_{i=1}^{M} \left[\frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i) \right]$$

$$\text{s.t.} \quad \forall(i,k), \quad \theta_{ik} \geq 0, \quad \forall(i,j),\ \mathbf{B}_{ji} \geq 0, \quad \forall i,\ \sum_{j=1}^{d} \mathbf{B}_{ji} = 1$$

Rewriting as matrix factorization problem:

$$\min_{\mathbf{B}} \min_{\boldsymbol{\Theta}=\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M} \frac{1}{2}\|\mathbf{X} - \mathbf{B}\boldsymbol{\Theta}\|_F^2 + \lambda \sum_{i=1}^{M} \Omega(\boldsymbol{\theta}_i)$$

$$\text{s.t.} \quad \forall(i,k), \quad \theta_{ik} \geq 0, \quad \forall(i,j),\ \mathbf{B}_{ji} \geq 0, \quad \forall i,\ \sum_{j=1}^{d} \mathbf{B}_{ji} = 1$$

What happens if we remove the constraints and regularization?

A link to LSI?...

$$\min_{\mathbf{B}} \min_{\mathbf{\Theta}=\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M} \sum_{i=1}^{M} \left[\frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i) \right]$$

$$\text{s.t.} \quad \forall(i,k), \quad \theta_{ik} \geq 0, \quad \forall(i,j),\ \mathbf{B}_{ji} \geq 0, \quad \forall i,\ \sum_{j=1}^{d} \mathbf{B}_{ji} = 1$$

Rewriting as matrix factorization problem:

$$\min_{\mathbf{B}} \min_{\mathbf{\Theta}=\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M} \frac{1}{2}\|\mathbf{X} - \mathbf{B}\mathbf{\Theta}\|_F^2$$

What happens if we remove the constraints and regularization?

A link to LSI?...

$$\min_{\mathbf{B}} \min_{\boldsymbol{\Theta}=\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M} \sum_{i=1}^{M} \left[\frac{1}{2}\|\mathbf{x}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda\Omega(\boldsymbol{\theta}_i) \right]$$

$$\text{s.t.} \quad \forall(i,k), \quad \theta_{ik} \geq 0, \quad \forall(i,j),\ \mathbf{B}_{ji} \geq 0, \quad \forall i,\ \sum_{j=1}^{d} \mathbf{B}_{ji} = 1$$

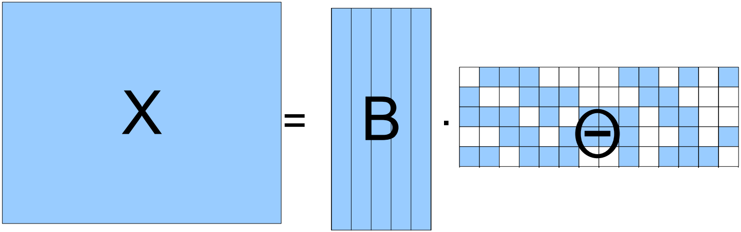
Rewriting as matrix factorization problem:

$$\min_{\mathbf{B}} \min_{\boldsymbol{\Theta}=\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M} \frac{1}{2}\|\mathbf{X} - \mathbf{B}\boldsymbol{\Theta}\|_F^2$$

What happens if we remove the constraints and regularization?
We get back **LSI**: $B = U_K$ and $\boldsymbol{\theta}_i = \tilde{x}_i$

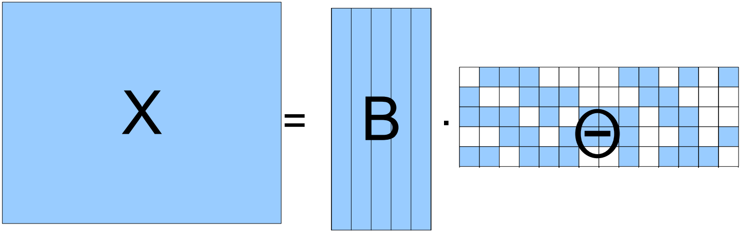
Topic models and matrix factorization

- $\mathbf{X} \in \mathbb{R}^{d \times M}$ 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 $\boldsymbol{\theta}_i$ associated each to one document and which encodes its "topic content".



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 $\boldsymbol{\theta}_i$ associated each to one document and which encodes its "topic content".



How about sparsity in topics?...

Ridge, penalization and sparsity

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

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

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

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

Ridge, penalization and sparsity

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

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

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

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 $\quad \|\theta\|_0 = |\{k \mid \theta_k \neq 0\}|$

Ridge, penalization and sparsity

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

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

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

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

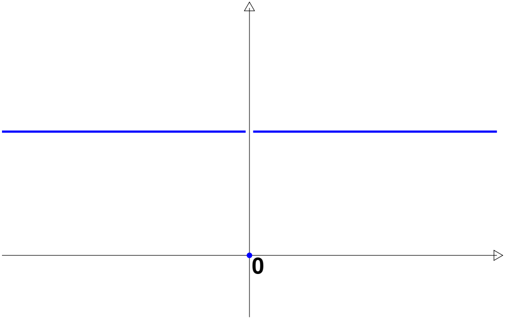
Can we choose Ω to obtain a sparse decomposition?

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

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

Relaxing the ℓ_0 penalization

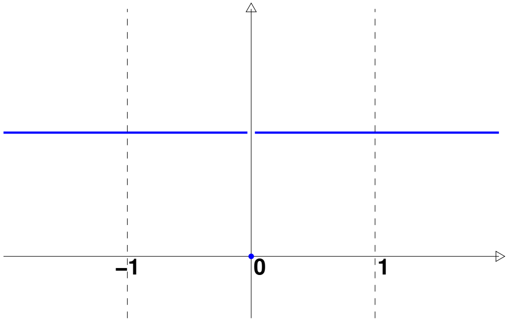
$$\|\theta\|_0 = \sum_{k=1}^{K} \mathbf{1}_{\{\theta_k \neq 0\}}$$



Relaxing the ℓ_0 penalization

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

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

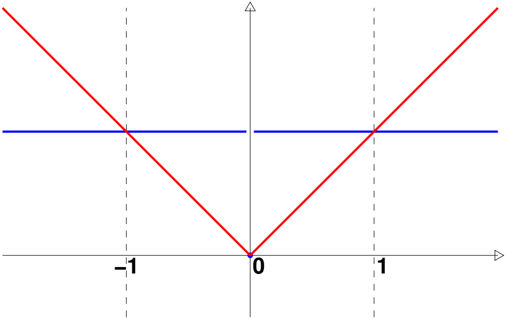


Relaxing the ℓ_0 penalization

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

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

Relax



Relaxing the ℓ_0 penalization

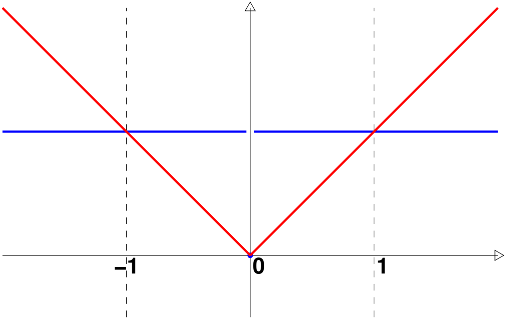
$$\|\theta\|_0 = \sum_{k=1}^{K} \mathbf{1}_{\{\theta_k \neq 0\}}$$

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

Relax

We obtain the ℓ_1-norm:

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



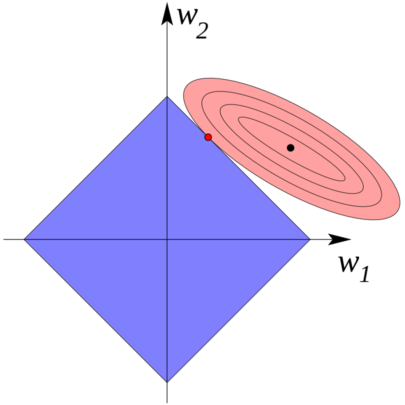
The LASSO (Tibshirani, 1996)

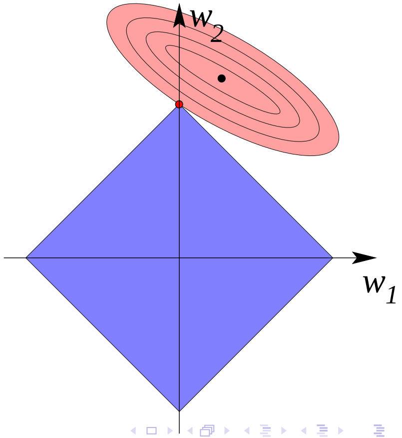
LASSO: Least Absolute Shrinkage and Selection operator

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

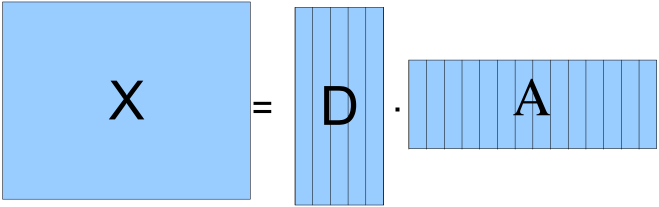
Why ℓ_1-norm constraints leads to sparsity?

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





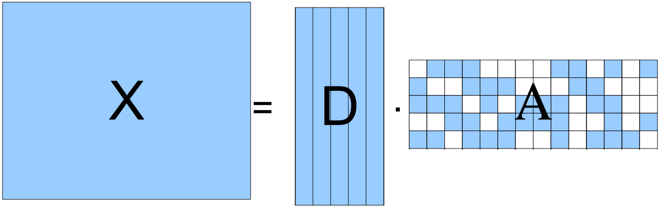
Decomposition of signals on a dictionary



- dictionary $\mathbf{D} = (\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(K)})$ 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 D and/or α**
  - Low-rank: $\mathbf{D}$ and $A^\top$ have few columns
  - Dictionary learning / sparse PCA: $\mathbf{D}$ or $\mathbf{A}$ has many zeros
  - Clustering (k-means): $\mathbf{A} \in \{0,1\}^{K \times M}$, $\mathbf{A1} = \mathbf{1}$
  - 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

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:

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

- Problem is convex in **D** and **A** separately, but not jointly.

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

Focus on previous formulation:

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

- Problem is convex in **D** and **A** separately, but not jointly.

  $\rightarrow$ Alternating scheme: optimize **D** and **A** in turn.

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

Focus on previous formulation:

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

- Problem is convex in **D** and **A** separately, but not jointly.
  $\rightarrow$ Alternating scheme: optimize **D** and **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:

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

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

$$\text{With} \quad \widetilde{\mathbf{X}} \;=\; \mathbf{X} - \sum_{j' \neq j} \mathbf{d}^{(k)} \boldsymbol{\alpha}_k,$$

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

Focus on previous formulation:

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

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

$$\text{With} \quad \widetilde{\mathbf{X}} = \mathbf{X} - \sum_{j' \neq j} \mathbf{d}^{(k)} \boldsymbol{\alpha}_k, \quad \text{we have} \quad \mathbf{d}^{(k)} \leftarrow \frac{\widetilde{\mathbf{X}} \boldsymbol{\alpha}_k^\top}{\|\widetilde{\mathbf{X}} \boldsymbol{\alpha}_k^\top\|}$$

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

Focus on previous formulation:

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

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

$$\text{With} \quad \widetilde{\mathbf{X}} = \mathbf{X} - \sum_{j' \neq j} \mathbf{d}^{(k)} \boldsymbol{\alpha}_k, \quad \text{we have} \quad \mathbf{d}^{(k)} \leftarrow \frac{\widetilde{\mathbf{X}} \boldsymbol{\alpha}_k^\top}{\|\widetilde{\mathbf{X}} \boldsymbol{\alpha}_k^\top\|}$$

$$\text{and} \quad \boldsymbol{\alpha}_k^\top \leftarrow \text{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^M} \|\mathbf{X}^\top \mathbf{d}^{(k)} - \boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1$$

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

Focus on previous formulation:

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

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

$$\text{With} \quad \widetilde{\mathbf{X}} = \mathbf{X} - \sum_{j' \neq j} \mathbf{d}^{(k)} \boldsymbol{\alpha}_k, \quad \text{we have} \quad \mathbf{d}^{(k)} \leftarrow \frac{\widetilde{\mathbf{X}} \boldsymbol{\alpha}_k^\top}{\|\widetilde{\mathbf{X}} \boldsymbol{\alpha}_k^\top\|}$$

$$\text{and} \quad \boldsymbol{\alpha}_k^\top \leftarrow \text{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^M} \|\mathbf{X}^\top \mathbf{d}^{(k)} - \boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1$$

  - 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!

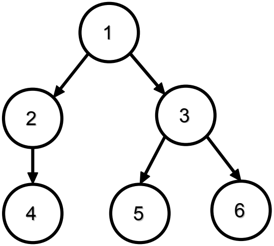
Another common choice for Ω:

- 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

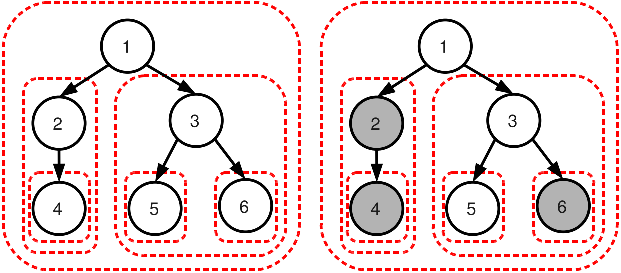
Hierarchical Norms (Zhao et al., 2009; Bach, 2008)



(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 $\boldsymbol{\alpha}$ (not on individual dictionary elements $\mathbf{d}_i$)
- 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

Hierarchical Dictionary Learning

Efficient Optimization

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

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

- Can we solve these efficiently?

Hierarchical Dictionary Learning

Efficient Optimization

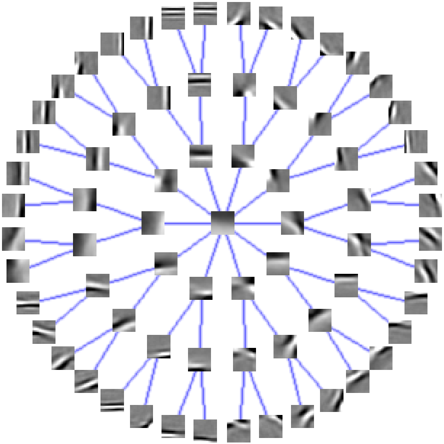
$$\min_{\substack{\mathbf{A}\in\mathbb{R}^{K\times M}\\ \mathbf{D}\in\mathbb{R}^{p\times K}}} \sum_{i=1}^{M} \|\mathbf{x}^{(i)} - \mathbf{D}\boldsymbol{\alpha}^{(i)}\|_2^2 + \lambda\Omega(\boldsymbol{\alpha}^{(i)}) \text{ s.t. } \forall k,\ \|\mathbf{d}^{(k)}\|_2 \leq 1.$$

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

- Can we solve these efficiently?

$\rightarrow$ Proximal methods

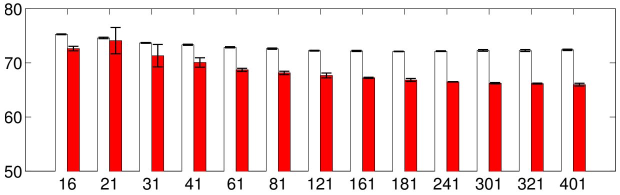
Hierarchical dictionary for image patches



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, $\quad \mathbf{1}^\top \mathbf{x}^{(i)} = N_i$,
$\boldsymbol{\theta}$=topic proportions, $\quad$ $\mathbf{B}$=topic word frequencies

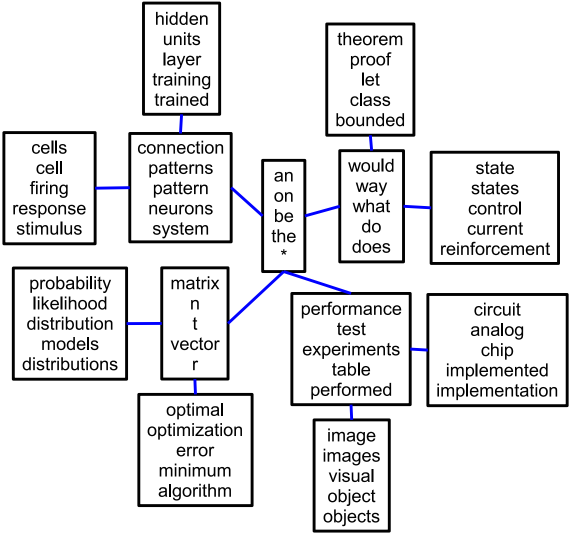
$$\text{Model} \quad x_i \quad \text{as.} \quad x_i \sim \mathcal{M}(\mathbf{B}\boldsymbol{\theta}, N_i)$$

- 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



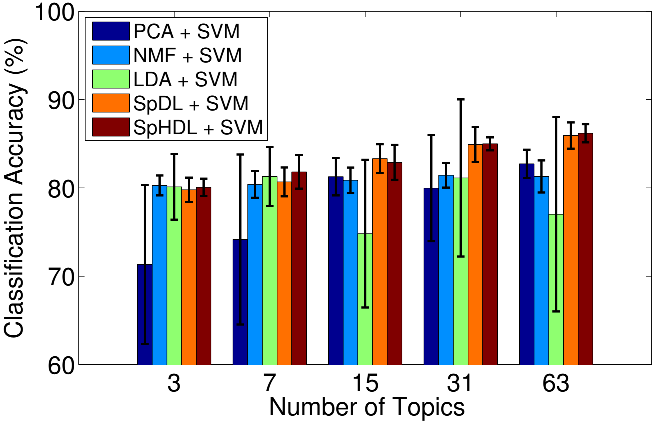
NIPS abstracts

- 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

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

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

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

First-order/proximal methods

- They require solving efficiently the proximal operator

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

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

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

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

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

- 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., $\forall g, h \in \mathcal{G}$,

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

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

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

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

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

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

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

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.