Machine/Statistical Learning

Lecture 6: Dictionnary Learning

Filière SICOM, 3A

Reminder of regression and sparsity

- $\boldsymbol{x} \in \mathbb{R}^p \leftarrow$ observations,
- $D \in \mathbb{R}^{p \times K} \leftarrow$ design/regression matrix,

Regression problem

Find coefficient vector $\boldsymbol{a} \in \mathbb{R}^K$ s.t. $\quad \boldsymbol{x} \approx D\boldsymbol{a}, \quad$ i.e. $\|\boldsymbol{x} - D\boldsymbol{a}\| \leq \epsilon$

☞ regularization/model selection : sparsity constraint

$$\min_{\boldsymbol{a}} \|\boldsymbol{x} - D\boldsymbol{a}\|_2^2 \quad \text{s.t. } \mathrm{Pen}(\boldsymbol{a}) \leq T, \quad \text{with e.g.}$$

- $\mathrm{Pen}(\boldsymbol{a}) = \|\boldsymbol{a}\|_1 \leftarrow \ell_1$ norm,
- $\mathrm{Pen}(\boldsymbol{a}) = \|\boldsymbol{a}\|_0 \equiv \#\{i \; : \; 1 \leq i \leq K \text{ and } \boldsymbol{a}_i \neq 0\} \leftarrow \ell_0$ *pseudo-norm*, i.e. number of non-zero components for $\boldsymbol{a}$

Dictionary problem

Assumptions

- Inputs : $\boldsymbol{x}_i \in \mathbb{R}^p$, for $i = 1, \ldots, n$, $\leftarrow$ learning set

Objective

"Inverse" the regression problem, i.e. find $D \in \mathbb{R}^{p \times K}$ ensuring a sparse approximation/representation of the learning set $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$

$$\boldsymbol{x}_i \approx D\boldsymbol{a}_i, \quad \text{for } i = 1, \ldots, n, \quad \text{s.t.}$$

- $K \ll n$ $\leftarrow$ sparsity of the representation 'basis' ($\sim$ low rank constraint)
- $\|\boldsymbol{a}_i\|_0 \leq T$, for $i = 1, \ldots, n$ $\leftarrow$ sparsity of the coefficients

Dictionary terminology

- $D \equiv \left(\boldsymbol{d}_1 | \quad \ldots \quad | \boldsymbol{d}_K\right) \in \mathbb{R}^{p \times K}$ $\leftarrow$ dictionary s.t. $K \ll n$ (low rank)
- $\boldsymbol{d}_k \in \mathbb{R}^p$ $\leftarrow$ atoms with $k = 1, \ldots, K$ (column vectors of D)
- $\boldsymbol{a}_i \in \mathbb{R}^K$ $\leftarrow$ coefficients of $\boldsymbol{x}_i$ s.t. $||\boldsymbol{a}_i||_0 \leq T$ (sparsity)

Dictionary vs Orthornormal Basis (ONB)

Orthornormal Bases (ONB)

Lots of good properties for ONB (e.g. Fourier, orthonormal wavelets,...) :

- uniqueness of the decomposition (n coefficients to represent a n-dimensional vector),
- existence of fast transforms,
- projections to obtain approximate representations, ...

But not necessarily good for sparse representations

☞ e.g. sums of sinusoids + spikes (+ steps) are not sparse in either Fourier or identity bases...

Dictionary

Change of mindset

- loose uniqueness of representation using overcomplete/redundant bases

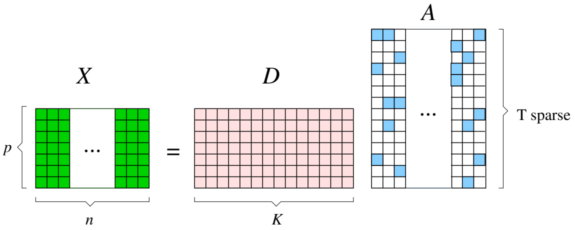
☞ get sparser representations in return !

Recap : Dictionary and Sparse representations

Hypotheses

- Inputs : $\boldsymbol{x}_i \in \mathbb{R}^p$, $1 \leq i \leq n$, $\quad \rightarrow X = \left(\boldsymbol{x}_1 | \ \dots \ | \boldsymbol{x}_n\right) \in \mathbb{R}^{p \times n}$

Objective : sparse representation/factorization $X \approx DA$



- $D \equiv \left(\boldsymbol{d}_1 | \dots | \boldsymbol{d}_K\right) \in \mathbb{R}^{p \times K} \leftarrow$ dictionary (redundant), with $K \ll n$.
- $\boldsymbol{d}_k \in \mathbb{R}^p \leftarrow$ atoms with $k = 1, \dots, K$
- $A \equiv \left(\boldsymbol{a}_1 | \dots | \boldsymbol{a}_n\right) \in \mathbb{R}^{K \times n} \leftarrow$ coefficients s.t. $\boldsymbol{x}_i \approx D\boldsymbol{a}_i$ with $||\boldsymbol{a}_i||_0 \leq T$

☞ only the observations $\boldsymbol{x}_i$ are known : unsupervised problem

Formalizing the dictionary learning problem

Cost criterion with sparsity constraints

$$\min_{D,\{\boldsymbol{a}_i\}} J\left(D,\{\boldsymbol{a}_i\}\right) \equiv \sum_{i=1}^{n} \|\boldsymbol{x}_i - D\boldsymbol{a}_i\|_2^2 = \|X - DA\|_F^2,$$
$$\text{s.t. } \mathrm{Pen}(\boldsymbol{a}_i) \leq T \quad \text{for } i = 1, \dots, n$$

- $X = \big(\boldsymbol{x}_1 | \ \dots \ | \boldsymbol{x}_n\big) \in \mathbb{R}^{p \times n}$, $\quad D = \big(\boldsymbol{d}_1 | \ \dots \ | \boldsymbol{d}_K\big) \in \mathbb{R}^{p \times K}$, $A = \big(\boldsymbol{a}_1 | \ \dots \ | \boldsymbol{a}_n\big) \in \mathbb{R}^{K \times n}$,
- $\|M\|_F^2 \equiv \sum_{i,j} m_{ij}^2 \leftarrow$ Frobenius norm

Identifiability issue

Pb : atoms $\boldsymbol{d}_k$ and their associate coefficients $\boldsymbol{a}^k \in \mathbb{R}^{1 \times n}$ (kth line of $\boldsymbol{a}$) are defined up to a scaling factor in the factorization criterion

☞ atoms are assumed to be normalized, i.e. $\|\boldsymbol{d}_k\|_2 = 1$, for $k = 1, \dots, K$

Optimizing the factorization criterion

Alternate minimization method

$$\min_{D,\{\boldsymbol{a}_i\}} J\left(D,\{\boldsymbol{a}_i\}\right) = \|X - DA\|_F^2, \quad \text{s.t. } \mathrm{Pen}(\boldsymbol{a}_i) \leq T$$

Joint minimization w.r.t. A and D split into 2 simpler separated steps

1. sparse coding to estimate the coefficients A for a given dictionary D,
2. dictionary D update performed atom-by-atom, for a given A

☞ repeat 1) and 2) until a stopping criterion

Optimization issues

- when $\mathrm{Pen}(\boldsymbol{a}_i) = \|\boldsymbol{a}_i\|_0 \leftarrow$ NP-hard combinatorics problem...
- joint minimization w.r.t. D and $\{\boldsymbol{a}_i\}_{1 \leq i \leq n} \leftarrow$ non-convex problem

☞ sub-optimal solution around an initial value for D (cf clustering)

Sparse coding : update of the coefficients $\boldsymbol{a}$

Assumption : D fixed

Objective : find sparse $A^\star$ minimizing $J(A) \equiv J(D, A) = \|X - DA\|_F^2$

☞ linear regression problem on each $\boldsymbol{a}_i$ with sparsity constraints

Basis Pursuit (BP)

Convex relaxation of the ℓ_0 pseudo-norm with the ℓ_1 norm :

$$A^\star = \min_A \|X - DA\|_F^2 + \lambda \sum_{i=1}^{n} \|\boldsymbol{a}_i\|, \quad \text{where } \lambda > 0,$$

- convex problem on A, with Lasso solutions for $\boldsymbol{a}_i$, $i = 1, \ldots, n$

☞ Online Dictionary Learning (ODL) algorithm (Mairal *et al.*, 2009)

(Orthogonal) Matching Pursuit (OMP)

Sparse approximate (sub-optimal) solution of the ℓ_0 minimization problem using a greedy algorithm to select the "best matching" atoms

Orthogonal Matching Pursuit

Let D_Γ be the sub dictionary composed of the only atoms $\boldsymbol{d}_k$ s.t. $k \in \Gamma$

1. **Initialization :** $\Gamma = \emptyset$, $\boldsymbol{r} = \boldsymbol{x}_i$ (sweep on the observations $\boldsymbol{x}_i$)
2. **for** $iter = 1, \ldots, T$ **do**
3. Select the best matching atom, i.e. most correlated with the residuals

$$\widehat{k} \leftarrow \max_{1 \leq k \leq K} |\boldsymbol{r}^T \boldsymbol{d}_k|$$

4. Update the active set : $\Gamma \leftarrow \Gamma \cup \{\widehat{k}\}$
5. Update the residuals : $\boldsymbol{r} \leftarrow \left(I_p - D_\Gamma (D_\Gamma^T D_\Gamma)^{-1} D_\Gamma^T\right) \boldsymbol{x}_i,$
6. **endfor**
7. **Output :** $\boldsymbol{a}_i \equiv$ coefficients of the orthogonal projection of $\boldsymbol{x}_i$ on the space spanned by D_Γ

☞ residuals orthogonal to the sparse signal approximation $D_\Gamma \boldsymbol{a}_i$

☞ by construction, card $(\Gamma) = T \Rightarrow \|\boldsymbol{a}_i\|_0 \leq T$

K-SVD algorithm outline

1. preliminary tools : SVD and low rank approximation
2. K-SVD dictionary update and K-SVD algorithm

Singular Value Decomposition (SVD)

Let $X \in \mathbb{R}^{n\times m}$ be a real rectangular matrix. There exists a factorization, called a singular value decomposition of X, of the form

$$X = U\Sigma V^T, \quad \text{where}$$

- $U \in \mathbb{R}^{n\times n}$ is orthonormal $(UU^T = U^T U = I_n) \leftarrow$ matrix of left-singular vectors $\boldsymbol{u}_k \in \mathbb{R}^n$ s.t. $U = (\boldsymbol{u}_1|\ldots|\boldsymbol{u}_n)$,
- $V \in \mathbb{R}^{m\times m}$ is orthonormal $(VV^T = V^T V = I_m) \leftarrow$ matrix of right-singular vectors $\boldsymbol{v}_k \in \mathbb{R}^m$ s.t. $V = (\boldsymbol{v}_1|\ldots|\boldsymbol{v}_m)$,
- $\Sigma \in \mathbb{R}^{n\times m}$ is a rectangular diagonal matrix with non negative diagonal entries $\Sigma_{ii} \equiv \lambda_i \geq 0$ for $i = 1,\ldots,\min(n,m)$

☞ the λ_i's are uniquely defined and called the singular values of X

☞ By convention, these singular values are sorted in descending order :

$$\lambda_1 \geq \ldots \geq \lambda_{\min(n,m)} \geq 0$$

Singular value decomposition (SVD) and principal components analysis (PCA)

$X \in \mathbb{R}^{n\times m}$ with SVD $X = U\Sigma V^T$

Eigendecomposition of Gram matrices

Gram matrices express as

$$\begin{aligned} X^T X &= V\Sigma^T U^T U \Sigma V^T &= V\left(\Sigma^T\Sigma\right)V^T, \\ XX^T &= U\Sigma V^T V \Sigma^T U^T &= U\left(\Sigma\Sigma^T\right)U^T, \end{aligned}$$

where $\Sigma^T\Sigma \in \mathbb{R}^{m\times m}$ and $\Sigma\Sigma^T \in \mathbb{R}^{n\times n}$ are square diagonal

- right-singular vectors of X are eigenvectors of $X^T X = V\left(\Sigma^T\Sigma\right)V^T$,
- left-singular vectors of X are eigenvectors of $XX^T = U\left(\Sigma\Sigma^T\right)U^T$,
- non-zero singular values λ_i's are the square roots of the non-zero eigenvalues of $X^T X$ or XX^T

☞ SVD yields principal components analysis (PCA) decomposition

Singular Value Decomposition (SVD) and low rank approximation

- $X \in \mathbb{R}^{n\times m}$ with SVD $X = U\Sigma V^T$
- ☞ left-singular vectors $\boldsymbol{u}_k \in \mathbb{R}^n$ of X are the columns of U, for $k = 1, \dots, n$
- ☞ right-singular vectors $\boldsymbol{v}_k \in \mathbb{R}^m$ of X are the columns of V, for $k = 1, \dots, m$

Eckart–Young–Mirsky theorem

For the Frobenius norm ($\|M\|_F^2 = \sum_{i,j} m_{i,j}^2$), the solution to the low-rank approximation problem $\min_{\hat{X}} \|X - \hat{X}\|_F^2 \quad \text{s.t. } \operatorname{rank}(\hat{X}) \le r$, is

$$\hat{X} = \sum_{k=1}^{r} \lambda_k \boldsymbol{u}_k \boldsymbol{v}_k^T,$$

where λ_k, $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ are the first singular values and left/right-singular vectors of X, for $k = 1, \dots, r \le \min(m, n)$.

Dictionary update : K-SVD algorithm (Aharon *et al.*, 2006)

Objective : For a given A, find D minimizing $J(D) \equiv J(D, A) = \|X - DA\|_F^2$

Sweep on the atoms $\boldsymbol{d}_k$ that are sequentially updated, for $k = 1, \ldots K$, as

$$\min_{\boldsymbol{d}_k} \|X - DA\|_F^2 = \|E^k - \boldsymbol{d}_k \boldsymbol{a}^k\|_F^2$$

- $\boldsymbol{a}^k \in \mathbb{R}^{1\times n}$ is the kth line of A,
- $E^k = X - \sum_{l\neq k} \boldsymbol{d}_l \boldsymbol{a}^l \in \mathbb{R}^{p\times n}$,
- $\boldsymbol{d}_k \boldsymbol{a}^k \in \mathbb{R}^{p\times n}$ is a rank 1 matrix

Atoms (and coefficients) update

Best rank 1 approximation of E^k obtained from its SVD as $\lambda_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T$,

- $\lambda_1 \equiv$ largest singular value, $\boldsymbol{u}_1$ and $\boldsymbol{v}_1^T \equiv$ associated singular vectors

☞ $\boldsymbol{d}_k = \boldsymbol{u}_1 \leftarrow$ unit norm vector by construction

☞ $\boldsymbol{a}^k = \lambda_1 \boldsymbol{v}_1^T \leftarrow$ Pb : there is no reason for the updated $\boldsymbol{a}^k$ to be sparse !

Dictionary update : K-SVD algorithm (Aharon *et al.*, 2006)

K-SVD solution to enforce sparsity of the coefficients

For updating the kth atom restrict attention to the only observations $\boldsymbol{x}_i$ for which the atom is active, i.e. data and dictionary columns with indexes in

$$\Gamma_k = \{i \;:\; \alpha_i^k \neq 0\}$$

- $\boldsymbol{a}_{\Gamma_k}^k \in \mathbb{R}^{|\Gamma_k|}$ is the vector of the only non zero components of $\boldsymbol{a}^k \in \mathbb{R}^n$
- $E_{\Gamma_k}^k \in \mathbb{R}^{n \times |\Gamma_k|}$ is the matrix obtained by retaining the only column vectors of $E^k = X - \sum_{l \neq k} \boldsymbol{d}_l \boldsymbol{a}^l$ with column index $i \in \Gamma_k$
- $\lambda_1 \equiv$ largest singular value, $\boldsymbol{u}_1$ and $\boldsymbol{v}_1^T \equiv$ associated singular vectors of the SVD of $E_{\Gamma_k}^k$

K-SVD atoms (and coefficients) update

- $\boldsymbol{d}_k \leftarrow \boldsymbol{u}_1$ (unit norm vector by construction)
- $\boldsymbol{a}_{\Gamma_k}^k \leftarrow \lambda_1 \boldsymbol{v}_1{}^T$ (hence $\boldsymbol{a}^k$ remains as sparse as sparse coding outputs)

☞ both atoms and coefficients are updated during the K-SVD dictionary update step

K-SVD algorithm

0. Initialize dictionary $D^{(0)}$ using union of ONBs elements / randomly by picking K observations
1. Sparse coding : computation of the coefficients $\boldsymbol{a}_i$ for each $\boldsymbol{x}_i$ given $D^{(t-1)}$ (usually using OMP with $\|\boldsymbol{a}_i\|_0 \leq T$)
2. Dictionary update : for each atom $\boldsymbol{d}_k \in D^{(t-1)}$, $1 \leq k \leq K$
    - Define the set of active entries $\Gamma_k = \{i \ : \ \alpha_i^k \neq 0\}$
    - Compute $E^k = X - \sum_{l \neq k} \boldsymbol{d}_l \boldsymbol{a}^l$ and its restriction $E^k_{\Gamma_k}$ to columns in Γ_k
    - Compute the first largest singular value $\lambda_1 > 0$ of $E^k_{\Gamma_k}$, and the associate left and right singular vectors $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$
    - Update the k-th atom and the associated coefficients
      $$\begin{aligned} \boldsymbol{d}_k &\leftarrow \boldsymbol{u}_1, \\ \boldsymbol{a}^k_{\Gamma_k} &\leftarrow \lambda_1 \boldsymbol{v}_1^T \end{aligned}$$

- Repeat steps 1 and 2 for $t = 1, \ldots$ until convergence

Properties of K-SVD algorithm

Convergence

Convergence to a local minimum guaranteed if the sparse coding step always decrease the empirical quadratic error $\|X - DA\|_F^2$

- no theoretical guaranties when using approximations like OMP
- in practice, condition empirically satisfied when $T \ll n$

Generalization of K-means

- For $T = 1$ sparsity constraints, normalizing the coefficients $\boldsymbol{a}_i$ rather than the atoms $\boldsymbol{d}_k$, then K-SVD reduces to the K-means algorithm
  - ☞ dictionary atoms are the cluster means/centroids
- In the general case $T \geq 1$, the computation of the mean is replaced by a SVD computation for updating the atoms
  - ☞ "K-SVD" algorithm

Implementation details of K-SVD

Sparse coding

- OMP is preferred for efficiency,
- for denoising problem, more convenient to solve (approximatively) the similar problem

$$\min_{\boldsymbol{a}_i} ||\boldsymbol{a}_i||_0 \quad \text{s.t. } \|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{a}_i\| \leq \epsilon,$$

where ϵ can be tuned from the noise power $\leftarrow$ change of stopping criterion in OMP

Dictionary update heuristics

- Pruning atoms that are not "used" enough,
- Removing atoms too coherent with each other,

☞ pruned or removed atom replaced by the least well-explained observation, i.e. observation $\boldsymbol{x}_l$ where $l = \arg\max_{1 \leq i \leq n} \|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{a}_i\|_2$

Examples of sparse representations for image restoration

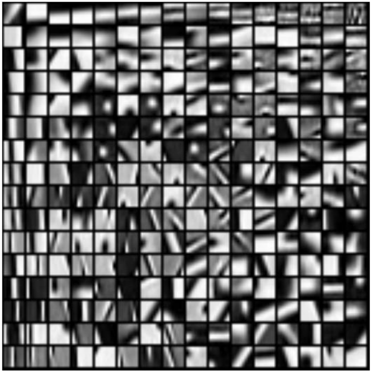
Solving the denoising problem [Elad and Aharon, 2006]

- Extract all overlapping 8×8 patches $\boldsymbol{x}_i$, for $1 \leq i \leq n$ with $n > 10^5$
- Solve a matrix factorization (dictionary learning) problem :

$$\min_{D,\{\boldsymbol{a}_i\}} \quad \sum_{i=1}^{n} ||\boldsymbol{x}_i - D\boldsymbol{a}_i||_2^2 = \min_{D,A} \quad ||X - DA||_F^2$$

  with sparsity constraints $\text{Pen}(\boldsymbol{a}_i) \leq T$, e.g. $\text{Pen}(\boldsymbol{a}_i) = ||\boldsymbol{a}_i||_0$
- Average the reconstruction of each patch.

K-SVD results for image restoration

- Dictionary trained on a noisy version of the image

From ICCV tutorial
`http://lear.inrialpes.fr/people/mairal/tutorial_iccv09/tuto_part2.pdf`

Applications : denoising

[Mairal, Bach, Ponce, Sapiro, and Zisserman, 2009c]

- Dictionary trained on a noisy version of the image

From ICCV tutorial

`http://lear.inrialpes.fr/people/mairal/tutorial_iccv09/tuto_part2.pdf`

Applications : inpainting (1)

Inpainting, [Mairal, Elad, and Sapiro, 2008a]

- Dictionary trained on the image with missing data

From ICCV tutorial http://lear.inrialpes.fr/people/mairal/tutorial_iccv09/tuto_part2.pdf

Applications : inpainting (2)

Inpainting, [Mairal, Elad, and Sapiro, 2008a]

- Dictionary trained on the image with missing data

From ICCV tutorial http://lear.inrialpes.fr/people/mairal/tutorial_iccv09/tuto_part2.pdf