An introduction to dictionary learning

Pierre CHAINAIS

Sept. 17th 2014
From high to low dimension

- Data live in a high dimensional space
- Significant information lives in a low dimensional space

How to capture the "substantifique moelle"?

- dimension reduction? ($K < p$)
- adaptive representation? ($K < p$, $K = p$, or $K > p$)
Motivations

Learning a 'good' dictionary is useful for:

- compression,
- classification or segmentation,
- denoising,
- inpainting,
- blind source separation,
- recommendation systems...

2 main approaches:

- **Parametric dictionaries**: combining existing functions...
- **Non parametric dictionaries**: matrix factorization
Adaptive representation

1. **Mathematical construction:**
   - DCT, Fourier: stationary signals...
   - Time-frequency atoms: non stationary signals
   - Wavelets: stationary + transients + multiscale
   - Curvelets: wavelets + contours...

2. **Statistical learning:**
   - Representative examples: vector quantization (K-means),
   - Orthogonal basis: PCA / SVD (Karhunen-Loève)
   - A family of functions for linear decomposition =

**Dictionary**
Dimension reduction
Feature selection

Select the most representative features (forward, backward...)

↓

Advantage: interpretability

Limitation: use predefined features

Other approach: build/learn a mapping to a new representation
Select the most representative features (forward, backward...)

↓

**Advantage:** interpretability

**Limitation:** use predefined features

Other approach: build/learn a mapping to a new representation
Objective: find representative examples for groups of data points.

11 vowels $\Rightarrow K = 11$ classes described by $D = 10$ features
(cf. time frequency analysis)

Classification based on $M=2$ Fisher discriminant components
Vectorial quantization of color images

Compression

1 pixel = $x_n = (x_R, x_V, x_B) \implies$ clustering 3d using $K$ colors only

$K = 2$

$K = 3$

$K = 10$

Original image
Principal Component Analysis (PCA)

How spread are data points?
Principal Component Analysis (PCA)

How spread are data points?
Principal Component Analysis (PCA)
Example: face recognition

- $D =$ number of pixels per image, for instance $19 \times 19 = 361$,
- $\mathbf{x}_n \in \mathbb{R}^D$ is an image of a face,
- $\mathbf{x}_{ni} =$ intensity of the $i$-th pixel of image $n$,

$$
\mathbf{X}_{D \times N}^T \simeq \mathbf{U}_{D \times M} \times \mathbf{Z}_{M \times N}
$$

Interest:
- extraction of general characteristics,
- use the $\mathbf{z}_j$ for classification (K-NN,...),
- reduction of dimension $\Rightarrow$ speed

[Turk & Pentland 1991]
Inverse problems in image processing

Denoising
Inverse problems in image processing

Denoising
Inverse problems in image processing

Deconvolution
Inverse problems in image processing

Deconvolution
Inverse problems & dictionary learning

Main problem:
ill posed inverse problems: unknown complexity / structure

Main purpose:
- to propose a suitable model (structure)
- to discover the number of degrees of freedom (complexity)

Main tools to promote sparsity:
- a wide range of penalized optimization formulations ($L^1\ldots$)
- model selection deals with discovering complexity

Main interest:
- efficient tools and algorithms in optimization,
- different approaches to promote (structured) sparsity
- theorems to control convergence properties...
ICA, the subspace vectors found by ICA in the example of causes is equivalent to the signal dimension in the standard to higher order statistics. However, since the maximal number of causes lies within different subspaces. Other methods such as sensors observe simultaneously two or more processes with second order correlations in a data set, but also with respect because it is able to separate sources not only with respect to the pixel space is 16, while the number of causes is 20 (total image processing and dictionary learning)

An intuitive way to approach this dimensionality reduction is to look at what generates the dimensionality of the signal. Therefore, they cannot find the underlying 20 letters. Sparse coding learns an overcomplete dictionary of atoms or subspaces that provide efficient representations of classes of signals. Sparsity constraints are keys to most of the algorithms that solve the dictionary learning problem.

An important question arises here: given the observed data, how to determine the subspaces where the data lie? The choice of these subspaces is crucial for efficient dimensionality reduction. Measurements. By identifying these few subspaces, we find the large, only few ones will contain data samples from sensor words, although the number of representation subspaces is much smaller: the observed sensors to observe only a limited number of processes? Why do we need to respect orthogonality constraints in the data representation in the reduced space.

The obvious question is: Why should we constrain our sensory world, although the number of possible processes is much larger, we can imagine that all the images of a single person in the world is smaller than the maximum number of people in the world. We cannot reasonably accept that the total number of images of all observable processes in nature is smaller than the maximum number of people in the world. We need to respect orthogonality constraints in the data representation in the reduced space. There is no reason to believe that the number of causes or the number of subspaces used need to respect orthogonality constraints in the data representation in the reduced space. The obvious question is: Why should we constrain our sensory world, although the number of possible processes is much larger, we can imagine that all the images of a single person in the world is smaller than the maximum number of people in the world. We cannot reasonably accept that the total number of images of all observable processes in nature is smaller than the maximum number of people in the world. We need to respect orthogonality constraints in the data representation in the reduced space.

Representation in the reduced space.

...
Emergence of simple-cell receptive field properties by learning a sparse code for natural images

Bruno A. Olshausen* & David J. Field

Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853, USA

The receptive fields of simple cells in mammalian primary visual cortex can be characterized as being spatially localized, oriented and bandpass (selective to structure at different spatial scales), comparable to the basis functions of wavelet transforms. One approach to understanding such response properties of visual neurons has been to consider their relationship to the statistical structure of natural images in terms of efficient coding. Along these lines, a number of studies have attempted to train unsupervised learning algorithms on natural images in the hope of developing receptive fields with similar properties, but none has succeeded in producing a full set that spans the image space and contains all three of the above properties. Here we investigate the proposal that a coding strategy that maximizes sparseness is sufficient to account for these properties. We show that a learning algorithm that

FIG. 1 Principal components calculated on 8 × 8 image patches extracted from natural scenes by using Sanger’s rule. The full set of 64 components is shown, ordered by their variance (by columns, then by rows). The oriented structure of the first few principal components does not arise as a result of the oriented structures in natural images, but rather because these functions are composed of a small number of low-frequency components (the lowest spatial frequencies account for the greatest part of the variance.
K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation

Michal Aharon, Michael Elad, and Alfred Bruckstein

Abstract—In recent years there has been a growing interest in the study of sparse representation of signals. Using an overcomplete dictionary that contains prototype signal-atoms, signals are described by sparse linear combinations of these atoms. Applications that use sparse representation are many and include compression, regularization in inverse problems, feature extraction, and more. Recent activity in this field has concentrated mainly on the study of pursuit algorithms that decompose signals with respect to a given dictionary. Designing dictionaries to better fit the above model can be done by either selecting one from a prespecified set of linear transforms or adapting the dictionary to a set of training signals. Both of these techniques have been considered, but this topic is largely still open. In this paper we propose a novel algorithm for adapting dictionaries in order to achieve sparse signal representations. Given a set of training signals, we seek the dictionary that leads to the best representation for each member in this set, under strict sparsity constraints. We present a new method—the K-SVD algorithm—generalizing the K-means clustering process. K-SVD is an iterative method that alternates between sparse coding of the examples based on the current dictionary and a process of updating the dictionary atoms to better fit the data. The update of the dictionary columns is combined with an update of the sparse representations, thereby accelerating convergence. The K-SVD algorithm is flexible and can work with any pursuit method (e.g., basis pursuit, FOCUSS, or matching pursuit). We analyze this algorithm and demonstrate its results both on synthetic tests and in applications on real image data.

The deviation are the \( \ell^p \)-norms for \( p = 1, 2, \) and \( \infty \). In this paper, we shall concentrate on the case of \( p = 2 \).

If \( n < K \) and \( \mathbf{D} \) is a full-rank matrix, an infinite number of solutions are available for the representation problem, hence constraints on the solution must be set. The solution with the fewest number of nonzero coefficients is certainly an appealing representation. This sparsest representation is the solution of either

\[
(P_0) \quad \min_{\mathbf{x}} \| \mathbf{x} \|_0 \quad \text{subject to} \quad \mathbf{y} = \mathbf{D}\mathbf{x}
\]

or

\[
(P_{0,\epsilon}) \quad \min_{\mathbf{x}} \| \mathbf{x} \|_0 \quad \text{subject to} \quad \| \mathbf{y} - \mathbf{D}\mathbf{x} \|_2 \leq \epsilon
\]

where \( \| \cdot \|_0 \) is the \( l^0 \) norm, counting the nonzero entries of a vector.

Applications that can benefit from the sparsity and overcompleteness concepts (together or separately) include compression, regularization in inverse problems, feature extraction, and more. Indeed, the success of the JPEG2000 coding standard can be attributed to the sparsity of the wavelet coefficients of natural images [1]. In denoising, wavelet methods and shift-invariant variations that exploit overcomplete representation are among the most widely used denoising procedures.
Inverse problems and dictionary learning

Restore initial image $\Rightarrow$ good representation $=$ linear regression

- $H$ : damaging operator (blur, mask...)
- $D$ : dictionary (cosines, wavelets, learnt atoms...)
- $\alpha$ : coefficients, $X_i = \sum_j \alpha_{ij} D_j$

\[
\begin{align*}
Y &= HX + n \\
X &= D\alpha
\end{align*}
\]

and $X$ is sparse on $D$ to be discovered... **Prior**: $\alpha$ is sparse
Inverse problems and dictionary learning

Restore initial image \(\Rightarrow\) good representation = linear regression

- \(\mathbf{H}\) : damaging operator (blur, mask...)
- \(\mathbf{D}\) : dictionary (cosines, wavelets, learnt atoms...)
- \(\alpha\) : coefficients, \(\mathbf{X}_i = \sum_j \alpha_{ij} \mathbf{D}_j\)

\[
\begin{align*}
\mathbf{Y} & = \mathbf{H}\mathbf{X} + \mathbf{n} \\
\mathbf{X} & = \mathbf{D}\alpha
\end{align*}
\]

and \(\mathbf{X}\) is sparse on \(\mathbf{D}\) to be discovered... **Prior** : \(\alpha\) is sparse
Inverse problems and dictionary learning

Restore initial image \( \Rightarrow \) good representation = linear regression

- \( \mathbf{H} \): damaging operator (blur, mask...)
- \( \mathbf{D} \): dictionary (cosines, wavelets, learnt atoms...)
- \( \alpha \): coefficients, \( \mathbf{X}_i = \sum_j \alpha_{ij} \mathbf{D}_j \)

\[
\begin{aligned}
\begin{cases}
\mathbf{Y} &= \mathbf{HX} + \mathbf{n} \\
\hat{\mathbf{X}} &= \mathbf{D}\alpha
\end{cases}
\end{aligned}
\]

and \( \hat{\mathbf{X}} \) is sparse on \( \mathbf{D} \) to be discovered... **Prior**: \( \alpha \) is sparse

**Optimization typical approach** :

\[
\begin{aligned}
\mathbf{Y} &= (\mathbf{HX} + \text{Gaussian noise}) + \text{regularization} \\
(\mathbf{D}, \alpha) &= \text{argmin}_{\mathbf{D}, \alpha} \| \mathbf{Y} - \mathbf{H} \hat{\mathbf{D}} \alpha \|_2^2 + \lambda \| \alpha \|_{L0} \\
&\downarrow \\
&\text{LASSO, Forward-Backward, proximal methods...}
\end{aligned}
\]
Inverse problems and dictionary learning

Restore initial image $\Rightarrow$ good representation $=$ linear regression

- $H$: damaging operator (blur, mask...)
- $D$: dictionary (cosines, wavelets, learnt atoms...)
- $\alpha$: coefficients, $X_i = \sum_j \alpha_{ij} D_j$

\[
\begin{align*}
Y &= HX + n \\
X &= D\alpha
\end{align*}
\]

and $X$ is sparse on $D$ to be discovered... **Prior**: $\alpha$ is sparse

**Optimization typical approach** :

\[
Y = (HX + \text{Gaussian noise}) + \text{regularization}
\]

\[
(D, \alpha) = \arg\min_{D,\alpha} \| Y - H D\alpha \|^2_{\hat{X}} + \lambda \| \alpha \|_{L1}^{\text{Laplace}}
\]

$\Downarrow$

LASSO, Forward-Backward, proximal methods...
Dictionary learning

- Searching for an adaptive representation
  - ... for dimension reduction (correlated atoms, $K < p$)
  - ... for an orthonormal basis (PCA, $K = p$),
  - ... for a redundant dict. / sparse representation ($K > p$)

- Optimization problem:

\[
Y = (HX + \text{Gaussian noise}) + \text{regularization}
\]

\[
(D, \alpha) = \arg\min_{D,\alpha} \| Y - H D \hat{\alpha} \|^2 + \lambda \| \hat{\alpha} \|_{L0}
\]

\[\downarrow\]

Alternate optimization (proximal methods...)

\[
\hat{\alpha}
\]
Dictionary learning

- Searching for an adaptive representation
  - ... for dimension reduction (correlated atoms, $K < p$)
  - ... for an orthonormal basis (PCA, $K = p$),
  - ... for a redundant dict. / sparse representation ($K > p$)

- Optimization problem:

$$
\mathbf{Y} = (\mathbf{HX} + \text{Gaussian noise}) + \text{regularization}
$$

$$(\mathbf{D}, \mathbf{\alpha}) = \arg\min_{\mathbf{D}, \mathbf{\alpha}} \| \mathbf{Y} - \mathbf{H} \hat{\mathbf{D}} \mathbf{\alpha} \|^2 + \lambda \| \mathbf{\alpha} \|_{L1}
$$

$\downarrow$

Alternate optimization (proximal methods...)

Dictionary learning

- Searching for an adaptive representation
  - to solve inverse problems,
  - for compression,
  - for classification,…

- Many open questions:
  - optimize the size of the dictionary?
  - optimize sparsity?…

- Many generalizations:
  - distributed setting,
  - task driven dictionaries,
  - multiscale dictionary,
  - fast transforms,
  - multispectral/hyperspectral,
  - multimodal dictionaries…

---

**Word “One”**

- Audio
- Video

---

In stereo vision, the same three-dimensional (3-D) scene is captured by different sensors. The left and the right images are well represented by a dictionary of local atom transforms, which is made feasible by the use of geometric dictionaries built on scaling, shearing, and the epipolar geometry constraints. Dictionaries can be learned such that they efficiently describe the content of natural stereo images. Through the projective properties of light rays, different measurement devices capture the same scene from different views. Audio-visual signals are generated by the brain to capture perceived sensory information and to solve inverse problems. For instance, in the auditory area of the cortex, a video part corresponding to the movement of the lips that generates the cause that is human speech is exploited. A multimodal dictionary learned with elements that have an audio part and a video part is shown in Figure 4. One important contribution of this work certainly lies in its benefits towards understanding and modeling the integration of audio and visual sensory information in the cortex.
Thèse de Hong Phuong Dang :

Méthodes bayésiennes non paramétriques pour l’apprentissage de dictionnaire (2013-2016)

Post-doc de Sylvain Rousseau (co-encad. Christelle Garnier) :


L’objectif du post-doc est d’explorer le potentiel des méthodes de représentation parcimonieuse pour repousser les limites du filtrage particulaire en grande dimension (tâche 4 du projet). L’application visée est le suivi multi-objets dans des séquences d’images.

ANR Bayesian Non Parametrics for Signal & Image processing (BNPSI, 2014-2018)