Sparse coding and dictionary learning

Domingo Mery

Departamento de Ciencia de la Computación
Universidad Católica de Chile
Santiago de Chile
September 24, 2013

Departamento de Ciencia de la Computación
Universidad Católica de Chile
Domingo Mery:
Av. Vicuña Mackenna 4860(143)
eMail: dmery@ing.puc.cl
http://dmery.ing.puc.cl

Contents

Chapter 1

Introduction

In recent years, sparse representation has been widely used in signal processing [30], neuroscience [27], statistics [6], sensors [40] and computer vision [37, 41]. In many computer vision applications, under assumption that natural images can be represented using sparse decomposition [26] state-of-the-art results have been significantly improved. In these applications, the performance can be improved by learning non-parametric dictionaries for the sparse representation (instead of using fixed dictionaries).

In signal processing, it is very convenient to estimate a new representation of a signal in order to analyze it efficiently. The idea, is that this representation captures a useful characterization of the signal for analysis tasks, *e.g.*, feature extraction for pattern recognition, frequency spectrum for denoising, etc. An appropriate representation, due to its simplicity, is obtained by a linear transform. Thus, a signal $\mathbf{y} \in \mathbb{R}^n$ can be expressed as a linear combination of a set of elementary signals $\mathbf{D} = [\mathbf{d}_1 \ \mathbf{d}_2 \ \dots \ \mathbf{d}_K] \in \mathbb{R}^{n \times K}$ as:

$$\mathbf{y} = \mathbf{D}\mathbf{x}, \tag{1.1}$$

where the vector $\mathbf{x} \in \mathbb{R}^K$ corresponds to the representation coefficients of signal $\mathbf{y}$. In this representation, matrix $\mathbf{D}$ and its columns $\mathbf{d}_k$ are commonly known as *dictionary* and *atoms* respectively.

Chapter 2

Traditional dictionaries

When every signal can be uniquely represented by a linear combination, the dictionary $\mathbf{D}$ corresponds to a *basis*. This is the case of Discrete Fourier Transform (DFT) for example, where the basis functions are sine and cosine waves with unity amplitude. In this case, the element j of atom $\mathbf{d}_k$ is defined as $d_{jk} = \exp(2\pi ijk/n)$ with $K = n$ and $i = \sqrt{-1}$ [11]. It is well known that for some applications, *e.g.*, signal filtering, instead of processing the signal $\mathbf{y}$, it can be more convenient to process the signal in frequency domain $\mathbf{x}$, because it can be used to separate low and high frequencies effectively. Nevertheless, the Fourier basis is very inefficient when representing for example a discontinuity, because its representation coefficients are over all frequencies and the analysis becomes difficult or even impossible. Other *predefined* basis, *i.e.*, where the atoms are *fixed*, are Discrete Cosine Transformation (DCT) and Wavelets (*e.g.*, Gabor) among others [11]. In many applications, since these dictionaries are fixed, they cannot represent more complex and high-dimensional signals satisfactorily [29].

In order to avoid the mentioned problem with fixed dictionaries, another way to represent a signal is using a *learned* dictionary, *i.e.*, a dictionary that is estimated from representative signal examples. This is the case of Principal Component Analysis (PCA), or Karhunen-Loève Transform (KLT) [17], where the dictionary $\mathbf{D}$ is computed using the first K eigenvectors of the eigenvalue decomposition of the covariance matrix $\mathbf{\Sigma}$, which is usually estimated from a set of zero-means signal examples $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$. The basis here represents K orthogonal functions (with $K \leq n$) that transforms $\mathbf{Y}$ in a set of linearly uncorrelated signals $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ called the K *principal*

components. This relationship is expressed as $\mathbf{Y} = \mathbf{DX}$. In this case, KLT represents a signal more efficiently than DFT because the dictionary is not fixed and it is *learned* from signal examples [29].

The mentioned dictionaries are *orthogonal*, *i.e.*, each atom $\mathbf{d}_i$ is orthogonal to atom $\mathbf{d}_j$ in $\mathbb{R}^n$ space $\forall i \neq j$. Therefore, a signal $\mathbf{y}$ is represented as a sum of orthogonal vectors $x_i\mathbf{d}_i$. In addition, most of these dictionaries are *orthonormal*, with $||\mathbf{d}_i|| = 1$ and $\mathbf{D}^\mathsf{T}\mathbf{D} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Hence, it is very simple to calculate $\mathbf{X} = \mathbf{D}^\mathsf{T}\mathbf{Y}$.

Chapter 3

Sparse dictionaries

Due to their mathematical simplicity, the orthogonal dictionaries dominated this kind of analysis for years. Nevertheless, there is no reason to accept as true that the number of atoms, required to characterize a set of signals, must be smaller than the dimension of the signal. Moreover, why the atoms of the dictionary should be orthogonal? The limited effectiveness of these dictionaries led to the development of newer dictionaries that can represent a wider range of signal phenomena, namely the *overcomplete* ones that have more atoms than the dimension of the signal ($K > n$) with no necessarily orthogonal atoms [33]. A seminal work in learning overcomplete dictionaries for image representation was presented by Olshausen and Field [25, 26]. They estimated –from small image patches of natural images– a *sparse* representation which were extremely similar to the mammalian simple-cell receptive fields (at that time, this phenomenon could only be described using Gabor filters). The key idea for representanting natural signals is that although the number of possible atoms in the overcomplete dictionary is huge, the number of those atoms required to represent a signal is much smaller, *i.e.*, the signals are sparse in the set of all possible atoms [33].

Sparse coding models a signal as a linear combination (1.1), or approximate, $\mathbf{y} \approx \mathbf{D}\mathbf{x}$, using a *sparse* linear combination of atoms from a learned dictionary, *i.e.*, only a few atoms from $\mathbf{D}$ are allowed to use in the linear combination (most of coefficients of $\mathbf{x}$ are zero) and the atoms are not fixed (the dictionary is adapted to fit a given set of signal examples). In this case, the basis are not orthogonal.

Thus, from a representative set of signals $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N}$, the idea is *i)* to learn a dictionary $\mathbf{D} = \{\mathbf{d}_k\}_{k=1}^{K}$ and *ii)* to estimate the corresponding sparse representations $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ of the original signals $\mathbf{Y}$.

In K-means algorithm –a very known algorithm used in clustering–, the sparsity is *extreme* because for the representation of $\mathbf{y}$ only one atom of $\mathbf{D}$ is allowed, and the corresponding coefficient of $\mathbf{x}$ is 1. In this case, the dictionary and coefficients are estimated by:

$$\mathbf{D}^*, \mathbf{X}^* = \underset{\mathbf{D},\mathbf{X}}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 \quad \text{subject to } \forall i, \mathbf{x}_i = \mathbf{e}_k \text{ for some } k \tag{3.1}$$

where $\mathbf{e}_k$ is a vector from the trivial basis, with all zero entries except a one in k-th position. In this equation, it is used the Frobenius norm defined as $||\mathbf{A}||_F^2 = \sum_{ij} a_{ij}^2$. In clustering problems, the atom $\mathbf{d}_k$ is the centroid of samples $\mathbf{y}_i$ that fulfill $\mathbf{x}_i = \mathbf{e}_k$. Thus, a signal $\mathbf{y}$ belongs to cluster k if it is closer to centroid k than any other centroids (in this case, its representation is $\mathbf{x} = \mathbf{e}_k$ and the corresponding atom is $\mathbf{d}_k$).

Sparsisty in general, can be expressed as:

$$\mathbf{D}^*, \mathbf{X}^* = \underset{\mathbf{D},\mathbf{X}}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 \quad \text{subject to } ||\mathbf{x}_i||_0 \leq T \tag{3.2}$$

where $||\mathbf{x}_i||_0$ is the ℓ^0 norm, counting the nonzero entries of $\mathbf{x}_i$. The goal is to express a new signal $\mathbf{y}$ as a linear combination of a small number of signals take from the dictionary. This optimization problem can be expressed as:

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}}(||\mathbf{y} - \mathbf{D}^*\mathbf{x}||_2^2 + \lambda||x||_1) \tag{3.3}$$

It can be demonstrated that the solution of the ℓ^0 minimization problem (3.2) is equivalent to the solution of the ℓ^1 minimization problem [5]:

$$\underset{\mathbf{D},\mathbf{X}}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 \quad \text{subject to } ||\mathbf{x}_i||_1 \leq T \tag{3.4}$$

Thus, on one hand, the *dictionary learning problem* is as follows: given a set of training signals $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N}$, find the dictionary $\mathbf{D}$ (and a set of representation coefficients $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$) that represents at best each signal using the sparsity constraint (3.3), where no more than T atoms are allowed in each decomposition $\mathbf{x}_i$. On the other hand, the *sparse coding problem* can

be stated as follows: given a signal $\mathbf{y}$ and a learned dictionary $\mathbf{D}$, find $\mathbf{x}$, the representation of signal $\mathbf{y}$, as:

$$\underset{\mathbf{x}}{\operatorname{argmin}} ||\mathbf{x}||_0 \quad \text{subject to } ||\mathbf{y} - \mathbf{D}\mathbf{x}||_2 < \epsilon \tag{3.5}$$

where ϵ is the error tolerance.

Chapter 4

Dictionary learning

There are three categories of algorithms used to learn dictionaries [33]: *i)* probabilistic methods, *ii)* methods based on clustering, and *iii)* methods with a particular construction. Probabilistic methods are based on maximum likelihood approach, *i.e.*, given the generative model (1.1), the objective is to maximize the likelihood that the representative samples have efficient, sparse representations in a redundant dictionary given by $\mathbf{D}$ [26, 19, 12, 39]. In clustering-based methods, the representative samples are grouped into patterns such their distance to a given atom is minimal. Afterwards, the atoms are updated such that the overall distance in the group of patterns is minimal. This schema follows a K-means algorithm. In order to generalize the K-means algorithm, the 'K-SVD' algorithm was developed [2]. The method has two steps: a) it uses orthogonal matching pursuit (OMP) algorithm for the sparse approximation[1], b) the columns of the dictionary are sequentially updated using single value decomposition (SVD) decomposition to minimize the approximation error- It is reported that dictionaries learned with K-SVD show excellent performance in image denoising [7, 24] among other applications. Finally, dictionaries with specific structures uses (instead of general forms of atoms) a set of *parametric functions* that can describe the atoms shortly, *i.e.*, the generating functions and the parameters build the dictionary functions. Thus, the problem is reduced to learn the parameters for one or more generating functions (see for example [21, 13]).

[1] OMP is a greedy algorithmm that iteratively selects locally optimal basis vectors [36].

Chapter 5

Classification using sparse representation

According to equation (3.3) it is possible to learn the dictionary $\mathbf{D}$ and estimate the most important constitutive components $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ of the representative signals $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$. In a supervised problem –with labeled data $(\mathbf{y}_i, c_i)$, where c_i is the class of sample $\mathbf{y}_i$–, naturally the classification problem can be stated as follows [3]: given training data $(\mathbf{y}_i, c_i)$, design a classifier h –with parameters θ– which maps the transformed samples $\mathbf{x}_i$ to its classification label c_i, thus, $h(\mathbf{x}_i, \theta)$ should be c_i. In order to classify a new sample data $\mathbf{y}$, it is transformed into $\mathbf{x}$ using dictionary $\mathbf{D}$ and then it is classified as $c = h(\mathbf{x}, \theta)$. Nevertheless, since $\mathbf{X}$ is estimated to represent the original data efficiently, there is no reason to accept as true that this new representation can ensure an optimal separation of the classes. Another classification strategy uses one dictionary $\mathbf{D}_k$ per class [22], that is learned using the set $\mathbf{Y}_k$, that contains only the samples of class k of the training data: $\mathbf{Y}_k = \{\mathbf{y}_i | c_i = k\}$. With this strategy, using (3.4) a test sample $\mathbf{y}$ is codified by $\mathbf{x} = \mathbf{x}_k$ with dictionary $\mathbf{D} = \mathbf{D}_k$ for all classes $k = 1 \dots C$, and a reconstruction error is computed as $e_k = ||\mathbf{y} - \mathbf{D}_k \mathbf{x}_k||$. Finally, sample $\mathbf{y}$ is assigned to the class c with the smallest reconstruction error: $c = \text{argmin}_k(e_k)$. This test strategy, however, does not scale well for a large number of classes.

For these reasons, new strategies have been developed in order to learn at the same time *reconstructive* and *discriminative* dictionaries (for robustness to noise and for efficient classification respectively) [33]. This can be achieved by adding a new discrimination term in the objective function that includes

the representation that is also the most different from the one of signals in other data classes:

$$\underset{\mathbf{D},\mathbf{X},\theta}{\operatorname{argmin}}[||\mathbf{Y}-\mathbf{D}\mathbf{X}||_2^2+\gamma J(\mathbf{D},\mathbf{X},\mathbf{c},\theta)] \quad \text{subject to } ||\mathbf{x}||_0 \leq T. \tag{5.1}$$

The discrimination term $J(\mathbf{D},\mathbf{X},\mathbf{c},\theta)$ depends on the dictionary, the coefficient vectors, the labels of the samples, and the parameters θ of the model used for classification. Parameter γ weights the tradeoff between approximation and classification performance. This strategy with a common dictionary, has the advantage of sharing some atoms of the dictionary when representing samples of different classes. Equation (5.1) can be solved efficiently by fixed-point continuation methods when the classifier is based on logistic regression methods [23].

Chapter 6

Dictionaries in multiple-view images

There is only a very few number of approaches for sparse coding of multiple view images reported in the literature: In [14], independent component analysis (ICA) is used in order to learn an orthogonal basis of stereo images that are represented as a linear combination of stereo basis functions. In [4], a learning method that uses the receptive fields of binocular neurons is proposed. The fields are learned by maximizing the mutual information between the stereo image model and the disparity. Recently, in [32, 34], a very interesting method for learning overcomplete dictionaries for stereo image representation is proposed. The approach learns stereo atoms from stereo image pairs under explicit geometric (*epipolar*) constraints. Thus, two dictionaries are defined –dictionary $\mathbf{D}_L$ for left view and dictionary $\mathbf{D}_R$ for right view–, in which the atoms of $\mathbf{D}_R$ are computed from atoms of $\mathbf{D}_L$ using local transformations like rotation, translation and anisotropic scalings. The dictionaries are estimated using the expectation-maximization algorithm, that optimizes a maximum-likelihood objective function. The method was tested in the case of omnidirectional images, and camera pose estimation, yielding promising results.

Chapter 7

Other applications

Applications that use sparse representations include denoising [7, 24], compression [31], inpaiting [10, 24, 9], image restoration [24], face recognition [38, 2, 33], medical imaging [35, 20, 1], audio [28, 16], dimensionality reduction [18], and classification [15, 37].

Bibliography

[1] C.K. Abbey, J.N. Sohl-Dickstein, B.A. Olshausen, M.P. Eckstein, and J.M. Boone. Higher-order scene statistics of breast images. In *Proceedings of SPIE*, volume 7263, page 726317, 2009.

[2] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[3] L. Bar and G. Sapiro. Hierarchical dictionary learning for invariant classification. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 3578–3581, 2010.

[4] D. Chen, D. Varodayan, M. Flierl, and B. Girod. Wyner-ziv coding of multiview images with unsupervised learning of disparity and gray code. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 1112–1115. IEEE, 2008.

[5] David L Donoho. For most large underdetermined systems of linear equations the minimal ??1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.

[6] D.L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

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

[8] K. Engan, S.O. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. In *Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings., 1999 IEEE International Conference on*, volume 5, pages 2443–2446. IEEE, 1999.

[9] M.J. Fadili and J.L. Starck. EM algorithm for sparse representation-based image inpainting. In *IEEE International Conference on Image Processing (ICIP-2005)*, volume 2, pages II–61, 2005.

[10] M.J. Fadili, J.L. Starck, and F. Murtagh. Inpainting and zooming using sparse representations. *The Computer Journal*, 52(1):64–79, 2009.

[11] R.C. Gonzalez, R.E. Woods, and S.L. Eddins. *Digital Image Processing*. Prentice Hall Press, 2007.

[12] I.F. Gorodnitsky and B.D. Rao. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45(3):600–616, 1997.

[13] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12):3320–3325, 2003.

[14] P.O. Hoyer and A. Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191–210, 2000.

[15] K. Huang and S. Aviyente. Sparse representation for signal classification. *Advances in neural information processing systems*, 19:609, 2007.

[16] M.G. Jafari and M.D. Plumbley. Speech denoising based on a greedy adaptive dictionary algorithm. In *Proc. European Signal Processing Conf*, 2009.

[17] I. Joliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[18] E. Kokiopoulou and P. Frossard. Semantic coding by supervised dimensionality reduction. *Multimedia, IEEE Transactions on*, 10(5):806–818, 2008.

[19] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T.W. Lee, and T.J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003.

[20] B. Mailhé, R. Gribonval, F. Bimbot, M. Lemay, P. Vandergheynst, and J.M. Vesin. Dictionary learning for the sparse modelling of atrial fibrillation in ecg signals. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 465–468. IEEE, 2009.

[21] B. Mailhé, S. Lesage, R. Gribonval, F. Bimbot, P. Vandergheynst, et al. Shift-invariant dictionary learning for sparse representations: extending k-svd. *Proc European Signal Processing Conference*, 4, 2008.

[22] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[23] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. Technical Report 6652, INRIA, September 2008.

[24] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, 2008.

[25] B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[26] B.A. Olshausen and D.J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[27] B.A. Olshausen and D.J. Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004.

[28] M.D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M.E. Davies. Sparse representations in audio and music: from coding to source separation. *Proceedings of the IEEE*, 98(6):995–1005, 2010.

[29] R. Rubinstein, A.M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[30] R. Rubinstein, M. Zibulevsky, and M. Elad. Double sparsity: Learning sparse dictionaries for sparse signal approximation. *Signal Processing, IEEE Transactions on*, 58(3):1553–1564, 2010.

[31] J.L. Starck, M. Elad, and D.L. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Transactions on Image Processing*, 14(10):1570–1582, 2005.

[32] I. Tošić and P. Frossard. Distributed multi-view image coding with learned dictionaries. In *Proceedings of the 5th International ICST Mobile Multimedia Communications Conference*, page 27. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2009.

[33] I. Tosic and P. Frossard. Dictionary learning. *Signal Processing Magazine, IEEE*, 28(2):27–38, 2011.

[34] I. Tosic and P. Frossard. Dictionary learning for stereo image representation. *Image Processing, IEEE Transactions on*, 20(4):921–934, 2011.

[35] I. Tosic, I. Jovanovic, P. Frossard, M. Vetterli, and N. Duric. Ultrasound tomography with learned dictionaries. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5502–5505. IEEE, 2010.

[36] J.A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[37] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

[38] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2009.

[39] M. Yaghoobi, T. Blumensath, and M.E. Davies. Dictionary learning for sparse approximations with the majorization method. *IEEE Transactions on Signal Processing*, 57(6):2178–2191, 2009.

[40] A.Y. Yang, M. Gastpar, R. Bajcsy, and S.S. Sastry. Distributed sensor perception via sparse representation. *Proceedings of the IEEE*, 98(6):1077–1088, 2010.

[41] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1794–1801, 2009.