View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Infoscience - École polytechnique fédérale de Lausanne

Ivana Tošić and Pascal Frossard

Dictionary Learning

What is the right representation for my signal?

© DIGITAL STOCK & LUSHPIX

Huge amounts of high-dimensional information are captured every second by diverse natural sensors such as the eyes or ears, as well as artificial sensors like cameras or microphones. This information is largely redundant in two main aspects: it often contains multiple correlated versions of the same physical world and each version is usually densely sampled by generic sensors. The relevant information about the underlying processes that cause our observations is generally of much reduced dimensionality compared to such recorded data sets. The extraction of this relevant information by identifying the generating causes within classes of signals is the central topic of this article. We present methods for determining the proper representation of data sets by means of reduced dimensionality subspaces, which are adaptive to both the characteristics of the signals and the processing task at hand. These representations are based on the principle that our observations can be described by a sparse subset of atoms taken from a redundant dictionary, which represents the causes of our observations of the world. We describe methods for learning dictionaries that are appropriate for the representation of given classes of signals and multisensor data. We further show that dimensionality reduction based on dictionary representation can be extended to address specific tasks such as data analysis or classification when the learning includes a class separability criteria in the objective function. The benefits of dictionary learning clearly show that a proper understanding of causes underlying the sensed world is key to task-specific representation of relevant information in high-dimensional data sets.

WHAT IS THE GOAL OF DIMENSIONALITY REDUCTION?

Natural and artificial sensors are the only tools we have for sensing the world and gathering information about physical processes and their causes. These sensors are usually not aware of the physical process underlying the phenomena they "see," hence they often sample the information with a higher rate than the effective dimension of the process. However, to store, transmit or analyze the processes we observe, we do not need such abundant data: we only need the information that is relevant to understand the causes, to reproduce the physical processes, or to make decisions. In other words, we can reduce the

Digital Object Identifier 10.1109/MSP.2010.939537
Date of publication: 17 February 2011

1053-5888/11/$26.00©2011IEEE

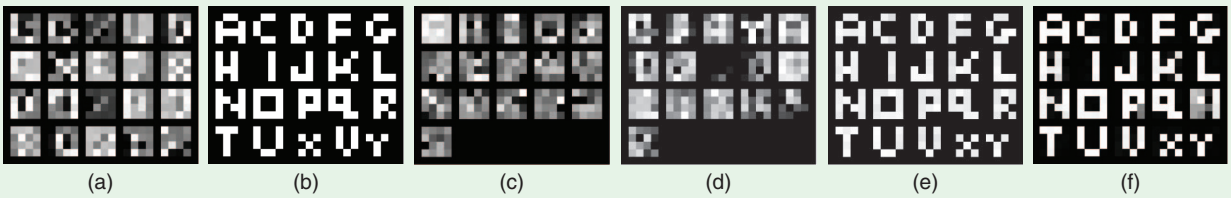
dimension of the sampled data to the effective dimension of the underlying process without sensible penalty in the subsequent data analysis procedure.

An intuitive way to approach this dimensionality reduction problem is first to look at what generates the dimensionality gap between the physical processes and the observations. The most common reason for this gap is the difference between the representation of data defined by the sensor and the representation in the physical space. In some cases, this discrepancy is, for example, a simple linear transform of the representation space, which can be determined by the well-known principal component analysis (PCA) [1] method. It may however happen that the sensors observe simultaneously two or more processes with causes lying within different subspaces. Other methods such as independent component analysis (ICA) [2] are required to understand the different processes behind the observed data. ICA is able to separate the different causes or sources by analyzing the statistical characteristics of the data set and minimizing the mutual information between the observed samples. However, ICA techniques respect some orthogonality conditions such that the maximal number of causes is often limited to the signal dimension. In Figure 1(a), we show some examples of noisy images whose underlying causes are linear combinations of two English letters chosen from a dictionary in Figure 1(b). These images are 4×4 pixels, hence their dimensionality in the pixel space is 16, while the number of causes is 20 (total number of letters). When applied to 5,000 randomly chosen noisy samples of these letters, PCA finds a linear transform of the pixels space into another 16-dimensional space represented by vectors in Figure 1(c). This is done by finding the directions in the original space with the largest variance. However, this representation does not identify the processes that generate the data, i.e., it does not find our 20 letters. ICA [2] differs from PCA because it is able to separate sources not only with respect to the second order correlations in a data set, but also with respect to higher order statistics. However, since the maximal number of causes is equivalent to the signal dimension in the standard ICA, the subspace vectors found by ICA in the example of Figure 1(d) do not explain the underlying letters.

The obvious question is: Why should we constrain our sensors to observe only a limited number of processes? Why do we need to respect orthogonality constraints in the data representation subspace? There is no reason to believe that the number of all observable processes in nature is smaller than the maximal dimension in existing sensors. If we look for an example in a 128×128 dimensional space of face images for all the people in the world, we can imagine that all the images of a single person belong to the same subspace within our 16,384-dimensional space, but we cannot reasonably accept that the total number of people in the world is smaller than our space dimension. We conclude that the representation of data could be overcomplete, i.e., that the number of causes or the number of subspaces used for data description can be greater than the signal dimension.

Where does the dimensionality reduction occur in this case? The answer to this question lies in one of the most important principles in sensory coding—efficiency, as first outlined by Barlow [3]. Although the number of possible processes in the world is huge, the number of causes that our sensors observe at a single moment is much smaller: the observed processes are sparse in the set of all possible causes. In other words, although the number of representation subspaces is large, only few ones will contain data samples from sensor measurements. By identifying these few subspaces, we find the representation in the reduced space.

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, but it is not trivial. This question has triggered the emergence of a new and promising research field called dictionary learning. It focuses on the development of novel algorithms for building dictionaries 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 problems; they enforce the identification of the most important causes of the observed data and favor the accurate representation of the relevant information. Figure 1(e) shows that one of the first dictionary learning methods called sparse coding [4] succeeds in learning all 20 letters that generate 5,000 observations



[FIG1] Learning underlying causes from a set of noisy observations of English letters. A subset of 20 noisy 4×4 images is shown in (a). These samples have been generated as linear combinations of two letters randomly chosen from the alphabet in (b), and they have been corrupted by additive Gaussian noise. When run of 5,000 such samples, PCA and ICA find the same number of components as the dimension of the signal. Therefore, they cannot find the underlying 20 letters. Sparse coding [4] learns an overcomplete dictionary of 20 components, thus it can separate these causes and find all 20 letters from the original alphabet. K-SVD [5] performs similarly, i.e., it finds almost all of the letters. However, since the implementation of K-SVD [5] uses MP for the sparse approximation step, it converges to a local minimum resulting in some repeated letters in the learned dictionary. (a) Noisy samples; (b) original causes; (c) PCA; (d) ICA; (e) sparse coding; and (f) KSVD.

in our simple example. In the course of the last decade, dictionary optimization has led to significant performance improvements in high-dimensional signal processing tasks such as audio, image, multiview, and multimodal data analysis.

This article presents the main challenges in the field of dictionary learning for dimensionality reduction. We first present a brief description of sparse approximations. Next, we give a tutorial overview of the main algorithms that permit the construction of dictionaries for the sparse representation of given classes of signals, possibly with properties such as large incoherence or model-based structures. In the section "Applications of Dictionary Learning," we present a few signal processing applications where the objectives of the learning algorithms is adapted to specific problems such as the joint analysis of correlated signals like audio-visual signals and stereo images. We later show in the section "Learning for Classification" that the construction of dictionaries can also be constrained in order to satisfy discriminative objectives; the dimensionality reduction steps not only lead to good approximation but also efficient classifications of signals.

SPARSE APPROXIMATIONS

The goal of sparse representation is to express a given signal $\mathbf{y}$ of dimension n as a linear combination of a small number of signals taken from a "resource" database, which is called the dictionary. Elements of the dictionary are typically unit norm functions called atoms. Let us denote the dictionary as $\mathcal{D}$ and the atoms as ϕ_k, $k = 1, \ldots, N$, where N is the size of the dictionary. The dictionary is overcomplete $(N > n)$ when it spans the signal space and its atoms are linearly dependent. In that case, every signal can be represented as a linear combination of atoms in the dictionary

$$\mathbf{y} = \mathbf{\Phi a} = \sum_{k=1}^{N} a_k \phi_k. \tag{1}$$

Because the dictionary is overcomplete, $\mathbf{a}$ is not unique. This is where the sparsity constraint comes into play. To achieve efficient and sparse representations, we generally relax the requirement for finding the exact representation. We look for a sparse linear expansion with an approximation error η of bounded energy ϵ. The objective is now to find a sparse vector $\mathbf{a}$ that contains a small number of significant coefficients, while the rest of the coefficients are close or equal to zero. In other words, we want to minimize the resources (atoms) that we use to accomplish the task of signal representation. This optimization problem can be formulated as follows:

$$\min_{\mathbf{a}} \|\mathbf{a}\|_0 \text{ subject to } \mathbf{y} = \mathbf{\Phi a} + \eta \text{ and } \|\eta\|_2^2 < \epsilon, \tag{2}$$

where $\|\cdot\|_p$ denotes the l_p norm. Unfortunately, this problem is NP-hard. However, there exist polynomial time approximation algorithms that find a suboptimal solution for the sparse vector $\mathbf{a}$. These algorithms can be classified in two main groups. The first group includes greedy algorithms such as the matching pursuit (MP) [6] and the orthogonal MP (OMP) [7], which iteratively select locally optimal basis vectors. In the second group, we find algorithms based on convex relaxation methods such as the basis pursuit denoising [8] or least absolute shrinkage and selection operator (LASSO) [9], which solve the following problem:

$$\min_{\mathbf{a}} \left(\|\mathbf{y} - \mathbf{\Phi a}\|_2^2 + \lambda \|\mathbf{a}\|_1\right). \tag{3}$$

The convex relaxation permits to replace the nonconvex l_0 norm in the original problem by the convex l_1 norm. The l_0 norm of a vector is equal to the number of nonzero elements in that vector. It is called a "norm" because it is the limit of p-norms as p approaches zero. However, note that it is not a true norm, unlike the l_1 norm that has all properties of a norm. Besides pursuit algorithms, there exist other sparse approximation algorithms such as the focal underdetermined system solver (FOCUSS) [10] and sparse Bayesian learning [11], for example. A recent review of the sparse recovery algorithms can be found in [12]. The performance of these algorithms in terms of the approximation quality and the sparsity of the coefficient vector $\mathbf{a}$ depends not only on the signal itself, but also on the overcomplete dictionary $\mathcal{D}$. Once the algorithms are used on a specific class of signals $\mathbf{y}$, we easily understand that not all dictionaries provide the same approximation performance. There exist dictionaries that are more likely to lead to sparse solutions than others. These are the dictionaries that include atoms explaining best the causes of the target data set. It is exactly the goal of dictionary learning methods to find such optimized dictionaries.

DICTIONARY LEARNING METHODS

The research in dictionary learning has followed three main directions that correspond to three categories of algorithms: i) the probabilistic learning methods; ii) the learning methods based on clustering or vector quantization; and iii) the methods for learning dictionaries with a particular construction. This construction is typically driven by priors on the structure of the data or to the target usage of the learned dictionary. This section presents the main principles of representative algorithms in each of these three dictionary learning categories.

PROBABILISTIC METHODS

Representation and coding of images have always been a great challenge for researchers because of the high dimensionality and complex statistics of such signals. Thus, it is not surprising that one of the earliest works addressing the problem of learning overcomplete dictionaries appeared exactly for image representation. In 1997, Olshausen and Field [4] developed a maximum likelihood (ML) dictionary learning method for natural images under the sparse approximation assumption. Their method is called sparse coding. The goal of the work was to give evidence that the coding in the primary visual area V1 in the human cortex probably follows a sparse coding model. In other words, their hypothesis was that the visual cortex reduces the high-dimensional representation of each retinal image into a reduced space defined by the receptive fields of a small number of active neurons. Given the linear generative image model in (1), the objective of the ML

learning method is to maximize the likelihood that natural images have efficient, sparse representations in a redundant dictionary given by the matrix $\mathbf{\Phi}$. Formally, the goal of learning is to find the overcomplete dictionary $\mathbf{\Phi}^*$ such that

$$\begin{aligned}\mathbf{\Phi}^* &= \arg\max_{\mathbf{\Phi}}[\log P(\mathbf{y}|\mathbf{\Phi})] \\ &= \arg\max_{\Phi}\left[\log\int_{\mathbf{a}} P(\mathbf{y}|\mathbf{a},\mathbf{\Phi})P(\mathbf{a})d\mathbf{a}\right]. \end{aligned} \tag{4}$$

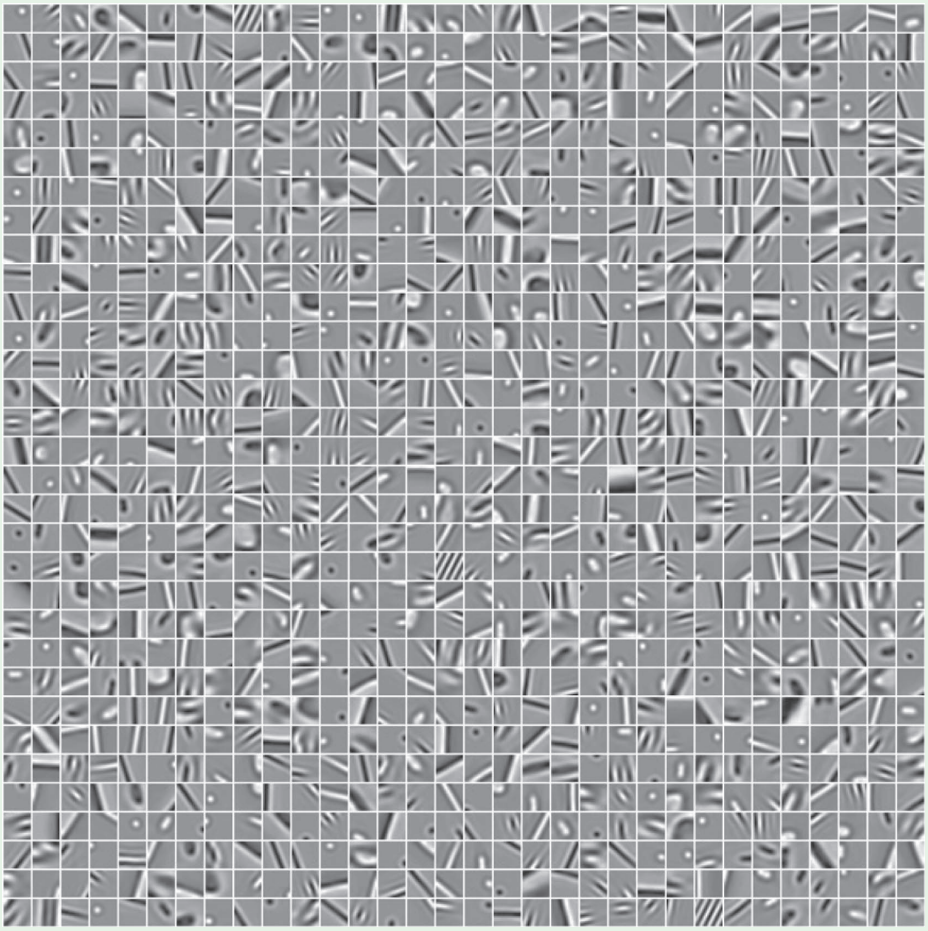
For high-dimensional vectors $\mathbf{a}$, the computation of the integral in (4) is extremely difficult. To simplify the problem and solve the ML optimization, Olshausen and Field introduced two main assumptions. First, the distribution $P(\mathbf{a})$ is assumed to be a product of Laplacian distributions for each coefficient, or equivalently that the coefficients a_i are independent. The Laplacian distribution is peaked at zero and presents a heavy tail, which nicely fits the probability distributions of coefficients a_i when the signal decomposition is sparse. Choosing the prior distribution on $\mathbf{a}$ to be tightly peaked at zero permits to approximate the integral in (4) only by its value at the maximum of $P(\mathbf{y}|\mathbf{a},\mathbf{\Phi})P(\mathbf{a})$. The second assumption is that the approximation noise η can be modeled as a Gaussian zero-mean noise. Under these two assumptions, the optimization problem in (4) can be reduced to an energy minimization problem

$$\begin{aligned}\mathbf{\Phi}^* &= \arg\min_{\Phi,\mathbf{a}} E(\mathbf{y},\mathbf{a}|\mathbf{\Phi}) \\ &= \arg\min_{\mathbf{\Phi},\,\mathbf{a}}[\|\mathbf{y}-\mathbf{\Phi}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1], \end{aligned} \tag{5}$$

where the energy function is defined as $E(\mathbf{y},\mathbf{a}|\mathbf{\Phi}) = -\log[P(\mathbf{y}|\mathbf{a},\mathbf{\Phi})P(\mathbf{a})]$. To take into account statistics of different images, the dictionary is usually learned by minimizing an average energy $\langle E(\mathbf{y}_i,\mathbf{a}_i|\mathbf{\Phi})\rangle$ over a set of randomly chosen images $\{\mathbf{y}_i\}$. The casted optimization problem can be solved by iterating between two steps. In the first step, $\mathbf{\Phi}$ is kept constant and the energy function is minimized with respect to a set of coefficient vectors $\{\mathbf{a}_i\}$. This inference step is essentially the sparse approximation problem defined by (3). It can be solved by convex optimization for each $\mathbf{y}_i$. The second step is called the learning step. It keeps the coefficients $\{\mathbf{a}_i\}$ constant, while performing the gradient descent on $\mathbf{\Phi}$ to minimize the average energy. Since the first step is computationally expensive, the probabilistic dictionary learning methods usually work with small image patches, i.e., the size of $\mathbf{y}_i$ is typically below 32×32 pixels. The algorithm iterates between the sparse approximation and the dictionary learning steps until convergence. This alternating optimization process does not necessarily find the global optimum solution of the considered problem. However, it has been shown to converge to a dictionary with atoms that resemble the receptive fields of simple neurons in V1. A ten times overcomplete dictionary learned on 16×16 image patches [13] is illustrated in Figure 2. The biggest part of the learned dictionary consists of atoms that are localized, oriented and bandpass. Interestingly, these types of features represent well the oriented edges in images. Moreover, the dictionary contains atoms that are center-surround and gratings, which better approximate textures in images. Dictionary learning here clearly meets our objectives: it identifies the most important building blocks in natural images, which permit to approximate the signals by a sparse series of causes or components. It also permits to build an interesting bridge between sparse image representation methods and the properties of the human visual cortex, which is undoubtedly a very efficient encoder for natural images.

The probabilistic inference approach in overcomplete dictionary learning has subsequently been adopted by other researchers. The two-step optimization structure has been preserved in most of these works, and the modifications usually appeared in either the sparse approximation step, or the dictionary update step, or in both. For example, the method of optimal directions (MOD) algorithm [14] optimizes iteratively the same objective ML function as in sparse coding. However, it uses the OMP algorithm to find a sparse vector $\mathbf{a}$ and introduces a closed-form solution for the dictionary update step. The two modifications render the MOD approach faster compared to the method of Olshausen and Field, but still does not guarantee to find the globally optimal solution. Moreover, it is not guaranteed to converge, neither to decrease the objective function at each iteration. The maximum a posteriori (MAP) dictionary learning method [15] belongs also to the family of two-step iterative algorithms based on probabilistic inference. Instead of maximizing the likelihood $P(\mathbf{y}|\mathbf{\Phi})$, the MAP method maximizes the posterior probability $P(\mathbf{\Phi}, \mathbf{a}|\mathbf{y})$. This essentially reduces to the same two-step algorithm, where dictionary update includes an additional constraint on the dictionary that can be for example the unit Frobenius norm of $\mathbf{\Phi}$ or the unit l_2 norm of all atoms in the dictionary. The sparse approximation step is here performed with FOCUSS [10]. Finally, the majorization method can also be used to minimize the objective function in both sparse approximation and dictionary update steps [16]. The sparse approximation step then reduces to the use of an iterative thresholding algorithm.

Naturally, the two assumptions introduced in the sparse coding method represent constraints that can be modified or even removed to learn better dictionaries or to extend the method to other signal models. Lewicki and Sejnowski have modified the first assumption and proposed a new way to approximate the integral in (4) with a Gaussian around the posterior estimate of the coefficient vector $\mathbf{a}$. This changes the update rule in the learning step [17]. They have shown that the ML dictionary learning method with the new estimate for $P(\mathbf{y}|\mathbf{\Phi})$ learns dictionaries that improve the efficiency of sparse coding. The efficiency is measured here in terms of the entropy of data given the overcomplete dictionary. This method actually represents a generalization of the independent component analysis (ICA) method to overcomplete dictionaries. On the other hand, one can also modify the second assumption on the existence of Gaussian noise. When the noise term is zero (i.e., $\eta = 0$), the sparse representation step



[FIG2] Overcomplete dictionary learned with sparse coding from a large data set of 16 × 16 natural image patches. [Used with permission from SPIE (B. A. Olshausen, C. F. Cadieu, and D. K. Warland, "Learning real and complex overcomplete representations from the statistics of natural images," *Proc. SPIE,* vol. 7446, 2009).]

is performed using the exact ℓ_1 sparse optimization [18]. In general, convergence is not guaranteed for the ℓ_1-constrained methods, although it can be proved in some conditions [19]. One could also introduce smoother sparsity priors to obtain more stable solutions. For example, the ℓ_1 constraint is replaced by a Kullback-Leibler (KL) divergence in [20], which shows that the sparsity is preserved, while the KL-regularization leads to efficient convex inference and stable coefficient vectors (i.e., stable representations).

Finally, fast online learning algorithms have been proposed recently [19]. As most of the learning methods based on alternate solutions of the sparse coding and dictionary updates steps use the whole training set at each iteration, these algorithms become rapidly expensive when the data set is large and mostly inappropriate for dynamic systems where data evolve over time. Online learning overcomes this limitation by increasing progressively the training set. An alternate optimization of sparse coding and dictionary update steps is performed with a subset of the training data. This subset is then augmented with a new training sample. The alternate optimization is run again on the new training data with the outcome of the previous iteration as initialization. The online algorithm repeats these iterations until all training data have been used. The resulting solution converges with efficient learning performance and drastically lower computational complexity.

CLUSTERING-BASED METHODS

A slightly different family of dictionary learning techniques is based on vector quantization (VQ) achieved by K-means clustering. The VQ approach for dictionary learning has been first proposed by Schmid-Saugeon and Zakhor in MP-based video coding [21]. Their algorithm optimizes a dictionary given a set of image patches by first grouping patterns such that their distance to a given atom is minimal, and then by updating the atom such that the overall distance in the group of patterns is minimal. The implicit assumption here is that each patch can be represented by a single atom with a coefficient equal to one, which reduces the learning procedure to a K-means clustering. Since each patch is represented by only one atom, the sparse approximation step becomes trivial.

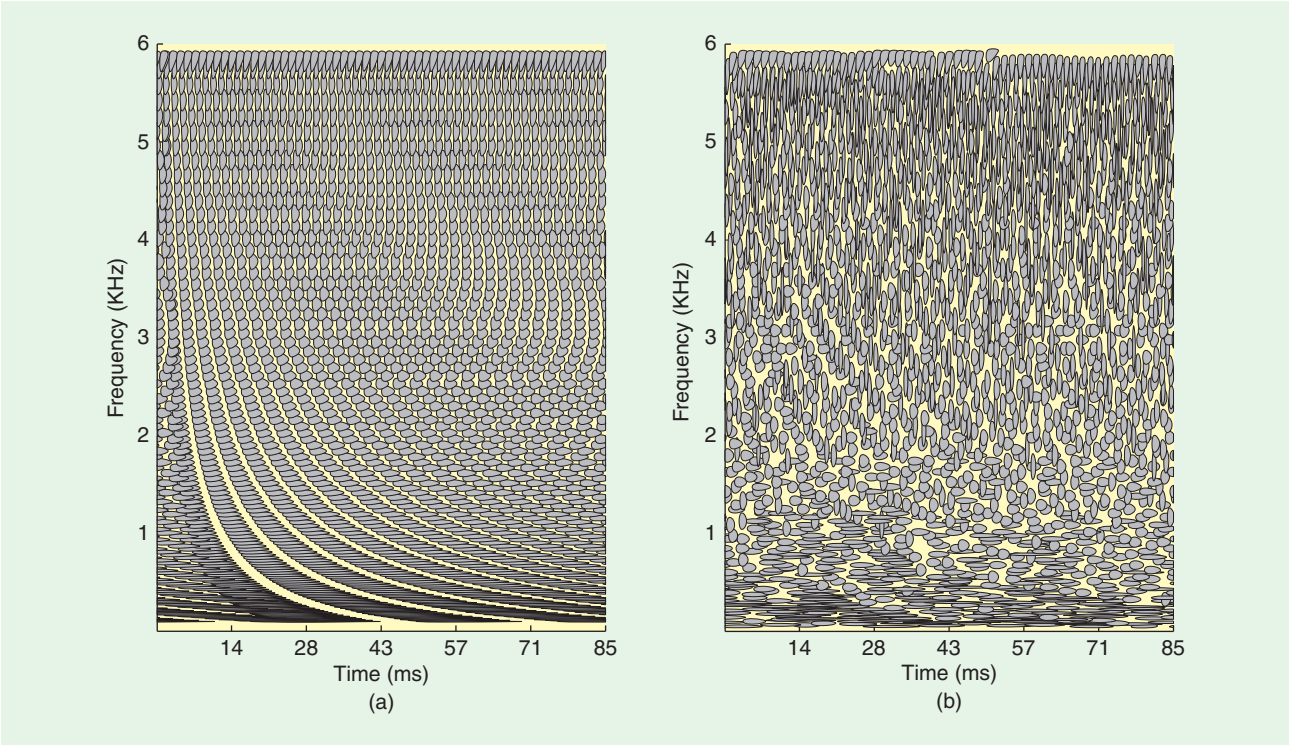
A generalization of the K-means algorithm for dictionary learning, called the K-SVD algorithm, has been proposed by Aharon et al. [5]. After the sparse approximation step with OMP, the dictionary update is performed by sequentially updating each column of $\mathbf{\Phi}$ using a singular value decomposition (SVD) to minimize the approximation error. The update step is hence a generalized K-means algorithm since each patch can be represented by multiple atoms and with different weights. This algorithm is not guaranteed to converge in general. However, in practice, dictionaries learned with K-SVD have shown excellent performance in image denoising. Figure 1(f) shows how K-SVD finds almost all 20 letters as the underlying causes of noisy letter samples. In this example, the sparse approximation step has been implemented by OMP, so it converges to a local minimum where letters "R" and "P" are not successfully separated.

LEARNING DICTIONARIES WITH SPECIFIC STRUCTURES

Many applications do not necessitate general forms of dictionary atoms but can rather benefit from a dictionary that is a set of parametric functions. In contrary to the generic dictionaries above, the advantages of parametric dictionaries reside in the short description of the atoms. The generating function and the atom parameters are sufficient for building the dictionary functions. This is quite beneficial in terms of memory requirements, communication costs or implementation complexity in practical applications.

Such generating functions can be built on prior knowledge about the form of signal causes or the target task. For example, some perceptual criteria can drive the choice of the generating functions in building the dictionary atoms, when the objective is to reconstruct data that are eventually perceived by the human auditory or visual system. Learning in such parametric dictionaries reduces to the problem of learning the parameters for one or more generating functions. Equivalently, it consists in finding a good discrete parametrization that leads to efficient sparse signal approximations. Parametric dictionaries are usually structured, so one can enforce some desired dictionary properties during learning such as minimal dictionary coherence; for example, one can optimize a parametric dictionary such that it gets close to an equiangular tight frame (ETF). In [22], a dictionary for audio signals is learned based on a Gammatone generating function, which has been shown to have similarities with the human auditory system. The method learns a dictionary with good coherence properties, which tiles the time-frequency plane more uniformly than the original Gammatone filter bank. The resulting dictionaries are shown in Figure 3.

Priors or models of the underlying signal causes can also lead to imposing properties such as shift-invariance [23] or multiscale [24] characteristics of the atoms. Such constraints typically limit the search space in the dictionary optimization problem, but lead to more accurate or task-friendly representations. Similarly, the target dictionary might present specific characteristics in particular recovery problems, such as a block-based structure [25], or orthogonality between subspaces [26]. These requirements



[FIG3] Time-frequency representations of structured dictionaries for audio signal representation. It can be observed that the learned dictionary (b) provides a more uniform tiling of the time-frequency plane than the original dictionary (a) designed from a Gammatone filter bank. This corresponds to a smaller coherence than in the original dictionary. Figure used with permission from [22].

considerably affect the design of learning strategies as well as the approximation performance.

APPLICATIONS OF DICTIONARY LEARNING

Dictionary learning for sparse signal approximation has found successful applications in several domains. For example, it has been applied to medical imaging and representation of audio and visual data. We overview here some of the main applications in these directions.

MEDICAL IMAGING

Dictionary learning has the interesting potential to reveal a priori unknown statistics of certain types of signals captured by different measurement devices. An important example are medical signals, such as electroencephalogram (EEG), electrocardiography (ECG), magnetic resonance imaging (MRI), functional MRI (fMRI), and ultrasound tomography (UST) where different physical causes produce the observed signals. It is crucial, however, that representation, denoising, and analysis of these signals are performed in the right signal subspace, such that the underlying physical causes of the observed signals can be identified. Learning of components in ECG signals facilitates ventricular cancellation and atrial modeling in the ECG of patients suffering from atrial fibrillation [27]. Overcomplete dictionaries learned from MRI scans of breast tissues have been shown to provide an excellent representation space for reconstructing images of breast tissue obtained by the UST scanner [28], which drastically reduces the imaging cost compared to MRI. Moreover, standard breast screening techniques, such as the X-ray projection mammography and computed tomography can potentially exploit highly sparse representations in learned dictionaries [29]. Analysis of other signals, such as neural signals obtained by EEG, multielectrode arrays, or two-photon microscopy could also largely benefit from adapted representations obtained by dictionary learning methods.

REPRESENTATION OF AUDIO AND VISUAL DATA

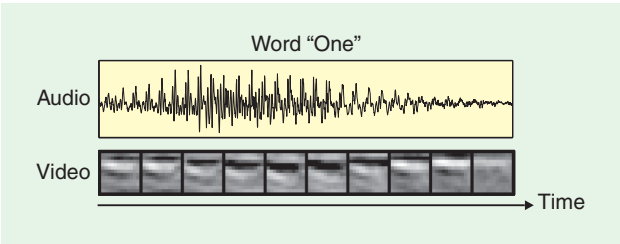
Dictionary learning has introduced significant progress in denoising of speech [30] and images [5], and in audio coding and source separation [16], [31], where it is very important to capture the underlying causes or the most important constitutive components of the target signals. The probabilistic dictionary learning framework has been also proposed for modeling natural videos. These methods explicitly model the separation of the invariant signal part given by the image content and the varying part represented by the motion. Learning under these separation constraints can be achieved using the bilinear model [32], [33], or the phase coding model [34]. In addition to learning the dictionary elements for the visual content, these methods also learn the sparse components of the invariant part (e.g., translational motion).

There exist many examples in nature where a physical process is observed or measured under different conditions. This results in sets of correlated signals whose common part corresponds to the underlying physical cause. However, different observation conditions introduce variability in the measured signals, such that the common cause is usually difficult to extract. Dictionary learning methods based on ML and MAP can be extended by modifying the objective function such that the learning procedures identify the proper subspace for the joint analysis of multiple signals. This permits to learn the underlying causes under different observation conditions. Such modified learning procedures have been applied to audio-visual signals [35] and to multiview imaging [36]. The synchrony between audio and visual signals is exploited in [35] to extract and learn the components of their generating cause that is human speech. A multimodal dictionary is learned with elements that have an audio part and a video part corresponding to the movement of the lips that generate the audio signal. An example of the learned atom for the word "one" 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.

In stereo vision, the same three-dimensional (3-D) scene is observed from different viewpoints, which produce correlated multiview images. Due to the projective properties of light rays, the correlation between multiview images has to comply with epipolar geometry constraints. Dictionaries can be learned such that they efficiently describe the content of natural stereo images and simultaneously permit to capture the geometric correlation between multiview images [36]. The correlation between images is modeled by the local atom transforms, which is made feasible by the use of geometric dictionaries built on scaling, rotation and shifts of a generating function. Learning is based on an ML objective that includes the probability that left image $\mathbf{y}_L$ and right image $\mathbf{y}_R$ are well represented by a dictionary Φ, and the probability that corresponding image components in different views satisfy the epipolar constraint

$$\mathbf{\Phi}^* = \arg\max_{\Phi}[\log P(\mathbf{y}_L, \mathbf{y}_R, D = 0|\mathbf{\Phi})], \tag{6}$$

where $D = 0$ denotes the event when the epipolar geometry is satisfied. This ML objective leads to an energy minimization learning method, where the energy function has three terms: image approximation error term (for both stereo images), the sparsity term, and the multiview geometry term. Dictionary learning is performed in two steps: sparse approximation step



[FIG4] Learned audio-visual atom representing the word "one." Figure used with permission from [35].

with the multiview MP algorithm [36], and dictionary update step with the conjugate gradient method. An illustrative example of a sparse decomposition of two stereo image patches with three correlated learned stereo atoms is shown in Figure 5. Learned stereo dictionaries can be applied to the joint or distributed coding of multiple correlated views or to the analysis and understanding of the geometry in 3-D scenes [36].

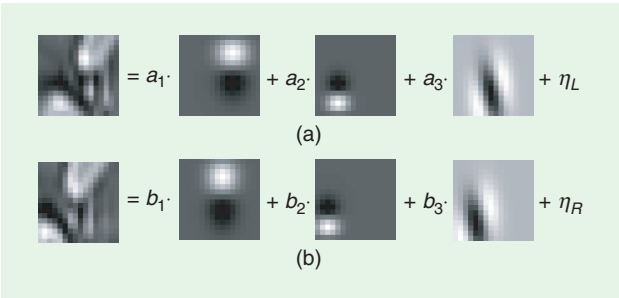
The above illustrations demonstrate the benefits of sparse approximations with learned dictionaries in very diverse applications. One of the main advantages of dictionary learning is that it allows for representing the underlying causes of signals or the main components of data. This is very important for proper understanding and analysis of data that are often the result of noisy measurements of physical processes.

LEARNING FOR CLASSIFICATION

DIMENSIONALITY REDUCTION AND CLASSIFICATION

Dimensionality reduction has been described so far from a pure approximation perspective, where a subspace or a dictionary is computed to explain the observed data with a sparse representation. Alternatively, dimensionality reduction can also target the analysis of data with the objective of distinguishing between different classes of signals or physical processes and to beat the curse of dimensionality and scale. Low-dimensional problems generally involve less complex and more efficient algorithms. The reduced subspace emphasizes in this case the most relevant information in the signal and permits to distinguish between different classes of observations.

Dimensionality reduction for signal analysis finds numerous applications in diverse domains such as sensor networks, computer vision, data mining, machine learning, or information retrieval. We can distinguish two main types of algorithms for computing the reduced subspace: the discriminative methods and the reconstructive methods that are illustrated in Figure 6. The main objective of the discriminative method is to find a mapping or an embedding between the original data space and a reduced dimension subspace, where data can then be efficiently analyzed or classified. This mapping can be either linear (e.g., linear discriminant analysis (LDA) [37]) or nonlinear (e.g., locally linear embedding (LLE) [38], Isomap [39]). The objective of the mapping is to clearly separate the data from different classes in the low-dimensional subspaces. The discriminative methods however aim at pure discrimination objectives and do not necessarily rely on the computation of meaningful features or specific components of the signal. These methods become unfortunately quite vulnerable to noise in the data, to missing data, or to imperfect testing conditions.



[FIG5] Sparse decomposition of a stereo image pair with three correlated learned stereo atoms. (a) Left image and its atoms. (b) Right image and its atoms. Stereo atoms in the two views (three right-most columns) are correlated by local geometric transforms that obey epipolar geometry constraints.

The reconstructive methods try to compute representations that enable analysis and labeling of the data and simultaneously capture its constitutive components to provide robustness to impairments. We focus here on representations that use linear subspaces as opposed to more generic manifold methods. The most common low-rank approximation methods used in signal analysis are based on ICA, PCA, or part-based representations such as nonnegative matrix factorization (NMF) algorithms [40]. The role of a dimensionality reduction algorithm consists here in simplifying the signal to its most meaningful components, such that it can be efficiently characterized in the reduced subspace. For example, PCA maximizes the variance of the data projected on the reduced subspace, which eventually reinforces the discrimination capabilities of the subspace representation. In most reconstructive methods, the projected data are eventually labeled based on nearest neighbor or nearest subspace criteria. However, the basis vectors that define the reduced dimension subspace might unfortunately be holistic, of global support, or with long description length. In the next sections, we describe methods that build linear subspaces from redundant dictionaries of functions with fine adaptation to the data under consideration toward effective signal classification.

SUBSPACE SELECTION FOR CLASSIFICATION

Dimensionality reduction can first be achieved by selecting a subset of functions from a large, fixed dictionary that is used for the analysis of particular signals. These functions then determine a subspace of reduced dimension, where classification can be performed by computing the nearest neighbor points among the projected data. A simple method to build such a subspace consists in modifying the sparse approximation methods described in the previous sections, such that the objective function is augmented with a discrimination term that represents the separability properties of the projection subspace. One can thus select a subset of functions in a dictionary (represented by the matrix $\mathbf{\Phi}$), which approximate the data samples and simultaneously encourage the separability of data in different classes. In other words, the reduced subspace can be computed by solving a problem like

$$\mathbf{a}^* = \arg\min_{\mathbf{a}} [\|\mathbf{y} - \mathbf{\Phi}\mathbf{a}\|_2^2 + \gamma J(\mathbf{\Phi},\mathbf{a})], \tag{7}$$

where the term $J(\mathbf{\Phi},\mathbf{a})$ measures the separability of the different classes when data is represented by atoms in $\mathbf{\Phi}$ and coefficients $\mathbf{a}$. It typically tries to maximize the variance

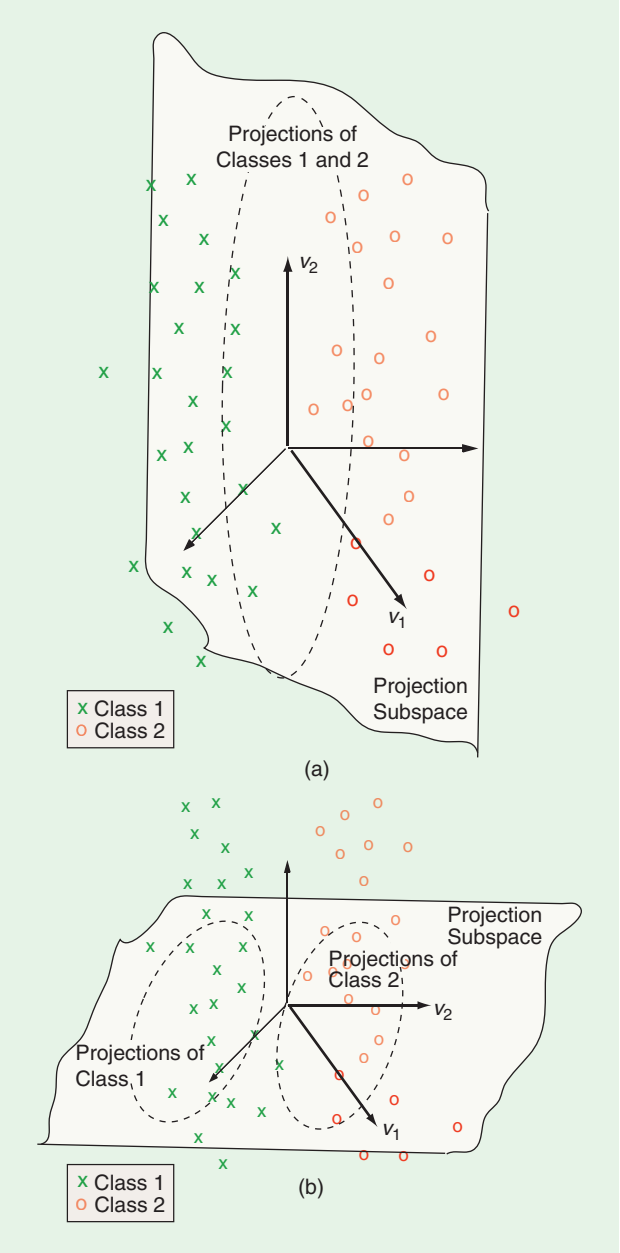
between the active atoms from $\mathbf{\Phi}$ that represent signals in different classes. The reduced subspace used for classification is finally formed by the subset of atoms in $\mathbf{\Phi}$ whose corresponding coefficients in $\mathbf{a}^*$ are nonzero. The subset selection problem can be interpreted as the inference step in the dictionary learning methods when the objective function is modified to include a discriminative term. Finally, the weight parameter γ controls the tradeoff between approximation and classification performance in the reduced dimension subspace. A subset of $\mathbf{\Phi}$ that solves the problem posed in (7) can be determined by iterative supervised atom selection built on OMP for example [41]. The idea mainly consists in selecting greedily the atoms from the dictionary that lead to the best tradeoff between approximation of the training data and discrimination between classes.

The minimum of the joint objective function above can be achieved with several distinct sets of functions: finding the best subspace for classification becomes nontrivial due to the redundancy of the dictionary. However, good subspaces for reconstructive dimensionality reduction are characterized by sparsity properties, where only a few significant components participate in the representation of the data. The method of sparse representation for signal classification in [42] thus explicitly includes sparsity constraints in the dimensionality reduction process. The reduced subspace is determined here by a simultaneous sparse approximation algorithm built on OMP, where the data separability term $J(\mathbf{\Phi},\mathbf{a})$ is given by a Fisher's discrimination criterion used in LDA. The reduced dimensionality subspace is therefore chosen as a compromise between approximation of data within classes, discrimination of data in different classes, and sparsity of the data representation as determined by an optimization problem of the following generic form:

$$\mathbf{a}^* = \arg\min_{\mathbf{a}} [\|\mathbf{y} - \mathbf{\Phi}\mathbf{a}\|_2^2 + \gamma_1 \|\mathbf{a}\|_0 + \gamma_2 J(\mathbf{\Phi},\mathbf{a})]. \tag{8}$$

SUPERVISED DICTIONARY LEARNING

An important advantage of redundant dictionaries for classification is that signal analysis can be performed with functions that are likely to match the data characteristics in different classes of signals. Similarly to data approximation problems, data analysis applications can further benefit from dictionary learning methods. The previous section describes subspace selection methods from predefined dictionaries. However, learning can improve the classification performance, as it leads to a better adaptation of the dictionary by enforcing sparsity in the representation of data in the different classes. The atoms in a dictionary $\mathcal{D}$ that is computed with dictionary learning methods generally capture the most important constitutive components of the signals. They naturally permit to classify the data into the corresponding linear subspace as shown in [5], for example. However, there is no guarantee that the subspace built on a learned dictionary $\mathbf{\Phi}$ is truly optimal for classification, as it targets efficient representation but not necessarily class separability. For example, one may define a set of functions that are good to (sparsely) approximate signals in a face image data set. However, there is no good reason why this same set of functions is also the best one for distinguishing different persons in this data set.



[FIG6] Illustration of dimensionality reduction of a two-class data set, by projection on a linear subspace defined by vectors $(\mathbf{v}_1,\mathbf{v}_2)$. (a) Purely reconstructive methods compute a representative subspace where the projections of the data are close to the original data points. The approximation of data by their projections is optimized, but the classification of the projected data is not trivial. (b) Purely discriminative methods compute the reduced dimensionality subspace so that the classification can be done efficiently from the data projections. Data approximation is quite poor in this case, which results in low robustness to data impairments. The optimal subspace has to be the result of a tradeoff between approximation and separability.

Dictionary learning methods should rather be modified so that they become simultaneously reconstructive (for robustness to noise) and discriminative (for efficient classification with the learned dictionary). The addition of a discriminative term into the dictionary learning algorithms requires supervision, where labels of training data are used to ensure that the data representation is sufficiently different in each class. It can be achieved by modifying the sparse coding step in the learning algorithms, so that it optimizes an objective function that favors the sparsest representation of a given signal and simultaneously the representation that is also the most different from the one of signals in other data classes. The supervised dictionary learning problem can be cast as a mixed formulation that minimizes the average value of the sparse approximation errors over different classes and also enforces discrimination between classes. For example, the dictionary optimization problem can be written as

$$\mathbf{\Phi}^* = \arg\min_{\Phi,\mathrm{a}}[\|\mathbf{y} - \mathbf{\Phi}\mathrm{a}\|_2^2 + \gamma_1\|\mathrm{a}\|_1 + \gamma_2 C(\mathrm{a},\mathbf{\Phi},\theta)], \quad (9)$$

where the function $C(\mathrm{a},\mathbf{\Phi},\theta)$ is a discrimination term that depends on the dictionary, the coefficient vectors, and the parameters θ of the model used for classification. Since the dictionary is learned, alternate inference and learning steps have to be used in solving (9). In contrary, the subspace selection problem in the section "Classification Subspace Selection" is solved only within the inference step. Note that the discrimination term is specific to the chosen classifier through the parameters θ so that the learning problem becomes highly dependent on the classification method and unfortunately nonconvex. Still, it can be solved efficiently by fixed-point continuation methods [43] when the classifier is based on logistic regression methods.

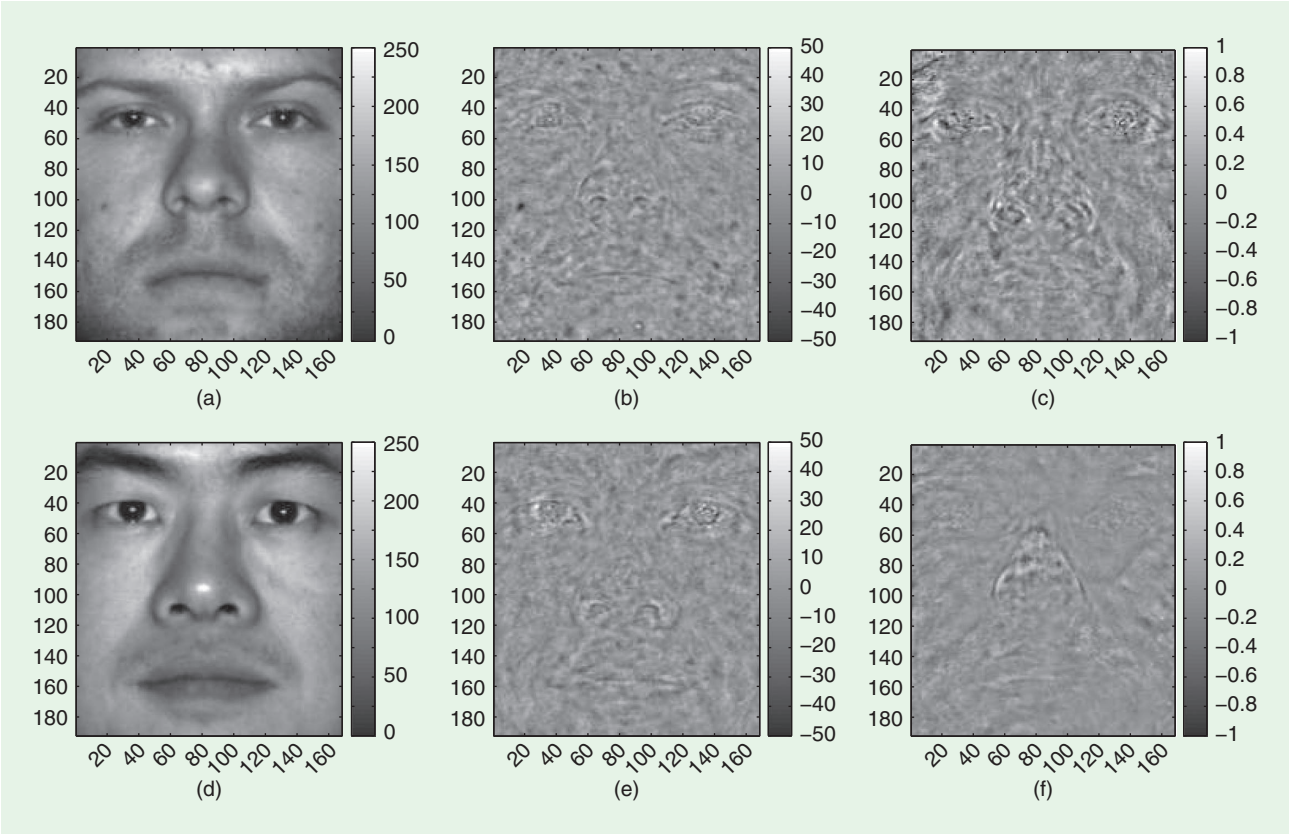
The use of one learned dictionary for all the data classes leads to a straightforward classification stage where the dictionary vectors and the coefficients in the signal representation are used directly to make classification decisions. Alternatively, one may want to improve the discrimination by building a distinctive projection subspace for each data class. Classification is then performed by selecting the subspace that is the nearest to the test signal, or equivalently the subspace that leads to the best representation of the test signal. A simple way to build adaptive dictionaries for each class is to use the signals in the training set for the class dictionary. Sparsity constraints are then rather applied within the classification process, where the sparsest representation of the test signal determines its class label. For example, Wright et al. [44] have proposed a face recognition method that uses training face images as dictionaries and an l_1 sparse optimization method in the classification stage. The authors show that the recognition task can be successfully accomplished even using random features at first. Furthermore, the algorithm is robust to a certain amount of noise due to the sparsity constraints.

It is often preferable, however, to construct adapted dictionaries that can lead to an efficient classification process based on simple subspace projections. The construction of class dictionaries can be performed with learning methods where the sparse coding step in the iterative learning algorithms is modified, so that sparse coding is computed independently within each class. Such a sparse coding stage can be implemented by class-supervised versions of simultaneous pursuit algorithms, for example, where a joint sparse representation of the training data is selected independently in each class. The subsequent dictionary update step further favors the reconstruction of signals with the functions selected in the modified sparse coding step. If the update step is based on an SVD algorithm, it simply leads to a supervised version of the K-SVD algorithm [45], where the K-SVD learning algorithm becomes adapted to classification tasks. As supervised dictionary learning should intuitively lead to subspaces that are good for approximating data in their own class but bad for representing data from any other class, the subspaces can also be computed with a hierarchical process that ensures that features selected in different dictionaries have only a minimal correlation [46]. Alternatively, global softmax discriminative functions can enforce that the learned dictionaries are better for representing data of their classes than data from any other classes. Such a discrimination can be achieved by modifying the dictionary update steps in the learning process with a modified version of MOD/K-SVD algorithm whose role is thus extended to ensuring data separability with the updated dictionary in addition to good approximation properties [47].

Finally, discrimination in dictionary learning can also be achieved by enforcing incoherency between the subspaces that represent data in different classes and not only by minimizing the correlation between the features in different subspaces. It relies on the intuition that some features might be relatively good in representing data in different classes, but several features taken together form a subspace that is mostly good in approximating data from the corresponding class. For example, the subspace formed by noses and eyes of persons in different classes are incoherent, even if these persons have similar eyes or the same nose. With the assumption that the residue of the subspace projection is minimal in the correct class, incoherent subspaces can be designed by an alternate projection method [48]. It builds on the natural conditions that the interplay between features of different classes should be small, while the interaction of training data with features in the correct class should be clearly higher than the interaction with features representing any other class (see Figure 7). With minimal assumptions on the signal models or sparsity features, such a dictionary learning method reaches state-of-the-art performance on a face classification experiment.

CONCLUSIONS

The goal of dimensionality reduction is to find efficient, low-dimensional data representations within the large dimensional space where the observed data lies. This article has presented some of the recent results supporting the idea that these representations are sparse within an overcomplete dictionary of atoms or subspaces. In this context, the methods for dictionary



[FIG7] Images of two subjects: Parts (a) and (d) show original projections onto the span of features from their own class. Parts (b) and (e) show projections onto the span of features of the (c) and (f) wrong class. The representation of signal with the subspace of the proper class is clearly more relevant than the representation with a subspace of another class: the scales and positions of projection components are close to the original signal. Figure used with permission from [48].

learning have much to offer since they are able to adapt the data representation to the underlying causes of the observations. We have given a broad overview of the main dictionary learning algorithms and shown their usage in various applications, such as audio-visual coding and stereo image approximation. We have also discussed the discriminative power of sparse representations and outlined the large potential benefits of dictionary learning in classification and face recognition applications.

Many challenges are still open in dictionary learning. Understanding the underlying causes of signals or the relevant information in observations becomes more challenging when the training samples are imperfect. In many applications, the training samples are noisy, distorted by the sensing process, or simply incomplete like in the case of occlusions in multiview imaging. The last example particularly makes us question the validity of linear representation models in vision where we usually encounter nonlinearities such as occlusions. Linear models also become invalid in advanced applications like medical imaging where the acquisition methods are typically nonlinear. In all these situations, dictionary learning still faces critical research questions. Similarly, signal analysis may require more complex models than linear subspaces for efficient classification. One can build dictionaries to be used in the definition of manifold models or graph-based representations that could potentially handle transformation-invariant classifications problems. In general, dictionaries offer a very flexible and powerful way to represent relevant information in high-dimensional signals. However, the proper modeling of the complex underlying causes of observations poses many exciting questions about the proper construction of these dictionaries.

ACKNOWLEDGMENTS

This work has been partly supported by the Swiss National Science Foundation under grant PBELP2-127847. The authors would like to thank the editors and anonymous reviewers for their insightful comments that have greatly helped to improve the quality of the work. Special thanks go also to Bruno Olshausen, Pierre Vandergheynst, Sofia Karygianni, Effrosyni Kokiopoulou, and Elif Vural for the careful reading of the manuscript and helpful feedbacks. We would also like to thank Gianluca Monaci, Fritz Sommer, Laurent Daudet, and Karin Schnass for providing some of the figures used in this article.

AUTHORS

Ivana Tošić (ivana@berkeley.edu) received the Dipl.Ing. degree in telecommunications from the University of Niš, Serbia, and the Ph.D. degree in computer and communication sciences from the Swiss Federal Institute of Technology (EPFL), Lausanne,

Switzerland. She is currently a postdoctoral researcher at the Redwood Center for Theoretical Neuroscience, University of California at Berkeley, United States, where she works on the intersection of image processing and computational neuroscience domains. She was awarded the Swiss National Science Foundation fellowship for prospective researchers. Her research interests include representation and coding of the plenoptic function, distributed source coding, binocular vision, and 3-D object representation. She is a Member of the IEEE.

Pascal Frossard (pascal.frossard@epfl.ch) received the M.S. and Ph.D. degrees, both in electrical engineering, from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1997 and 2000, respectively. Between 2001 and 2003, he was a member of the research staff at the IBM T.J. Watson Research Center, Yorktown Heights, New York. Since 2003, he has been an assistant professor at EPFL, where he heads the Signal Processing Laboratory (LTS4). His research interests include image representation and coding, visual information analysis, distributed image processing and communications, and media streaming systems. He is a Senior Member of the IEEE.

REFERENCES

[1] I. Jolliffe, *Principal Component Analysis.* Springer-Verlag: New York, 1986.

[2] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput*., vol. 7, no. 6, pp. 1129–1159, 1995.

[3] H. B. Barlow, "Possible principles underlying the transformations of sensory messages," *in Sensory Communication*, W. Rosenblith, Ed. Cambridge, MA: MIT Press, 1961, ch. 13, pp. 217–234.

[4] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vis. Res*., vol. 37, no. 23, pp. 3311–3325, 1997.

[5] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process*., vol. 54, no. 11, pp. 4311–4322, 2006.

[6] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process*., vol. 41, no. 12, pp. 3397–3415, 1993.

[7] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[8] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput*., vol. 20, no. 1, pp. 33–61, 1999.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. B (Method.)*, vol. 58, no. 1, pp. 267–288, 1996.

[10] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process*., vol. 45, no. 3, pp. 600–616, 1997.

[11] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process*., vol. 52, no. 8, pp. 2153–2164, 2004.

[12] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems," *Proc. IEEE* (Special Issue on Applications of Sparse Representation and Compressive Sensing), to be published.

[13] B. A. Olshausen, C. F. Cadieu, and D. K. Warland, "Learning real and complex overcomplete representations from the statistics of natural images," *Proc. SPIE*, vol. 7446, 2009.

[14] K. Engan, S. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," *in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'99)*, 1999.

[15] K. Kreutz-Delgado, J. Murray, B. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comput*., vol. 15, no. 2, pp. 349–396, 2003.

[16] M. Yaghoobi, T. Blumensath, and M. E. Davies, "Dictionary learning for sparse approximations with the majorization method," *IEEE Trans. Signal Process*., vol. 57, no. 6, pp. 2178–2191, 2009.

[17] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Comput*., vol. 12, no. 2, pp. 337–365, 2000.

[18] M. D. Plumbley, "Dictionary learning for L1-exact sparse coding," *Lect. Notes Comput. Sci*., vol. 4666, pp. 406–413, 2007.

[19] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.,* to be published.

[20] D. Bradley and J. Bagnell, "Differentiable sparse coding," *Proc. NIPS*, vol. 11, pp. 19–60, Jan. 2009.

[21] P. Schmid-Saugeon and A. Zakhor, "Dictionary design for matching pursuit and application to motion-compensated video coding," *IEEE Trans. Circuits Syst. Video Technol*., vol. 14, no. 6, pp. 880–886, 2004.

[22] M. Yaghoobi, L. Daudet, and M. E. Davies, "Parametric dictionary design for sparse coding," *IEEE Trans. Signal Process*., vol. 57, no. 12, pp. 4800–4810, 2009.

[23] B. Mailhé, S. Lesage, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Shift-invariant dictionary learning for sparse representations: Extending K-SVD," in *Proc. European Signal Processing Conf*., vol. 4, 2008.

[24] P. Sallee and B. A. Olshausen, "Learning sparse multiscale image representations," in *Proc. Conf. Neural Information Processing Systems*, 2003.

[25] K. Engan, K. Skretting, and J. Husřy, "Family of iterative LS-based dictionary learning algorithms, ILS-DLA, for sparse signal representation," *Dig. Signal Process*., vol. 17, no. 1, pp. 32–49, 2007.

[26] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, 2003.

[27] 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 *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'09)*, 2009.

[28] I. Tošić, I. Jovanović, P. Frossard, M. Vetterli, and N. Durić, "Ultrasound tomography with learned dictionaries," *in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'10)*, 2010.

[29] 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 *Proc. Soc. Photo-Optical Instrumentation Engineers Conf. Series (SPIE'09)*, vol. 7263, 2009.

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

[31] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: From coding to source separation," *Proc. IEEE*, no. 99, pp. 1–11, 2009.

[32] D. B. Grimes and R. P. N. Rao, "Bilinear sparse coding for invariant vision," *Neural Comput*., vol. 17, no. 1, pp. 47–73, 2005.

[33] B. A. Olshausen, C. Cadieu, B. J. Culpepper, and D. K. Warland, "Bilinear models of natural images," in *Proc. SPIE Conf. Human Vision and Electronic Imaging*, 2007.

[34] C. Cadieu and B. A. Olshausen, "Learning transformational invariants from time-varying natural images," in *Proc. Conf. Neural Information Processing Systems*, 2008.

[35] G. Monaci, P. Vandergheynst, and F. T. Sommer, "Learning bimodal structure in audio-visual data," *IEEE Trans. Neural Netw*., vol. 20, no. 12, pp. 1898–1910, 2009.

[36] I. Tošić and P. Frossard, "Dictionary learning for stereo image representation," *IEEE Trans. Image Process.,* to be published.

[37] A. Webb, *Statistical Pattern Recognition*, 2nd ed. Hoboken, NJ: Wiley, 2002.

[38] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[39] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[40] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, pp. 11–126, 1994.

[41] E. Kokiopoulou and P. Frossard, "Semantic coding by supervised dimensionality reduction," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 806–818, 2008.

[42] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Proc. Conf. Neural Information Processing Systems*, 2007.

[43] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Supervised dictionary learning," in *Proc. Conf. Neural Information Processing Systems*, 2008.

[44] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 31, no. 2, pp. 210–227, 2009.

[45] F. Rodriguez and G. Sapiro, "Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries," *IMA Preprint*, 2007.

[46] A. Destrero, C. De Mol, F. Odone, and A. Verri, "A sparsity-enforcing method for learning face features," *IEEE Trans. Image Process*., vol. 18, no. 1, pp. 188–201, 2009.

[47] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[48] K. Schnass and P. Vandergheynst, "A union of incoherent spaces model for classification," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'10)*, 2010.

[SP]