SCHOOL OF ENGINEERING - STI
SIGNAL PROCESSING INSTITUTE
Effrosyni Kokiopoulou and Pascal Frossard

CH-1015 LAUSANNE

Telephone: +4121 6932601
Telefax: +4121 6937600
e-mail: {effrosyni.kokiopoulou,pascal.frossard}@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

DIMENSIONALITY REDUCTION WITH SIMULTANEOUS SPARSE APPROXIMATIONS

Effrosyni Kokiopoulou and Pascal Frossard

Swiss Federal Institute of Technology Lausanne (EPFL)

Signal Processing Institute Technical Report

TR-ITS-2006.010

October 21st, 2006

Part of this work has been submitted to IEEE TMM.

This work has been supported by the Swiss NSF, under grants PP-002-68737, and NCCR IM2.

Dimensionality reduction with simultaneous sparse approximations

Effrosyni Kokiopoulou and Pascal Frossard
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Institute - ITS
CH- 1015 Lausanne, Switzerland
{effrosyni.kokiopoulou,pascal.frossard}@epfl.ch

Abstract

We propose a dimensionality reduction method for structured signals and its application in classification. The training phase implements a learning process that forms a parts-based representation of signals. Signals are jointly represented in a common subspace extracted from a redundant dictionary of basis functions, using greedy pursuit algorithms for simultaneous sparse approximations. A small set of basis functions is generally sufficient to characterize a particular signal, and distinguish it from its peers in classification tasks. The dimensionality reduction method is further extended into a supervised algorithm, which enforces the separability between classes, and provides improved classification performances. Interestingly, the proposed algorithms are generic in terms of target signals, and dictionary functions. The design of the dictionary stays particularly flexible, which allows for a direct control on the characteristics of the basis functions that can incorporate a priori and application-driven knowledge into the basis vectors during the learning process. We compare our dimensionality reduction method with Non-negative Matrix Factorization (NMF) and its variants, in the context of handwritten digit image recognition and face recognition. The experimental results suggest that the proposed dimensionality reduction method is competitive with NMF in terms of classification error rate, but advantageously provides meaningful features with high discriminant value.

Index Terms

Dimensionality Reduction, Redundant Dictionaries, Simultaneous Sparse Approximation.

I. Introduction

Recent years have witnessed a large volume of high dimensional multimedia data. It becomes increasingly important to design effective algorithms for pattern analysis and knowledge discovery from the data, in order to respond to the various information requests from diverse users and applications. A pattern of interest is usually observed in a high dimensional ambient space but it is typically of much lower intrinsic dimension. For instance, all the possible appearances of a facial image span only a small part of the high dimensional image space. The purpose of dimensionality reduction techniques is exactly to discover the intrinsic dimension of the data and to extract the low dimensional meaningful features that can accurately characterize the useful information.

Subspace analysis helps to reveal the latent low dimensional structures from the observed high dimensional data. It simply consists in computing the subspace of reduced dimension, which best characterize the relevant information contained in the data of interest. Non-negative Matrix Factorization (NMF) [1] and Principal Component Analysis (PCA) [2] are certainly among the most popular subspace methods for dimensionality reduction. NMF methods has been proposed for learning a parts-based representation, and advantageously provides sparser, spatially localized and therefore more interpretable basis vectors than those computed by PCA-based algorithms, which are holistic and of global support. Recently, a lot of variants of NMF (see e.g., [3], [4] and references therein) have been proposed in order to provide more control over the properties of the basis vectors and/or coefficients vectors, by introducing additional (possibly non convex) constraints into the NMF optimization problem. However, in some cases, this results in sophisticated non-convex optimization problems that are hard to solve in practice.

In this paper, we propose a subspace method which formulates the dimensionality reduction problem as a matrix factorization problem, where the basis vectors are extracted from a redundant dictionary of localized basis functions.

This work has been partly supported by the Swiss National Science Foundation, under grants PP-002-68737, and NCCR IM2.

The flexibility in the design of the dictionary provides direct control on the shape and the properties of the basis functions, such as spatial locality and sparse support. Moreover, it provides naturally the potential to incorporate a priori and application-driven knowledge into the learning process, without resorting to sophisticated constraints. Our dimensionality reduction method attempts to solve the factorization problem using the Simultaneous Orthogonal Matching Pursuit (SOMP) [5] algorithm, which has been previously proposed in a different context for simultaneous sparse approximation of signals. SOMP is a greedy suboptimal algorithm which selects in each step that basis function from the dictionary that will provide the largest reduction of the approximation error.

We further extend the proposed method to classification problems, and we propose a supervised dimensionality reduction that exploits the available class labels information. We present a variant of the SOMP algorithm that encourages the separability between classes. In particular, the selection of the basis function from the dictionary is now driven by a trade-off between the approximation error and class separability. We use the ratio of inter-class over the intra-class variance as a class separability cost function. This modification of the basis function selection step, results into features with more discriminating value than the unsupervised dimensionality reduction algorithm. We analyze the properties of the supervised greedy decomposition algorithm, whose convergence rate is obviously penalized by the class separability constraint. Experimental results in the context of face and digit recognition demonstrate the efficiency of the supervised dimensionality reduction algorithm, which is competitive with the NMF-based methods. When the size of the subspaces increases, SOMP algorithms even outperform NMF solutions, and advantageously provide meaningful features with high discriminant value.

The rest of the paper is organized as follows. In Section II we review the related work and discuss briefly the standard NMF algorithm and the local NMF, which is one of its most popular variants. In Section III, we introduce our subspace method for dimensionality reduction using redundant dictionaries, and in Section IV we discuss the supervised method and establish its convergence properties. Finally, Section V provides experimental results that demonstrate the properties of the proposed schemes, and compare them with NMF and its variants in the context of face and digit recognition.

II. Related Work

Dimensionality reduction is a very broad concept which encompasses numerous methods proposed in the literature. One may distinguish though three main families of methods (a) linear methods (e.g., LPP [6], ONPP [7] etc), (b) nonlinear methods (e.g., LLE [8], Laplacian Eigenmaps [9], Isomap [10] etc) and (c) low rank approximation methods (e.g., PCA [2], NMF [1], [11] etc). The first two categories employ a mapping from the high dimensional space to a low dimensional space, which is linear in the former case and nonlinear in the latter case. The third family includes those methods that use a low rank approximation of the data matrix. In other words, they use only a small number of basis vectors to approximate the high dimensional data of interest. The dimensionality reduction method that is proposed in this paper typically belongs to the third category. We now discuss in more details the most popular methods for low rank approximation.

The most popular subspace method for dimensionality reduction is Principal Component Analysis [2]. In PCA, a subspace is constructed from the eigenvectors of the sample covariance matrix and dimensionality reduction is accomplished by discarding the eigenvectors corresponding to its smallest eigenvalues. The obtained basis vectors from PCA are holistic and of global support. However, they generally fail to identify features that are spatially localized. This represents a clear drawback for applications that rely on parts-based representations of data objects, or where the most relevant information is contained in localized features.

Non-negative Matrix Factorization (NMF), introduced in [1], [11], is another popular dimensionality reduction method with empirical success in real life data sets. It has been proposed as a subspace method for a parts-based representation of objects by imposing non-negativity constraints, typical to digital imaging applications. Given a data matrix $S \in \mathbb{R}^{m \times n}$ with non-negative entries, NMF seeks two non-negative factors $W \in R^{m \times r}$ and $H \in R^{r \times n}$ such that

$$S \approx WH. \tag{1}$$

The columns of the matrix W contain the basis vectors and the matrix H contains the corresponding coefficients (or encoding) vectors for the approximation of the columns of S. Consider the generalized Kullback-Leibler (KL)

divergence between X and Y

$$D(X||Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} [x_{ij}\log\frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij}]. \quad (2)$$

The KL divergence is the most popular objective function used in NMF algorithms. In what follows, we describe briefly the standard NMF algorithm and one of its variants, which will be used in our experimental evaluation.

Standard NMF The NMF can be formulated as the following optimization problem

Optimization problem: **NMF**
$\min_{W,H} D(S||WH)$,
subject to
$W, H \geq 0$,
$\sum_{i=1} w_{ij} = 1,\ \forall j$.

A local minimum solution to the above problem can be obtained by iterating the multiplicative rules introduced in [1].

Local NMF Local NMF (LNMF) [3] is a variant of NMF, which tries to enforce the spatial locality of the basis vectors. In particular it differs from the standard NMF by imposing three additional constraints expressed by the following rules: (a) the number of basis components should be minimized, (b) different basis vectors should be as orthogonal as possible and (c) only the most important components are retained. In particular, LNMF can be formulated as the following optimization problem.

Optimization problem: **LNMF**
$\min_{W,H} D(S||WH) + \alpha\sum_{i,j} u_{ij} - \beta\sum_i z_{ii}$,
subject to
$W, H \geq 0$,
$\alpha, \beta > 0$,
$U = W^\top W$,
$Z = HH^\top$.

In the objective function we have introduced the scalars α and β, which are the Lagrange multipliers corresponding to the additional constraints on spatial locality of features. A local minimum solution to the above problem can be obtained by iterating the three multiplicative rules introduced in [3].

Other variants of NMF have also been proposed recently. For example, a sparsity controlled NMF algorithm based on a measure of sparsity that is a combination of the L1 and L2 norm, has been proposed in [4]. Along the same ideas of controlling sparsity of the reduced subspaces, NMF variants using convex programming have been proposed in [12], [13]. Yet another variant of NMF has been presented in [14], where the authors describe an extension of standard NMF by imposing smoothness constraints on the non-negative factors. In particular, they apply their algorithm for the analysis of non-negative spectral data generated from astronomical spectrometers. Finally, in [15], the NMF model is modified by introducing a smoothing symmetric matrix which controls the sparsity of both non-negative factors.

Although the NMF optimization problem is convex with respect to W or H individually, it is however non-convex with respect to both of them. Thus, all algorithms that have been proposed in the literature are not guaranteed to converge to the global minimum and they are prone to local minima. Moreover, it has been observed that they are also sensitive to the initializations of the two non-negative factors. If the initialization is not good it may happen that the algorithm gets trapped in a bad local minimum, which leads to clearly suboptimal performances.

Finally, extension to classification problems have been proposed with supervised variants of NMF, which takes into account class labels information. The authors in [16] and [17] independently propose a supervised NMF algorithm by incorporating the Fisher constraints into the objective function of NMF and they propose multiplicative update rules. Class separability criterion for basis function selection has also been proposed in [18], in the context of face identification.

Algorithm: **SOMP**
Input: Signal matrix $S \in R^{m \times n}$ and `tol`: approximation error tolerance.
Output: Set of selected atoms Ψ, approximation A and residual matrix R.
1. Initialize the residual $R_0 = S$, $\Psi = []$, $t = 1$.
2. Find index γ_t which solves the optimization problem
$\max_{\gamma \in \Gamma} \|R_t^\top \phi_\gamma\|_1$
3. Augment $\Psi = [\Psi,\ \phi_{\gamma_t}]$
4. Compute an orthonormal basis $V = [v_1, \ldots, v_t]$ of the span$\{\Psi\}$.
5. Compute the orthogonal projector $P_t = V_t V_t^\top$ on the span$\{\Psi\}$.
6. Compute the new approximation and residual
$A_t = P_t S$
$R_t = (I - P_t)S$
7. If $\|R\|_F \leq$ `tol`, then stop. Otherwise, increment iteration $t = t + 1$, and go to step (2).

TABLE I
THE SOMP ALGORITHM.

III. DIMENSIONALITY REDUCTION USING SOMP

We assume the existence of a redundant dictionary $\mathcal{D}$ that spans the Hilbert space $\mathcal{H}$ of the signals of interest. Redundancy offers flexibility in the construction of the dictionary, and in general improves the approximation rate, especially for multidimensional signals. A redundant dictionary is an overcomplete basis in the sense that it includes a number of vectors that is larger than the dimension of the subspace. The elements of the dictionary, which are indexed by $\gamma \in \Gamma$ i.e.,

$$\mathcal{D} = \{\phi_\gamma,\ \gamma \in \Gamma\}, \tag{3}$$

are usually called *atoms*. The atoms have unity norm i.e., $\|\phi_\gamma\|_2 = 1,\ \forall \gamma \in \Gamma$, where $\|\cdot\|_2$ denotes the L2 norm. It is important to note that we do not set any particular assumption on the dictionary design, and that the following analysis holds for any redundant dictionary (i.e., overcomplete basis). The only assumption that we make is that the dictionary spans the signal space $\mathcal{H}$.

Then, we consider a signal s_i as an element of $\mathcal{H} \subseteq R^m$. The training data forms a signal matrix

$$S = [s_1, s_2, \ldots, s_n] \in \mathbb{R}^{m \times n}, \tag{4}$$

where s_i denotes the i-th column of S. For dimensionality reduction, our goal is to decompose S in the following form

$$S = \Psi C,\ \Psi \in \mathbb{R}^{m \times r},\ C \in \mathbb{R}^{r \times n}, \tag{5}$$

where Ψ are the basis vectors drawn from the dictionary and C are the corresponding coefficients. In other words, every column of S is represented in the same set of basis functions Ψ using different coefficients. This is a dimensionality reduction step where each signal (column of S) is represented in the subspace spanned by the columns of Ψ, using only $r \ll m$ coefficients.

If the columns of Ψ are spatially localized basis functions then the decomposition given in Eq. (5) results in a parts-based representation. Note that the design of the dictionary determines the properties of Ψ. Therefore, one has direct control on the shape and the properties of the basis functions due to the flexible design of the dictionary. Recall that in NMF and its variants, one has however only implicit control on the properties of the basis functions, which is accomplished via additional constraints that are introduced in the optimization problem.

If we denote by $\|\cdot\|_F$ the Frobenius norm, then we formulate the above problem as the following optimization problem.

Optimization problem: **OPT1**
$\min_{\Psi,C} \|S - \Psi C\|_F^2$
subject to
$\Psi \subseteq \mathcal{D}$.

In order to solve OPT1 one may employ suboptimal algorithms that have been proposed in the context of simultaneous sparse signal approximations [5], [19], [20], [21]. We have chosen to use the Simultaneous Orthogonal Matching Pursuit (SOMP) algorithm [5], since it lends itself as an efficient algorithm for solving OPT1 in practice. Interestingly, an algorithm called M-OMP, which is identical to SOMP, has been independently proposed in [21]. However, for notational convenience we will keep using the term SOMP while referring to any of these two algorithms.

SOMP is a greedy algorithm that extracts a subset Ψ of the dictionary, such that all the columns of S are simultaneously approximated. SOMP is a generalization of Matching Pursuit [22] to the case of simultaneous approximation of several signals. In each step, SOMP greedily selects the atom from the dictionary, which best matches all the residual signals at each iteration. Initially, SOMP sets the residual matrix $R = S$. Once the best matching atom ϕ_γ has been selected, the algorithm updates the residual matrix by projection on its orthogonal complement, i.e.,

$$R = (I - \phi_\gamma \phi_\gamma^\top)S,$$

where $I - \phi_\gamma \phi_\gamma^\top$ is the projector on the orthogonal complement of span$\{\phi_\gamma\}$. The above step will remove the components of ϕ_γ from R.

In the next steps, the algorithm applies the same procedure on the updated residual matrix. Thus, it greedily selects in step t, the best matching atom ϕ_{γ_t} by solving the simple optimization problem

$$\gamma_t = \max \arg_{\gamma \in \Gamma} \|R^\top \phi_\gamma\|_1, \tag{6}$$

and includes the selected ϕ_{γ_t} in Ψ. The residual matrix is updated by $R = (I - P)S$, where P is the orthogonal projector on the $\mathrm{span}\{\Psi\}$. The main steps of the SOMP algorithm are summarized in Table I.

Note that the Orthogonal Matching Pursuit (OMP) converges in a finite number of iterations [23, Sec.9.5.3] since the norm of the residual is decreasing strictly monotonically in each step. This can be generalized to the case of SOMP, as stated in the following proposition.

***Proposition** 1:* [21] In each step of SOMP, the norm of the residual decreases strictly monotonically and SOMP converges in a finite number of steps.

Therefore, a greedy solution based on SOMP is not prone to be trapped in local minima, and not sensitive to initializations, contrarily to the NMF algorithms.

IV. Supervised dimensionality reduction

A. Supervised SOMP

We now propose to extend the previous algorithm to classification, and we propose a supervised learning solution when class labels are available. In order to develop a supervised dimensionality reduction method, we modify the objective function in OPT1 by including an additional term that encourages the separability between different classes. First, let us denote the number of classes by c and assume without loss of generality that

$$S = [S^{(1)}, \ldots, S^{(c)}] \in \mathbb{R}^{m \times n}, \tag{7}$$

where $S^{(i)} \in R^{m \times n_i}$ denotes the data samples that belong to the i-th class of cardinality n_i. Then we formulate a supervised dimensionality reduction problem by modifying the optimization problem OPT1 as follows.

Optimization problem: **OPT2**
$\min_{\Psi,C} \|S - \Psi C\|_F^2 + \lambda J(\Psi)$
subject to
$\Psi \subseteq \mathcal{D}$.

In the above optimization problem $J(\Psi)$ denotes the cost function that captures the separability of different classes. The scalar λ drives the trade-off between the approximation error and the class separability. In order to

Algorithm: **S-SOMP**
Input: Signal matrix $S \in R^{m \times n}$ and `tol`: approximation error tolerance.
Output: Set of selected atoms Ψ, approximation A and residual matrix R.
1. Initialize the residual $R_0 = S$, $\Psi = []$, $t = 1$.
2. Find index γ_t which solves the optimization problem $\gamma_t = \max \arg_{\gamma \in \Gamma} \|R_t^\top \phi_\gamma\|_1 + \lambda(\|G_b^\top \phi_\gamma\|_1 - \|G_w^\top \phi_\gamma\|_1)$
3. Augment $\Psi = [\Psi,\ \phi_{\gamma_t}]$
4. Compute an orthonormal basis $V = [v_1, \ldots, v_t]$ of the span$\{\Psi\}$.
5. Compute the orthogonal projector $P_t = V_t V_t^\top$ on the span$\{\Psi\}$.
6. Compute the new approximation and residual
$A_t = P_t S$
$R_t = (I - P_t)S$
7. If $\|R\|_F \leq$ `tol`, then stop. Otherwise, increment iteration $t = t + 1$, and go to step (2).

TABLE II
THE SUPERVISED SOMP (S-SOMP) ALGORITHM.

solve OPT2, we propose to modify the atom selection step of SOMP by including the class separability term. The intuition is that in each step, the algorithm selects the atom, which best explains all signals and at the same time it discriminates between signals of different classes. We call the modified supervised algorithm S-SOMP.

The separability cost function is chosen to capture the difference between the projected within-class variance and the projected between-class variance. By the projected class variance, we mean the restrictions of the within-class scatter matrix S_w and the between-class scatter matrix S_b, on the candidate atom ϕ. They are respectively given as $\phi^\top S_w \phi$ and $\phi^\top S_b \phi$. In order to reflect the differences between the projected variances, we define the cost function as

$$J(\phi) = \|G_w^\top \phi\|_1 - \|G_b^\top \phi\|_1, \tag{8}$$

where G_w and G_b respectively represents the transposes of the square roots of the scatter matrices S_w and S_b. The scatter matrices are defined as

$$S_w = \frac{1}{n} \sum_{i=1}^{c} \sum_{s \in S^{(i)}} (s - \mu^{(i)})(s - \mu^{(i)})^\top \tag{9}$$

$$S_b = \frac{1}{n} \sum_{i=1}^{c} n_i (\mu^{(i)} - \mu)(\mu^{(i)} - \mu)^\top, \tag{10}$$

where

$$\mu^{(i)} = \frac{1}{n_i} \sum_{j=1}^{n_i} s_j^{(i)} \tag{11}$$

denotes the centroid of the i-th class (the notation $s_j^{(i)}$ denotes the j-th sample of the i-th class), and

$$\mu = \frac{1}{n} \sum_{j=1}^{n} s_j \tag{12}$$

represents the global centroid. Then, if we define $e^{(i)} = [1, \ldots, 1]^\top \in R^{n_i \times 1}$ and $e = [1, \ldots, 1]^\top \in R^{n \times 1}$, we can write the matrices $G_w \in \mathbb{R}^{m \times n}$ and $G_b \in \mathbb{R}^{m \times c}$ as

$$G_w = \frac{1}{\sqrt{n}} [S^{(1)} - \mu^{(1)}(e^{(1)})^\top, \ldots, S^{(c)} - \mu^{(c)}(e^{(c)})^\top]$$

$$G_b = \frac{1}{\sqrt{n}} [\sqrt{n_1}(\mu^{(1)} - \mu), \ldots, \sqrt{n_c}(\mu^{(c)} - \mu)],$$

where we observe that $S_w = G_w G_w^\top$ and $S_b = G_b G_b^\top$. It implies that both scatter matrices are symmetric and positive semi-definite. Note that the L1 norm has been chosen in the cost function in order to be in accordance with the non-supervised algorithm as given in Eq. (6). Finally, we can rewrite the optimization problem that we solve in each step of the supervised S-SOMP as,

$$\gamma_t = \max \arg_{\gamma \in \Gamma} \|R^\top \phi_\gamma\|_1 + \lambda(\|G_b^\top \phi_\gamma\|_1 - \|G_w^\top \phi_\gamma\|_1). \tag{13}$$

Table II summarizes the main steps of the S-SOMP algorithm.

B. Analysis of S-SOMP

As shown before, the residual of SOMP converges to zero as the number of iteration increases. In S-SOMP, the class separability is strengthened, which results in an effective algorithm for classification tasks. However, the separability cost function introduces a penalty on the convergence rate of the S-SOMP algorithm. In the extreme case where the penalty term is very large, it can even cause temporary stagnation of the residual energy, as is explained in the following remark.

Remark 1: The residual of the S-SOMP algorithm decreases strictly monotonically in each step t, if the following condition is satisfied,

$$\|R_t^\top \phi_{\gamma_t}\|_2^2 > 0, \quad \forall t. \tag{14}$$

Proof: Assume that in iteration t, the condition (14) is violated and the selected atom ϕ_{γ_t} is orthogonal to all columns of the residual matrix. In other words,

$$\|R_t^\top \phi_{\gamma_t}\|_2^2 = 0. \tag{15}$$

First, note that condition (15) implies that

$$\|R_t^\top v_{t+1}\|_2^2 = 0. \tag{16}$$

Indeed, it holds that

$$v_{t+1} = \phi_{\gamma_t} - \sum_{i=1}^{t} \zeta_i v_i, \tag{17}$$

where $\zeta_i = v_i^\top \phi_{\gamma_t}$ are the weights of the linear combination that make v_{t+1} orthogonal to $v_1, \ldots, v_t$. Note that they can be also computed using the Gram Schmidt orthogonalization process [24]. In the same time $R_t \perp \text{span}\{v_1, \ldots, v_t\}$, due to the construction of the algorithm. Combined with relation (17), it leads to the condition given in Eq. (16).

Then, we call $V_{t+1} = [v_1, \ldots, v_{t+1}]$ an orthogonal basis for the span$\{\Psi \bigcup \phi_{\gamma_t}\}$ obtained in the first $t+1$ iterations. The orthogonal projector on the span$\{\Psi \bigcup \phi_{\gamma_t}\}$ is

$$P_{t+1} = V_{t+1} V_{t+1}^\top = \sum_{i=1}^{t+1} v_i v_i^\top = P_t + v_{t+1} v_{t+1}^\top. \tag{18}$$

Using the above formula, we observe that

$$\begin{aligned} R_{t+1} &= S - P_{t+1} S = (I - P_{t+1}) S \\ &= (I - P_t - v_{t+1} v_{t+1}^\top) S \\ &= R_t - v_{t+1} v_{t+1}^\top S. \end{aligned} \tag{19}$$

However it holds that $v_{t+1} v_{t+1}^\top S$ because

$$\begin{aligned} v_{t+1} v_{t+1}^\top S &= v_{t+1} v_{t+1}^\top (R_t + A_t) \\ &= v_{t+1} v_{t+1}^\top R_t + v_{t+1} v_{t+1}^\top A_t \\ &= 0, \end{aligned} \tag{20}$$

where the first term is zero because of (16) and the second term is zero due to the fact that A_t belongs in the span$\{V_t\}$ and $v_{t+1} \perp \text{span}\{V_t\}$ (see also equation (17)). In this case we therefore have $R_{t+1} = R_t$ due to Eq. (19) and (20). We conclude that if condition (14) is violated, the residual stays unchanged. ■

The above remark suggests that one should be mindful with the selection of the Lagrange multiplier λ which drives the trade-off between approximation and class separability. It should be chosen carefully in order to make sure that the condition (14) is never violated and to keep the right balance between the two objectives. Notice also that during the course of the algorithm, the approximation error decreases, and keeping a constant λ, will result in shifting the emphasis from the approximation error to the separability criterion. In order to avoid this phenomenon and to reduce the probability that condition (14) is violated, we use an adaptive rule for updating the parameter λ:

$$\lambda_{t+1} = \|R_t \phi_{\gamma_t}\|_1, \text{ with } \lambda_0 = 1.$$

In other words, we set λ to be equal to the L1 norm of the residual of the previous step. This rule takes into account the observation that from iteration to iteration the residual drops. The adaptive strategy decreases λ in the same rate as the residual and keeps the right balance between approximation and class separability.

Concerning the convergence properties of S-SOMP, we can build a proposition analogous to Proposition 1 of Section III. In other words, S-SOMP converges in a finite number of steps. However, the decay of the residual may not be monotonic, and the approximation rate is mostly driven by the weight of the separability constraints. As was shown in Remark 1, the approximation rate is driven by λ, and the signal representation becomes a judicious compromise between good approximation, and effective discrimination between classes in supervised dimensionality reduction problems.

V. Application to Image Classification

A. Dictionary design

We first discuss in detail how one may build structured dictionaries for dimensionality reduction in the context of digital images. A structured dictionary $\mathcal{D}$ is built by applying geometric transformations to a generating mother function ϕ. The parameters of the geometric transformations are carefully sampled such that the resulting dictionary forms an overcomplete basis of the image space. A geometric transformation $\gamma \in \Gamma$ is represented by a unitary operator $U(\gamma)$ and in the simplest case it may be one of the following three types.

- *Translation* by $\vec{b} = [b_1 \ b_2]^\top$. $U(\vec{b})$ moves the generating function across the image

$$U(\vec{b})\phi(x, y) = \phi(x - b_1, y - b_2).$$

- *Rotation* by θ. $U(\theta)$ rotates the generating function by angle θ i.e.,

$$\begin{aligned} U(\theta)\phi(x, y) &= \phi(x', y'). \\ x' &= \cos(\theta)x + \sin(\theta)y \\ y' &= \cos(\theta)y - \sin(\theta)x \end{aligned}$$

- *Anisotropic scaling* by $\vec{a} = [a_1 \ a_2]^\top$. $U(\vec{a})$ scales the generating function anisotropically in the two directions i.e.,

$$U(\vec{a})\phi(x, y) = \phi(\frac{x}{a_1}, \frac{y}{a_2}).$$

Composing all the above transformations yields a transformation $\gamma = \{\vec{b}, \vec{a}, \theta\} \in \Gamma$. Finally, an atom in the structured dictionary

$$D = \{U(\gamma)\phi, \ \gamma \in \Gamma\}$$

is built as

$$\begin{aligned} U(\gamma)\phi(x, y) &= \phi(x', y'), \\ x' &= \frac{\cos(\theta)(x - b_1) + \sin(\theta)(y - b_2)}{a_1} \\ y' &= \frac{\cos(\theta)(y - b2) - \sin(\theta)(x - b1)}{a_2}. \end{aligned}$$

In image classification applications, we consider three different structured dictionaries generated by ϕ, where ϕ is

- *Gaussian* function:
$$\phi(x,y) = \frac{1}{\sqrt{\pi}} \exp(-(x^2+y^2)) \tag{21}$$
- *Anisotropic refinement* (AR) function [25]. This generating function has an edge-like form and has been successfully used for image coding. It is Gaussian in one direction and the second derivative of Gaussian in the orthogonal direction. It can be mathematically expressed as,
$$\phi(x,y) = \frac{2}{\sqrt{3\pi}}(4x^2-2)\exp(-(x^2+y^2)). \tag{22}$$
- *Gabor* function. This generating function is very popular in face recognition. It consists of a Gaussian envelope modulated by a complex exponential. We have used the real part of a simplified version of the Gabor function,
$$\phi(x,y) = \cos(2\pi x)\exp(-(x^2+y^2)). \tag{23}$$

Note that one of advantages of structured dictionaries lies in the fact that they enable a fast FFT-based implementation of the SOMP algorithms. Recall that in each step of the SOMP algorithm, we need to compute the inner product of the candidate atom with the residual signals. In practice we construct the atoms only in their centered position. The inner product of a residual signal r with all translated versions of an atom g, is computed via 2D convolution which can be effectively computed using 2D FFT. Using this computational trick the algorithm becomes computationally attractive, even in the context of high dimensional signals, like digital images.

B. Experimental Setup

In both SOMP algorithms described earlier, the construction of the atoms in each dictionary proceeds by sampling uniformly 10 orientation angles in $[0,\pi]$ and 5 logarithmically equi-distributed scales in $[1, N/6]$ horizontally and $[1, N/4]$ vertically, where N is the image size. For our experimental comparisons, we use the implementations of NMF and LNMF provided in `nmfpack` [4] which is a MATLAB software package developed by P. Hoyer. Note that the codes do not come with a stopping criterion. Thus, we run the NMF methods up to maximum number of iterations, which was set to 1000 for both NMF algorithms.

In the learning stage of both SOMP algorithms, which produces the matrix Ψ of basis vectors, we use 4 samples per class. For classification, each training signal s_i is projected using the basis vectors Q, where Q denotes Ψ for the SOMP methods and W for the NMF methods. In particular, we project the samples in the reduced space using the transpose of Q i.e.,
$$y_i = Q^\top s_i, \quad i = 1,\ldots,n.$$
Note that we have chosen to use the transpose of Q instead of its pseudo-inverse in order to avoid numerical problems. Then classification is accomplished in the reduced space by simple nearest neighbor (NN) classification. In other words, the test signal s_t is also projected by $y_t = Q^\top s_t$ and then classified by assigning it the label of its nearest neighbor, among all the training signals. We measure performance in terms of classification error rate, which is the percentage of the test samples that have been misclassified.

In our experiments, we use the following data sets:

a) Handwritten digit image collection: We use the handwritten digit collection that is publicly available at S. Roweis web page[1]. This collection contains 20×16 bit binary images of "0" through "9", and each class contains 39 samples. Hence the signal matrix is of size 320×390. We form the training set by a random subset of 10 samples per class and the remaining 29 samples are assigned in the test set.

b) ORL face database: The ORL (formerly Olivetti) database [26] contains 40 individuals and 10 different images for each individual including variation in facial expression (smiling/non smiling) and pose. Figure 1 illustrates two sample subjects of the ORL database along with variations in facial expression and pose. The size of each facial image is 112×92. However, we downsampled each facial image to 28×23 for computational efficiency. Hence the signal matrix is of size 644×400. We form the training set by a random subset of 5 different facial expressions/poses per subject and use the remaining 5 as a test set.

[1]http://www.cs.toronto.edu/∼roweis/data/binaryalphadigs.mat

	Size of S	Samples/class
Handwritten digits	320×390	39
ORL face data set	644×400	10
CBCL face data set	361×2429	-

TABLE III

THE DATA SETS USED IN THE EXPERIMENTAL EVALUATION.

c) CBCL face database: The CBCL face database [27] consists of 2,429 facial images of size 19×19. Hence the signal matrix is of size 361×2429. Note that for this data set, there are no class labels available for the individuals. All data sets that are used in the experimental evaluation are summarized in Table III, along with their main properties.

Fig. 1. Sample face images from the ORL database. There are 10 available facial expressions and poses for each subject.

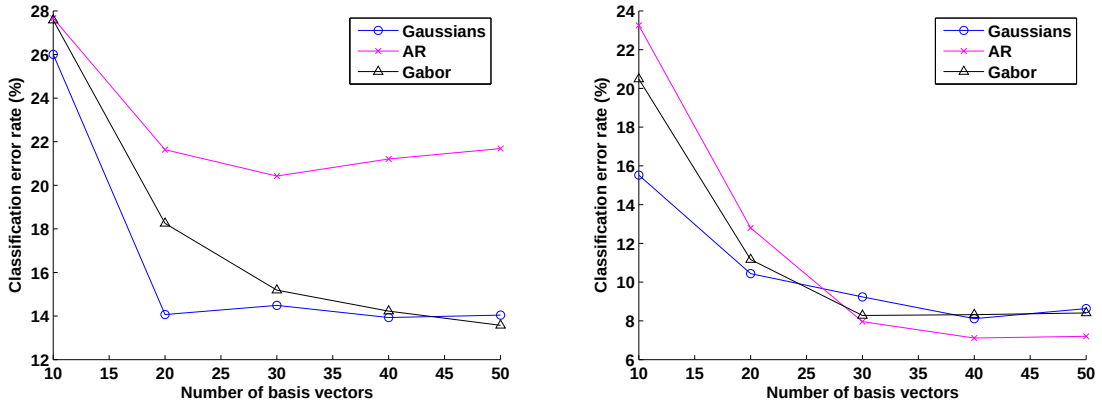
C. Classification performances

In the first experiment we investigate the impact of the dictionary on the classification performance by comparing the effectiveness of the three generating functions presented earlier. We run SOMP on both digit and ORL face data sets and compare the classification performance with respect to different dimensions $r = [10 : 10 : 50]$ (in MATLAB notation) of the reduced space. Sub-figures 2(a) and 2(b) depict the classification error rates obtained via the different dictionaries, for the digits and the face data set respectively. Note that for each value of r we report the average classification error rate across 100 random realizations of the training/test set. Observe that for the digits data set the dictionary built from Gaussian functions is the best performer. However, for the face data set the behavior is quite different and the AR dictionary seems to be competitive and even superior to the other dictionaries, especially for large dimensions. This is likely due to the fact that the AR atoms can represent the edge-like fine details of facial characteristics like the eyes, the mouth and so on. Therefore, in the following experiments and in both SOMP algorithms, we have chosen to use the Gaussian dictionary for the digit data set and the AR dictionary for the face data set.

Then we analyze and compare the classification performance obtained by the proposed algorithms, with several variants of NMF methods. The basis functions obtained from SOMP, NMF and LNMF algorithms are given in Figures 3 and 4, for the digits and faces data set respectively. The basis functions in sub-figures 3(a) and 4(a) are obtained from SOMP using the Gaussian dictionary. Similarly, the basis function in panels 3(b) and 4(b) are obtained from SOMP using the AR dictionary. The figures also depict the recovered basis functions from NMF and LNMF. Note that the features obtained from NMF are not localized and seem to be of global support. On the contrary, the features of LNMF are spatially localized and for the digits data set they seem quite similar to the Gaussian atoms.

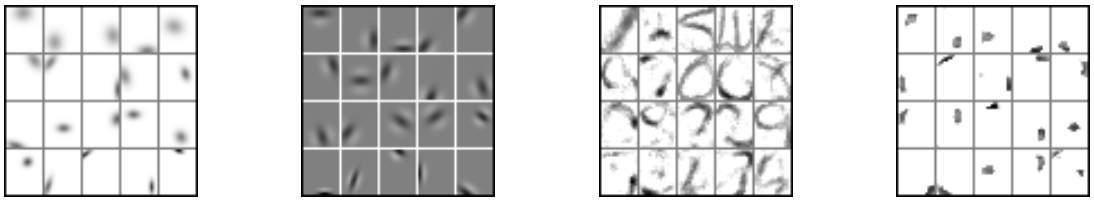
We now compare SOMP and S-SOMP with NMF and LNMF in terms of classification performance. In both data sets, we experiment with the dimension of the reduced space $r = [10 : 10 : 50]$ and in the classification experiments, for each value of r, we report the classification performance in terms of average error rate across 50 random realizations of the training/test set.

Figure 5(a) first depicts the average classification error rate for various values of the dimension r of the reduced space, for the handwritten digit image recognition task. The average is computed over 50 random realizations of the training/test set. Recall that in both SOMP algorithms we use the Gaussian dictionary. We observe that the SOMP algorithms are superior to the NMF algorithms. Furthermore, the supervised SOMP seems to be slightly better than its unsupervised counterpart.



(a) Digits data set (b) ORL face data set

Fig. 2. Impact of different dictionaries on the classification performance.



(a) Gaussian atoms (b) AR atoms (c) NMF (d) LNMF

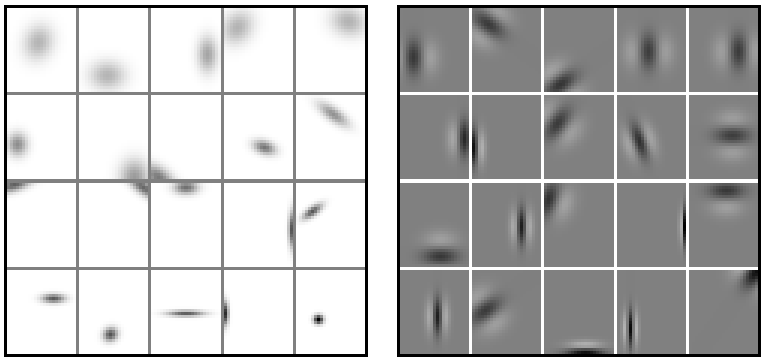
Fig. 3. Recovered basis vectors from the handwritten digit collection.

Then, Figure 5(b) depicts the average classification error rate across 50 random realizations of the training/test set for the face recognition task. Recall that for this data set we use the AR dictionary, in both SOMP algorithms. Observe that for small dimensions r of the reduced space, the SOMP algorithms are slightly inferior to the NMF algorithms. However, as r increases the SOMP methods become superior to the NMF methods. This can be explained by the greedy nature of the SOMP methods. In the first steps, the SOMP algorithms usually select atoms of large scale in order to reduce quickly the approximation error, but that do not consist in highly discriminating functions. The large scale atoms typically correspond to low frequency information which may not contribute a lot to the classification task.

It has to be noted that NMF may be combined with subsequent supervised methods such as Linear Discriminant Analysis (LDA)[28, ch.4] and yield an effective hybrid method which is among the state-of-the-art in face recognition and/or verification. In [17] the authors show competitive results of hybrid NMF methods in the context of face verification. Therefore we expect an analogous hybrid method of the proposed dimensionality reduction scheme to be competitive as well, with the state-of-the-art in these applications.

D. Discussion

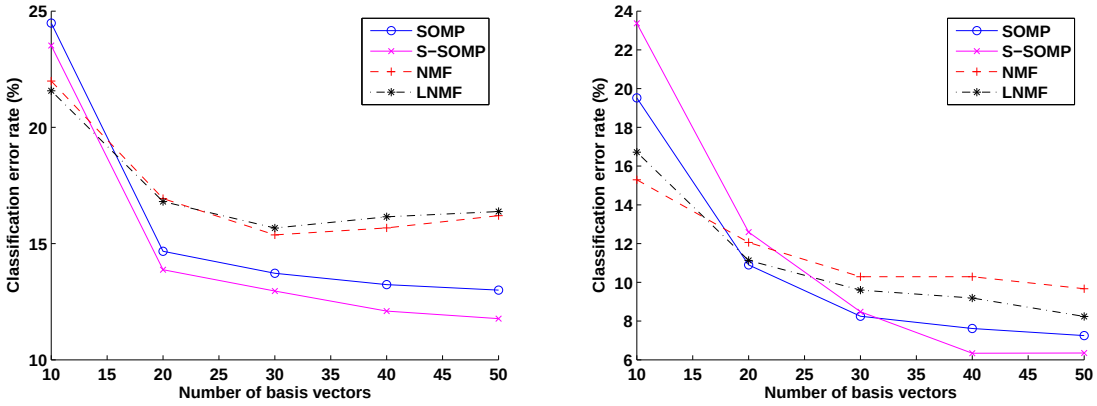
In this section, we finally discuss in more details the properties of SOMP algorithms in terms of convergence. Figure 6 first illustrates the behavior of both SOMP algorithms in terms of approximation rate and class separability, captured by $\hat{J}(\Psi) = \text{tr}(\Psi^\top S_b \Psi) - \text{tr}(\Psi^\top S_w \Psi)$. We run both SOMP algorithms on the digits data set and in S-SOMP, we use $\lambda_{t+1} = 10\|R_t \phi_{\gamma_t}\|_1$, with $\lambda_0 = 10$, for updating the parameter λ. Note that, as expected, the approximation rate is smaller for the supervised version of the SOMP algorithm, as discussed previously. Indeed,



(a) Gaussian atoms (b) AR atoms

(c) NMF (d) LNMF

Fig. 4. Recovered basis vectors from the ORL face data set.

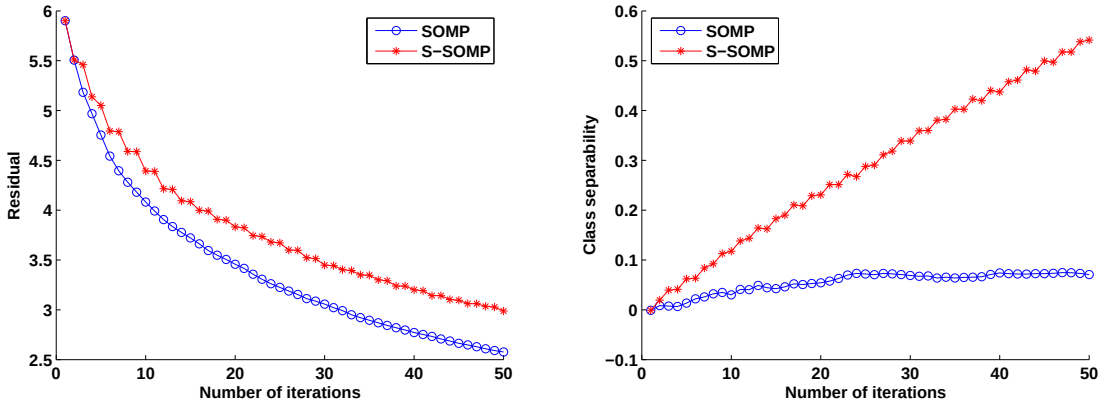


(a) Digits data set (b) ORL face data set

Fig. 5. Image recognition experiments.

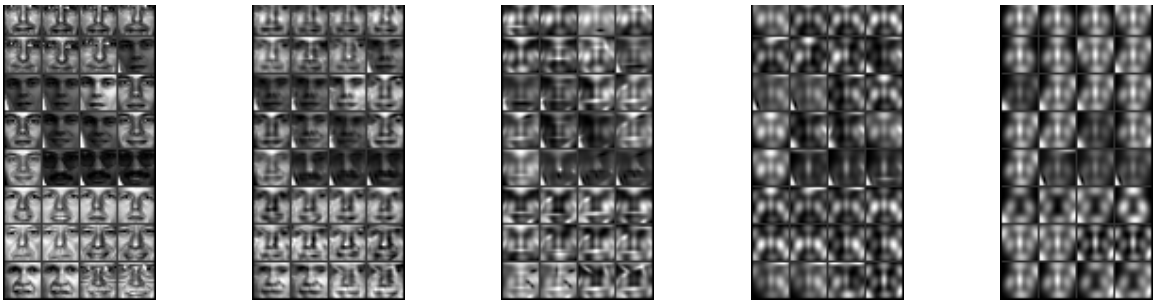
the algorithm does not select any more the “best” atom with respect to the approximation error, and is penalized by the separability cost function. However, the S-SOMP algorithm achieves higher discrimination among the different classes and hence, it offers better classification performances, as illustrated in Figure 5.

We finally illustrate the robustness of redundant expansions against errors in the signal representation. We compare



(a) Residual $\|R\|_F$ (b) Class separability $\hat{J}(\Psi)$

Fig. 6. Residual error and class separability versus number of iterations in S-SOMP.



(a) Original (b) No quantization (c) $q = 0.1$ (d) $q = 0.2$ (e) $q = 0.3$

Fig. 7. Quantization effects on facial images from the CBCL face data set, using SOMP.

(a) Original (b) No quantization (c) $q = 0.1$ (d) $q = 0.2$ (e) $q = 0.3$

Fig. 8. Quantization effects on facial images from the CBCL face data set, using NMF.

the behavior of facial representations of both SOMP and NMF under quantization noise. In particular, we test the robustness of the representation quality of human faces with respect to uniform quantization of the coefficient vectors C. Denote by q the quantization step of the uniform quantizer denoted in what follows as $Q(\cdot)$. We select $n = 32$ facial images $[s_1, \ldots, s_n]$ from the CBCL face database and run 50 steps of SOMP using a redundant dictionary of Gaussian atoms, yielding a common basis $\Psi \in R^{361\times 50}$. Then, for each facial image s_i, we perform the following steps,

1) Compute the coefficient vectors: $c_i = \Psi^{\dagger} s_i, \ \forall i$,
2) Quantize the coefficients: $\hat{c}_i = Q(c_i), \ \forall i$,
3) Reconstruct the facial images: $\hat{s}_i = \Psi \hat{c}_i, \ \forall i$.

We repeat the above process for different steps q of the quantizer. Figure 7(a) shows the facial images that were used in the approximation and 7(b) represents the obtained approximation with 50 Gaussian atoms, without quantization. Next, sub-figures 7(c)-7(e) illustrate the quantization effects for step sizes $q = 0.1, q = 0.2$ and $q = 0.3$. Figure 8 shows the analogous facial representations obtained with NMF, using 50 basis vectors. Notice that as the quantization becomes cruder (i.e, q increases) the reconstruction quality of SOMP facial representations degrades gracefully (in terms of perceptual quality). Thus, SOMP representations seem to be quite robust against quantization noise and the faces look remarkably natural. Even for $q = 0.3$ they still look like faces.

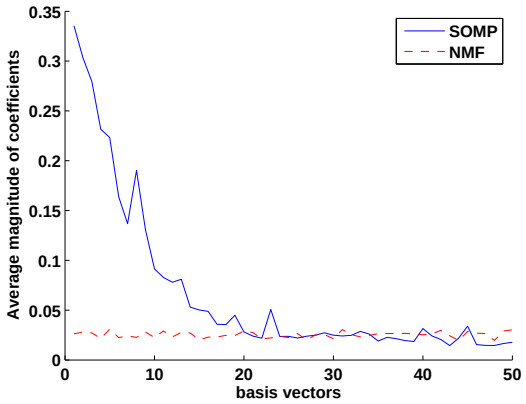


Fig. 9. Average magnitude of coefficients for both SOMP and NMF algorithms.

On the other hand, the facial modelling from NMF seems to be quite sensitive to quantization noise. Notice that in the majority of cases for $q = 0.3$, the NMF facial representations become identically zero (depicted as black blocks in the figure). This occurs due to the fact that the basis vectors in NMF representations are equally important, which implies that the corresponding coefficients are of similar order of magnitude. In SOMP representations however, due to the greedy nature of the algorithm, the first atoms are the most important ones. Usually they are of large scale and they capture the main geometric characteristics of the facial shape. The magnitude of their corresponding coefficients is quite higher than the magnitude of the coefficients of the remaining atoms. This is illustrated also in Figure 9, where we plot the magnitude of the coefficients of the basis vectors for both algorithms, averaged over the 32 facial images.

VI. Conclusions

We have proposed a method for dimensionality reduction using redundant dictionaries. We use greedy algorithms from simultaneous sparse signal approximation to extract meaningful features from overcomplete dictionaries. We have extended the algorithm to classification problems, where we proposed a supervised dimensionality reduction strategy. It includes a class separability penalty term in the objective function of the dimensionality reduction problem, which improves on the classification performance. The experimental results presented in the context of classification of handwritten digits, and face images, demonstrate the effectiveness of the proposed scheme. They suggest that the extracted features are meaningful and provide high discriminating value. This allows for a dimensionality reduction solution that offers jointly a good signal approximation, and interesting classification performances.

References

[1] D.D. Lee and H.S. Seung, "Algorithms for Non-negative Matrix Factorization", *Advances in Neural Information Processing Systems*, vol. 13, pp. 556-562, 2001.

[2] I.T. Jolliffe, "Principal Component Analysis", *Springer Verlag*, New York, 1986.
[3] S.Z. Li, X. Hou, H. Zhang and Q. Cheng, "Learning Spatially Localized, Parts-Based Representation", *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1-6, 2001.
[4] Patrik O Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints", *Journal of Machine Learning Research*, vol. 5, pp. 1457-1469, 2004.
[5] J. Tropp, A. Gilbert and M. Strauss, "Algorithms for Simultaneous Sparse Approximation. Part I: Greedy pursuit", *Signal Processing*, special issue "Sparse approximations in signal and image processing" , vol. 86, April 2006, pp. 572-588.
[6] X. He and P. Niyogi, "Locality Preserving Projections", *Advances in Neural Information Processing Systems 16 (NIPS 2003)*, Vancouver, Canada, 2003.
[7] E. Kokiopoulou and Y. Saad, "Orthogonal Neighborhood Preserving Projections", *IEEE Int. Conf. on Data Mining*, New Orleans, Louisiana, USA, Nov. 26-30, 2005.
[8] S. Roweis and L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding", *Science*, vol. 290, pp. 2323-2326, 2000.
[9] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation", *Neural Comput.*, vol. 15(6), pp. 1373-1396, 2003.
[10] J.B. Tenenbaum, V. de Silva and J.C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction ", *Science*, vol. 290(5500), pp. 2319-2323, 2000.
[11] 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.
[12] M. Heiler and C. Schnörr, "Learning Non-Negative Sparse Image Codes by Convex Programming", *IEEE Intl. Conf. on Comp. Vision (ICCV)*, Beijing, China, 2005.
[13] M. Heiler and C. Schnörr, "Learning Sparse Representations by Non-Negative Matrix Factorization and Sequential Cone Programming", *Journal of Machine Learning Research*, vol. 7, pp. 1385–1407, July 2006.
[14] V.P. Pauca , J. Piper and R.J. Plemmons, "Non-Negative Matrix Factorization for Spectral Data Analysis", *Linear Algebra and its Applications*, vol. 416, pp. 29-47, 2006.
[15] A. Pascual-Montano, J. Carazo, K. Kochi, D. Lehmann and R. Pascual-Marqui, "Nonsmooth Nonnegative Matrix Factorization (nsNMF)", *IEEE Transactions on Pattern Analysis and Machine Inteligence*, vol. 28(3), pp. 403-415, March 2006.
[16] Y. Wang, Y. Jia, C. Hu and M. Turk, "Fisher Non-negative Matrix Factorization for Learning Local Features", *Asian Conference on Computer Vision*, Jan. 27-30th, Jeju Island, Korea, pp. 806-811.
[17] S. Zafeiriou, A. Tefas, I. Buciu and I. Pitas, "Exploiting Discriminant Information in Non-negative Matrix Factorization with Application to Frontal Face Verification", *IEEE Transactions on Neural Networks*, vol. 17(3), pp. 683-695, May 2006.
[18] P. Jonathon Phillips, "Matching Pursuit Filters Applied to Face Identification", *2nd IEEE Transactions on Image Processing*, vol. 7(8), pp. 1150-1164, August, 1998.
[19] J. Tropp, "Algorithms for Simultaneous Sparse Approximation. Part II: Convex relaxation", *Signal Processing*, special issue "Sparse approximations in signal and image processing" , vol. 86, April 2006, pp. 589-602.
[20] D. Leviatan and V.N. Temlyakov., "Simultaneous Approximation by Greedy Algorithms", *Advances in Computational Mathematics, Springer*, vol. 25(1), pp. 73-90, June 2006.
[21] S.F. Cotter, B.D. Rao, K. Engan and K. Kreutz-Delgado, "Sparse Solutions to Linear Inverse Problems with Multiple Measurement Vectors", *IEEE Transactions on Signal Processing*, vol. 53(7), pp. 2477-2488, July 2005.
[22] S. Mallat and Z. Zhang., "Matching Pursuit with Time-Frequency Dictionaries", *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397415, Dec. 1993.
[23] S. Mallat, *A Wavelet Toor of Signal Processing, 2nd edn*, Academic Press, 1998.
[24] G. H. Golub and C. Van Loan, *Matrix Computations, 3rd edn*, The John Hopkins University Press, Baltimore, 1996.
[25] R. Figueras i Ventura, P. Vandergheynst and P. Frossard, "Low Rate and Flexible Image Coding with Redundant Representations", *IEEE Transactions on Image Processing*, vol. 15(3), pp. 726-739, March 2006.
[26] F. Samaria and A. Harter, "Parameterisation of a Stochastic Model for Human Face Identification", *2nd IEEE Workshop on Applications of Computer Vision*, Sarasota, FL, Dec. 1994.
[27] CBCL Face Database #1, MIT Center for Biological and Computation Learning, `http://www.ai.mit.edu/projects/cbcl`.
[28] A. Webb, *Statistical Pattern Recognition, 2nd edn*, Wiley, 2002.