On the Invariance of Dictionary Learning and Sparse Representation to Projecting Data to a Discriminative Space

Mehrdad J. Gangeh*, Senior Member, IEEE, Ali Ghodsi,

Abstract—In this paper, it is proved that dictionary learning and sparse representation is invariant to a linear transformation. It subsumes the special case of transforming/projecting the data into a discriminative space. This is important because recently, supervised dictionary learning algorithms have been proposed, which suggest to include the category information into the learning of dictionary to improve its discriminative power. Among them, there are some approaches that propose to learn the dictionary in a discriminative projected space. To this end, two approaches have been proposed: first, assigning the discriminative basis as the dictionary and second, perform dictionary learning in the projected space. Based on the invariance of dictionary learning to any transformation in general, and to a discriminative space in particular, we advocate the first approach.

Index Terms—Dictionary learning, discriminative projected space, linear transformation, sparse representation, supervised learning.

I. INTRODUCTION

There are many mathematical models with varying degrees of success to describe data, among which dictionary learning and sparse representation (DLSR) have attracted the interest of many researchers in various fields. Dictionary learning and sparse representation are two closely-related topics that have roots in the decomposition of signals to some predefined basis, such as the Fourier transform. Representation of signals using predefined basis is based on the assumption that these bases are sufficiently general to represent any kind of signal. However, recent research shows that learning the basis from data, instead of using predefined ones, leads to state-of-the-art results in many applications such as texture classification [11–13], face recognition [4–6], image denoising [7, 8], biomedical tissue characterization [9–11], motion and data segmentation [12, 13], data representation and column selection [14], and image super-resolution [15].

In classical dictionary learning and sparse representation (DLSR), the main goal is to represent a signal/data sample using a few dictionary atoms, i.e., sparsely, by minimizing a loss function. The most common loss function is the mean-squared error between the original signal and the reconstructed one. To induce the sparsity, a sparsity inducing penalty is introduced into the formulation. Although the most natural sparsity inducing function is $\ell_0$ norm, since it causes the resulting optimization problem to be non-convex, $\ell_1$ norm, as the convex envelop of $\ell_0$ norm, is usually used instead [16].

To provide a mathematical formulation, let $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{p \times n}$ be a set of $n$ finite training data samples with the dimensionality of $p$. The classical DLSR optimization problem using mean-squared error as the loss function and $\ell_1$ norm as the sparsity regularization term is as follows:

$$L(X, D, \alpha) = \min_{D, \alpha} \sum_{i=1}^{n} (\frac{1}{2} \|x_i - D\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1),$$

where $D \in \mathbb{R}^{p \times k}$ is the dictionary with $k$ atoms, $\alpha \in \mathbb{R}^{k \times n}$ is the sparse coefficients, and $\alpha_i$ is the $i^{th}$ column of $\alpha$.

In [1], the main goal is to minimize the reconstruction error between the original signal and its sparse representation. While this is important in applications such as denoising, inpainting, and coding, it does not lead to an optimal dictionary when classification is the primary task [17].

To compute a discriminative dictionary, which is more desired when the ultimate goal is to classify the signals in hand, a branch of DLSR has been recently introduced, which is appropriately called supervised dictionary learning (SDL) [11, 5, 6, 17–30]. Most SDL methods try to include the category information into the learning of the dictionary to enhance the discrimination power of the learned dictionary. For a recent review on the SDL methods, the interested reader is urged to refer to [27].

Among the proposed SDL approaches in the literature, there have recently been some approaches that suggest learning the dictionary and corresponding sparse coefficients in a more discriminative projected space to fulfil the needs for a supervised dictionary learning. To this end, while both Gangeh et al. [27] and Zhang et al. [28] propose transforming/projecting the data to a discriminative space, the former uses the projection basis as the dictionary elements, whereas the latter suggests performing dictionary learning and sparse representation in the

\[ \|x\|_0 = \#\{i : x_i \neq 0\}. \]
projected space. In other words, in [27], the transformation itself is considered as the dictionary whereas in [28], the dictionary is learned in the projected space after the transformation is applied. Although the criteria used for making the projected space discriminative are different in the two approaches, our focus is not on the discrimination criteria in this paper, but on the way to learn the dictionary.

In this paper, we prove that dictionary learning is invariant to a linear transformation. This means that performing DLSR in a linearly transformed space does not lead to any different solution than the original space. In the proof, no assumption (such as orthonormality, etc.) has been made on the transformation and hence, it is general and includes any linear transformation. This subsumes the special case of projecting the data to a discriminative (in any sense) space. Therefore, projecting the data to a discriminative space first and then learning the dictionary in the projected space can be, for example, used as the dictionary method in [28] does not make any difference with learning the dictionary in the original space. Instead, as proposed in [27], the basis in the projected space can be, for example, used as the dictionary elements.

The invariance of learning the dictionary and subsequently the sparse coefficients to a linear (possibly more discriminative) transformation is an important property of DLSR. The main contribution of this paper is providing the mathematical proof for this invariance.

The organization of the rest of the paper is as follows: in Section II we provide the mathematical proof that dictionary learning and sparse representation is invariant to any linear transformation. In Section III we use the invariance property of DLSR to a linear transformation and discuss it in the context of supervised dictionary learning by presenting a case study for two different suggested approaches in the literature for learning the dictionary in a (discriminative) projected space. Finally in Section IV we provide some suggestions on how to compute dictionary in a projected space and conclude the paper.

II. INVARIANCE OF DLSR TO A LINEAR TRANSFORMATION

In this section, we show that learning the dictionary and sparse coefficients is invariant to a linear transformation of the data to another space. To this end, we compute the dictionary and coefficients in the original space as well as the transformed space and show that these computations lead to same solutions.

To learn the dictionary in the original space, we suppose that \( X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{p \times n} \) are \( n \) finite \( p \)-dimensional training samples, from which the dictionary \( D_o \in \mathbb{R}^{p \times k} \) (\( D_o \) is the dictionary in the original space) and coefficients \( \alpha \in \mathbb{R}^{k \times n} \) are to be learned. This can be done using the formulation provided in (1). However, since (1) is non-convex when both the dictionary and coefficients are unknown, as is common in DLSR literature [20], [31], the solution for the optimization problem given in (1) alternates between sparse coding and dictionary updating. In the step of dictionary updating, using the given set of training samples \( X \) in the original space and a fixed set of sparse coefficients \( \alpha \), we need to solve the least square problem from the reconstruction term in (1) as follows

\[
\min_{D_o} \sum_{i=1}^{n} \| x_i - D_o \alpha_i \|_2^2, \tag{2}
\]

which can be rewritten in matrix form as

\[
\min_{D_o} \| X - D_o \alpha \|_F^2, \tag{3}
\]

where \( \| . \|_F \) is the Frobenius norm.\(^3\)

To solve the optimization problem given in (3), we take its derivative in respect to the dictionary \( D_o \), and perform a few manipulations as follows

\[
\frac{\partial}{\partial D_o} \| X - D_o \alpha \|_F^2 = \frac{\partial}{\partial D_o} \text{tr}( (X - D_o \alpha)^T (X - D_o \alpha) ) \\
= \frac{\partial}{\partial D_o} [ \text{tr}(X^T X) - 2\text{tr}(X^T D_o \alpha) + \text{tr}(\alpha^T D_o^T D_o \alpha) ] \\
= -2X \alpha^T + 2D_o \alpha \alpha^T,
\]

where \( \text{tr} \) is the trace operator.

By equating the above derivative to zero and solving for \( D_o \), we obtain

\[
D_o = X \alpha^T (\alpha \alpha^T)^{-1}, \tag{4}
\]

which is the dictionary in the original space.

Now, we address the problem of learning the dictionary \( D_i \in \mathbb{R}^{p \times k} \) in the transformed space. Suppose that there exists a transformation \( U \in \mathbb{R}^{p \times m} \). To perform dictionary learning and sparse representation in the transformed space \( U^T X \), we need to solve the following optimization problem

\[
\min_{D_o, \alpha} \sum_{i=1}^{n} \left( \| U^T x_i - U^T D_o \alpha_i \|_2^2 + \lambda \| \alpha_i \|_1 \right), \tag{5}
\]

which has to be again solved iteratively and alternately between the sparse coding and dictionary updating steps. At the dictionary update step, with a given set of sparse coefficients \( \alpha \), in the transformed space of \( U^T X \), we need to minimize the reconstruction error in the transformed space, i.e.,

\[
\min_{D_i} \| U^T X - U^T D_i \alpha \|_F^2. \tag{6}
\]

To solve the optimization problem given in (6), similar to the steps followed for the computation of the dictionary in the original space, we take the derivative in respect to the dictionary \( D_i \), and perform a few manipulations as follows

\[
\frac{\partial}{\partial D_i} \| U^T X - U^T D_i \alpha \|_F^2 = \frac{\partial}{\partial D_i} \text{tr}( (U^T X - U^T D_i \alpha)^T (U^T X - U^T D_i \alpha) ) \\
= \frac{\partial}{\partial D_i} [ \text{tr}(X^T U U^T X) - 2\text{tr}(X^T U U^T D_i \alpha) + \text{tr}(\alpha^T D_i^T U U^T D_i \alpha) ] \\
= -2UU^T X \alpha^T + 2UU^T D_i \alpha \alpha^T.
\]

\(^3\)The Frobenius norm of a matrix \( X \) is defined as \( \| X \|_F = \sqrt{\sum_{i,j} (x_{i,j}^2)} \).
By equating the above derivative to zero we obtain
\[
UU^T D_t \alpha \alpha^T = UU^T X \alpha^T. \tag{7}
\]

Now consider the system of equations \( AD_t C = B \), where the matrices \( A, B, \) and \( C \) are known and \( D_t \) is unknown. It is known that if there exist Moore-Penrose pseudo-inverses \( [32] \) (or in short pseudo-inverses) \( A^\dagger \) and \( C^\dagger \) of \( A \) and \( C \), respectively, such that the following condition is satisfied
\[
AA^\dagger BC^\dagger C = B, \tag{8}
\]
then the system of equations \( AD_t C = B \) can be solved for \( D_t \), and
\[
D_t = AA^\dagger B + Y - A^\dagger AYCC^\dagger \tag{9}
\]
is a solution for any matrix \( Y \in \mathbb{R}^{p \times k} \). [33], [34].

Now, we show that the condition provided in (8) is held. As shown in (7), suppose \( A = UU^T, B = UU^T X \alpha^T \), and \( C = \alpha \alpha^T \), where \( A \in \mathbb{R}^{p \times p}, B \in \mathbb{R}^{p \times k} \), and \( C \in \mathbb{R}^{k \times k} \). Replacing these values of \( A, B, \) and \( C \) into the left side of condition given in (8), it converts to
\[
UU^T(UU^T)^\dagger UU^T X \alpha^T (\alpha \alpha^T)^\dagger (\alpha \alpha^T) = UU^T X \alpha^T. \tag{10}
\]
According to the definition of pseudo-inverse of a matrix \( [32] \), \( AA^\dagger A = A \) and hence, \( UU^T(UU^T)^\dagger UU^T = UU^T \). On the other hand, \( C = \alpha \alpha^T \) is square and full rank as \( k < n \), and therefore, \( C^\dagger = C^{-1} \). Thus, \( UU^T(UU^T)^\dagger UU^T X \alpha^T (\alpha \alpha^T)^\dagger (\alpha \alpha^T) = UU^T X \alpha^T, \) i.e., the condition \( AA^\dagger BC^\dagger C = B \) given in (8) is held and the solution provided in (9) is indeed a valid solution for \( D_t \).

By replacing the values of \( A, B, \) and \( C \) in (9), we obtain the dictionary in the transformed space as follows
\[
D_t = (UU^T)^\dagger UU^T X \alpha^T (\alpha \alpha^T)^{-1} + Y - (UU^T)^\dagger UU^T Y. \tag{10}
\]

To derive (10), we have used the identity \( CC^\dagger = I \).

Two cases can be considered for the solution provided in (10) as described below:

1. If \( (UU^T)^\dagger UU^T = I \) \((m \geq p)\), then \( D_t = X \alpha^T (\alpha \alpha^T)^{-1} = D_o \). The solution is unique and the same as the dictionary computed in the original space.
2. However, in case of \( (UU^T)^\dagger UU^T \neq I \), i.e., \( m < p \), there are infinitely many solutions for \( D_t \) depending on the matrix \( Y \in \mathbb{R}^{p \times k} \), including \( Y = X \alpha^T (\alpha \alpha^T)^{-1} \). Replacing this particular \( Y \) into the solution provided for \( D_t \) in (10) leads to a solution as follows
\[
D_t = (UU^T)^\dagger UU^T X \alpha^T (\alpha \alpha^T)^{-1} + X \alpha^T (\alpha \alpha^T)^{-1} - (UU^T)^\dagger UU^T X \alpha^T (\alpha \alpha^T)^{-1},
\]
in which the first and last terms cancel each other out and hence, \( D_t = X \alpha^T (\alpha \alpha^T)^{-1} = D_o \). This means that \( D_t = D_o \) is still a solution, but there are infinitely many other solutions that are as good as \( D_o \).

In both cases, the dictionary learned in the transformed space is the same as the dictionary learned in the original space. In the latter case, however, the solution is not unique and there exist infinitely many other solutions, which are as good as the dictionary learned in the original space.

It is worthwhile to mention that in the proof, no assumption (such as orthonormality) has been made on the transformation and hence, it can be any linear transformation. This includes any transformation that projects the data to a possibly discriminative space and we discuss it in the context of supervised dictionary learning in next section. However, the conclusion made in this paper is general and includes any linear transformation whether it is discriminative or not.

### III. Dictionary in Projected Space-A Case Study

In this section, we provide an overview of two recently proposed SDL methods in the literature that suggest projection to a more discriminative space. The two methods introduce two different techniques for projecting the data to a more discriminative space, one by using Hilbert-Schmidt independence criterion (HSIC) to project the data to a space where the dependency between the data and class labels is maximized, the other by projecting the data to a space where the ratio of intra-class to inter-class reconstruction errors is minimized. Although understanding the methods to project to a more discriminative space may not be strictly essential here – as our emphasis is on how to learn the dictionary and corresponding sparse coefficients in the projected space – a brief description of projecting methods provide an insight on the two SDL approaches, as described in the following two subsections. It is important, however, to note that the two methods propose two different dictionary learning in the projected space: the former by using the basis of the discriminative space as the dictionary (Subsection III-A); and the latter learning the dictionary in a discriminative projected space (Subsection III-B).

#### A. Basis of Projected Space as Dictionary-HSIC-Based Dictionary Learning

In [27], Gangeh et al. have proposed to assign the (discriminative) transformation to the dictionary. To make the projected space discriminative, Hilbert-Schmidt independence criterion (HSIC) [35] has been used, using which the dependency between the data samples and the corresponding labels has been maximized. HSIC is a kernel-based independence measure between two random variables \( X \) and \( Y \) [36]. It computes the Hilbert-Schmidt norm of the cross-covariance operators in reproducing kernel Hilbert Spaces (RKHSs) [36], [37].

In practice, HSIC is estimated using a finite number of data samples. Let \( Z := \{(x_1, y_1), ..., (x_n, y_n)\} \subseteq X \times Y \) be \( n \) independent observations drawn from \( p := p_{X \times Y} \). The empirical estimate of HSIC can be computed using [33], [36]
\[
\text{HSIC}(Z) = \frac{1}{(n-1)^2} \text{tr}(KHLH), \tag{11}
\]
where \( \text{tr} \) is the trace operator, \( H, K, L \in \mathbb{R}^{n \times n}, K_{i,j} = k(x_i, x_j), L_{i,j} = I(y_i, y_j), \) and \( H = I - n^{-1}ee^T \) (\( I \) is the identity matrix, and \( e \) is a vector of \( n \) ones, and hence, \( H \) is the centering matrix). According to (11), maximizing the empirical estimate of HSIC, i.e., \( \text{tr}(KHLH) \), will lead to the maximization of the dependency between the two random variables \( X \) and \( Y \).
To compute the projected space of maximum dependency between the data and class labels, it has been proposed in [27] to solve the following optimization problem

$$\max_U \text{tr}(U^\top XHLHX^\top U),$$

subject to $U^\top U = I$ \hspace{1cm} (12)

where $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{p \times n}$ is $n$ data samples with the dimensionality of $p$; $H$ is the centering matrix, and its function is to center the data, i.e., to remove the mean from the features; $L$ is a kernel on the labels $y \in \mathbb{R}^n$; and $U$ is the transformation that maps the data to the space of maximum dependency with the labels. According to the Rayleigh-Ritz Theorem, the solution for the (12) is the top eigenvectors of $\Phi = XHLHX^\top$ corresponding to its largest eigenvalues.

To explain how the optimization problem provided in (12) leads to a dictionary in the space of maximum dependency with the labels, using a few manipulations, we note that the objective function given in (12) has the form of the empirical HSIC given in (11), i.e.,

$$\max_U \text{tr}(U^\top XHLHX^\top U) = \max_U \text{tr}(X^\top UU^\top XHLH) = \max_U \text{tr}
\left(\left(U^\top X \right)^\top U^\top X \right)HLH = \max_U \text{tr}(KHLH),$$

(13)

where $K = (U^\top X)^\top U^\top X$ is a linear kernel on the transformed data in the subspace $U^\top X$. To derive (13), it is noted that the trace operator is invariant under cyclic permutation.

Now, it is easy to observe that the form given in (13) is the same as the empirical HSIC in (11) up to a constant factor and therefore, it can be easily interpreted as transforming centered data $X$ using the transformation $U$ to a space where the dependency between the data and class labels is maximized.

As described in [27], the computed transformation $U$, which is the basis of the space of maximum dependency between the data and class labels, constructs the dictionary. Based on the proof provided in Section II, no further dictionary learning and sparse representation is needed in the transformed/projected space as learning the dictionary (and sparse coefficients) are invariant to this transformation. However, the transformation provides the directions of maximum discrimination in the sense described above and hence, by assigning them to the dictionary, the dictionary atoms are indeed the directions of maximum separation (in the same sense), which is the ultimate goal in the design of a discriminative dictionary for classification tasks.

After finding the dictionary $D = U$, the sparse coefficients can be computed using the formulation given in (1) [27].

B. Learning Dictionary in Projected Space-Simultaneous Discriminative Projection and Dictionary Learning

Similar to the HSIC-based SDL, Zhang et al. [28] have also proposed to project the data to a more discriminative (in some sense, which will be defined in next lines) space. However, instead of assigning the discriminative projected space to the dictionary as has been proposed in HSIC-based SDL, the dictionary is learned in the projected space. To this end, they propose to first project the data to an orthogonal space where the intra- and inter-class reconstruction errors are minimized and maximized, respectively, and subsequently learn the dictionary and the sparse representation of the data in this space. Intra-class reconstruction error for a data sample $x_i$ is defined as the reconstruction error using the dictionary atoms in the ground-truth class of $x_i$ under the metric $UU^\top$ ($U$ is the projection to be learned), whereas inter-class error is defined as the reconstruction error using the dictionary atoms other than ground-truth class of $x_i$ under the same metric.

To provide the mathematical formulation, given a set of training set $X \in \mathbb{R}^{p \times n}$, the task is to learn a discriminative transformation/projection $U \in \mathbb{R}^{p \times m}$, where $m \leq p$ is the number of basis, and dictionary $D \in \mathbb{R}^{p \times k}$, using the optimization problem given as

$$\min_{U, D} \frac{1}{n} \sum_{i=1}^{n} \left( S_\beta(R(x_i)) + \lambda \|\alpha_i\|_1 \right)$$

subject to $U^\top U = I$ \hspace{1cm} (14)

where $S_\beta(x) = \frac{1}{1+\beta^x}$ is a sigmoid function centered at 1 with the slope of $\beta$, and $R(x_i)$ is the ratio of intra- to inter-class reconstruction errors. $S_\beta(R(x_i))$ can be intuitively considered as the inverse classification confidence and by minimizing this term over the training samples in the objective function of (14), the discriminative projections $U$ and dictionary $D$ are empirically learned subject to a sparsity constraint imposed as the second term in (14).

In (14), $\alpha_i$ is the sparse representation of the projected data sample $U^\top x_i$ in the space of dictionary learned in the projected space $U^\top D$, i.e.,

$$\hat{\alpha}_i = \min_{\alpha} \left( \|U^\top x_i - U^\top D\alpha_i\|^2_2 + \lambda \|\alpha\|_1 \right).$$

(15)

The optimization problem given in (14) and (15) has to be solved alternately between sparse coding (using (15) with $U$ and $D$ fixed) and learning the dictionary and projected space (using (14) with fixed sparse coefficients $\alpha$).

As can be seen, it is proposed here to learn the dictionary and sparse coefficients in the projected space. However, since the transformation does not have a closed-form solution, its learning is also performed simultaneously using gradient descent, which may lead to a suboptimal solution as it may get stuck in a local minima. Based on the invariance of DLSR to projecting data to a discriminative space, this approach may not lead to any different solution other than the one learned in the original space, and hence is not recommended.

IV. DISCUSSION AND CONCLUSION

In this paper, we provided a mathematical proof for the invariance of the dictionary learning to a linear transformation. In the proof, no assumption, such as orthonormality, was made on the transformation/projection and therefore, the projection can be any linear transformation, which makes the conclusion sufficiently general to include a broad range of transformations including discriminative projections.
Since projection to a discriminative space has been recently introduced in the literature of supervised dictionary learning (SDL) as a means to enhance the discriminative power of the learned dictionary, which is highly desired when the ultimate goal is the classification of provided data samples, we investigated two different recently proposed SDL methods as special cases that suggest such a transformation. While the two methods propose two different criteria for making the projected space discriminative, i.e., one based on the maximizing the dependency between the data and the class labels using Hilbert-Schmidt independence criterion, and the other based on minimizing the ratio of intra-to inter-class reconstruction errors, the focus of the paper is not on the discrimination criterion but on how to learn the dictionary and subsequently the sparse coefficients in the projected space.

W.r.t the learning dictionary in the projected space, HSIC-based SDL [27] suggests assigning the basis of learned projected space to the dictionary, whereas simultaneous discriminative projection and dictionary learning [28] proposes to learn the dictionary in the projected space. In this paper, we showed that the latter leads to the same solution for the dictionary as in the original space and hence, the dictionary learned is not more discriminative than the original space. In other words, we showed that learning the dictionary (and subsequently the sparse coefficients) is invariant to projection to another (discriminative) space.

It is worthwhile to highlight here that in our proof in Section 11 it was assumed that the optimal projection space has already been computed and is available (like HSIC-based method). However, in Zhang et al. method [28], the projected space is learned along with the dictionary in an alternating minimization scheme using gradient descent. This leads to a suboptimal solution for the projected space as the optimization formulation when both the dictionary and projected space are unknown is non-convex and by finding the solution using gradient descent, a local minima may be found.

Based on the invariance of learning the dictionary and sparse representation to projecting data to a discriminative space, we recommend that instead of learning the dictionary in the projected space, the learned basis for the projected space is used as the dictionary directly. This has another advantage as the dictionary is the same as the basis of the transformation/projection. Therefore, the optimal projected space can be computed using the optimization problem defined based on the discrimination criterion for the projected space not in a suboptimal way using gradient descent by alternating between learning the dictionary and the projected space.

REFERENCES

[1] M. J. Gangeh, A. Ghodsii, and M. S. Kamel, “Dictionary learning in texture classification,” in Proceedings of the 8th international conference on Image analysis and recognition - Volume Part I. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 335–343.

[2] J. Xie, L. Zhang, J. You, and D. Zhang, “Texture classification via patch-based sparse texton learning,” in 17th IEEE International Conference on Image Processing (ICIP), 2010, pp. 2737–2740.

[3] M. Varma and A. Zisserman, “A statistical approach to material classification using image patch exemplars,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 11, pp. 2032–2047, Nov. 2009.

[4] C. Zhong, Z. Sun, and T. Tan, “Robust 3D face recognition using learned visual codebook,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–6.

[5] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[6] M. Yang, L. Zhang, X. Feng, and D. Zhang, “Fisher discriminative dictionary learning for sparse representation,” in 13th IEEE International Conference on Computer Vision (ICCV), 2011, pp. 543–550.

[7] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” IEEE Transactions on Image Processing, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[8] J. Mairal, M. Elad, and G. Sapiro, “Sparse representation for color image restoration,” IEEE Transactions on Image Processing, vol. 17, no. 1, pp. 53–69, Jan. 2008.

[9] M. J. Gangeh, L. Sørensen, S. B. Shaker, M. S. Kamel, M. de Bruijne, and M. Loog, “A texture-based approach for the classification of lung parenchyma in CT images,” in Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). Berlin, Heidelberg: Springer-Verlag, 2010, pp. 595–602.

[10] L. Sørensen, M. J. Gangeh, S. B. Shaker, and M. de Bruijne, “Texture classification in pulmonary CT,” in Lung Imaging and Computer Aided Diagnosis, A. El-Baz and J. S. Sure, Eds. CRC Press, 2007, pp. 343–367.

[11] M. J. Gangeh, A. Sadeghi-Naini, M. S. Kamel, and G. Czarnota, “Assessment of cancer therapy effects using texton-based characterization of quantitative ultrasound parametric images,” in Proceedings of the International Symposium on Biomedical Imaging: From Nano to Macro (ISBI), 2013, pp. 1360–1363.

[12] S. Rao, R. Torn, R. Vidal, and Y. Ma, “Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[13] E. Elhamifar and R. Vidal, “Sparse subspace clustering,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 2790–2797.

[14] E. Elhamifar, G. Sapiro, and R. Vidal, “See all by looking at a few: Sparse modeling for finding representative objects,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1600–1607.

[15] J. Yang, J. Wright, T. Huang, and Y. Ma, “Image super-resolution via sparse representation,” IEEE Transactions on Image Processing, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[16] A. M. Bruckstein, D. L. Donoho, and M. Elad, “From sparse solutions of systems of equations to sparse modeling of signals and images,” SIAM Review, vol. 51, no. 1, pp. 34–81, 2009.

[17] K. Huang and S. Aviyente, “Sparse representation for signal classification,” in Advances in Neural Information Processing Systems (NIPS), 2007, pp. 609–616.

[18] M. Yang, L. Zhang, J. Yang, and D. Zhang, “Metaface learning for sparse representation based face recognition,” in 17th IEEE International Conference on Image Processing (ICIP), 2010, pp. 1601–1604.

[19] I. Ramirez, P. Sprechmann, and G. Sapiro, “Classification and clustering via dictionary learning with structured incoherence and shared features,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 3501–3508.

[20] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, “Supervised dictionary learning,” in Advances in Neural Information Processing Systems (NIPS), 2008, pp. 1033–1040.

[21] Q. Zhang and B. Li, “Discriminative K-SVD for dictionary learning in face recognition,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2691–2698.

[22] D. S. Pham and S. Venkatesh, “Joint learning and dictionary construction for pattern recognition,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[23] J. Mairal, F. Bach, and J. Ponce, “Task-driven dictionary learning,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pp. 791–804, Apr. 2012.

[24] Z. Jiang, Z. Lin, and L. S. Davis, “Learning a discriminative dictionary for sparse coding via label consistent K-SVD,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1697–1704.

[25] W. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin, “Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images,” IEEE Transactions on Image Processing, vol. 21, no. 1, pp. 130–144, Jan. 2012.
[26] B. Babagholami-Mohamadabadi, A. Jourabloo, M. Zolfaghari, and M. Manzuri-Shalmani, “Bayesian supervised dictionary learning,” in Proceedings of the Uncertainty in Artificial Intelligence (UAI) Workshop: Big Data Meet Complex Models, 2013, pp. 11–19.

[27] M. J. Gangeh, A. Ghodsi, and M. S. Kamel, “Kernelized supervised dictionary learning,” IEEE Transactions on Signal Processing, vol. 61, no. 19, pp. 4753–4767, Oct. 2013.

[28] H. Zhang, Y. Zhang, and T. S. Huang, “Simultaneous discriminative projection and dictionary learning for sparse representation based classification,” Pattern Recognition, vol. 46, no. 1, pp. 346 – 354, 2013.

[29] S. Lazebnik and M. Raginsky, “Supervised learning of quantizer codebooks by information loss minimization,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 7, pp. 1294–1309, July 2009.

[30] F. Rodriguez and G. Sapiro, “Sparse representations for image classification: learning discriminative and reconstructive non-parametric dictionaries,” in IMA Preprint Series 2213, 2007.

[31] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” IEEE Transactions on Signal Processing, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[32] C. R. Rao and S. K. Mitra, Generalized Inverse of Matrices and Its Applications. New York: Wiley, 1971.

[33] H. Lütkepohl, Handbook of Matrices. John Wiley & Sons, 1996.

[34] R. M. Pringle and A. A. Rayner, Generalized inverse matrices with applications to statistics. New York: Hafner Pub. Co., 1971.

[35] A. Gretton, R. Herbrich, A. J. Smola, O. Bousquet, and B. Schölkopf, “Kernel methods for measuring independence,” Journal of Machine Learning Research, vol. 6, pp. 2075–2129, Dec. 2005.

[36] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, “Measuring statistical dependence with hilbert-schmidt norms,” in Proceedings of the 16th international conference on Algorithmic Learning Theory (ALT), 2005, pp. 63–77.

[37] N. Aronszajn, “Theory of reproducing kernels,” Transactions of the American Mathematical Society, vol. 68, no. 3, pp. 337–404, 1950.