A New Sparse Dimension Reduction Method for Hyperspectral Images

Jing-zun ZHANG1,2,*, Li-ping LIAO1,3, Guang-mei XU1, Rui-zhe ZHANG1 and Ning HE1

1 Smart City College, Beijing Union University, Beijing, CO 100101 China
2 Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, CO 10094 China
3 Beijing Jiaotong University, Beijing, CO 10044 China

*Corresponding author

Keywords: Hyperspectral image, Dimension reduction, Double sparse& unsupervised.

Abstract. Motivated by the development of sparse representation, many sparsity-based methods have been successfully applied in Hyperspectral images (HSI) domain. These methods are beneficial to find the intrinsic structure from the high dimension data. Sparse dimension reduction is becoming one of the research hotspots, more efforts are needed to be further. In this paper, we present a new unsupervised dimension reduction (DR) method, which is called unsupervised double sparse learning method (UDSDL). UDSDL inherits the merits of sparse representation, it can support the possibility of data compressing while preserving more discriminative features. Experiments on a real HSI data set collected by the Airborne visible/Infrared Imaging Spectrometer are performed to demonstrate the effectiveness of the proposed UDSDL method.

Introduction

With its numerous and contiguous spectral bands, Hyperspectral Images (HSI) can provide detailed spectral information of the observed materials. It is nowadays considered as one of the most powerful remote sensing tools for acquiring precise information of the Earth’s surface. However, the increased volume of data contained in the HSI entails inconvenience and inefficiency in information storage, display, transmission, and processing for researchers [1][2]. Chang found that up to 90% of the spectral band is not necessary and can be neglected, without affecting the accuracy of the overall information interpretation [3]. It is desirable to develop a DR technique that can automatically find a discriminative and physically meaningful representative of the whole image cube without loss of information.

Sparsity-based methods [4] are the new data mining algorithms that can overcome the classical DR ones’ insufficiency under very high scenario but with very few applications in HSI domain. Therefore, we make a further study of sparsity-based models, present a new unsupervised HSI DR method by introducing $l_1 – norm$ regularization on loadings and elastic net constraint on the learned features. This method can provides additional data compression mechanism while preserving the discriminative information. The reminder of this paper is organized as follows: Section 2 presents the notations and briefly reviews the sparsity-based DR methods from a matrix factorization point of view and introduces two representatives of them. By analyzing the pros and cons of the above two, we give details of our model in Section 3. To demonstrate the effectiveness of the proposed model, extensive experimental results on real HSI dataset are reported and analyzed in Section 4. Section 5 concludes this paper and discusses paths for future research.

Related Work

Notation: In this paper, vectors are denoted by bold lower-case letters, matrices by bold upper-case ones, scalars by lower-case. All vectors are treated as column vectors. For instance, we consider in the rest of this paper a vector $x$ in $R^n$ and a matrix $X$ in $R^{m \times n}$. The columns of $X$ are represented by indexed vectors $x_1, \ldots, x_n$ such that we can write $X = [x_1, \cdots, x_n]$. The i-th entry is denoted by $x[i]$.
and the i-th entry of j-th column of X is represented by x[i,j]. Same as the other literature, we define the \( l_0 - norm \) of vector x as \( \|x\|_0 = \#\{i, s.t. x[i] \neq 0\} \), the \( l_1 - norm \) of vector x is: \( \|x\|_1 = \sum_{i=1}^{m} |x_i| \), the Frobenius norm of matrix X as \( \|X\|_F = (\sum_{i=1}^{m} \sum_{j=1}^{n} x_{i,j}^2)^{1/2} \).

For a given matrix \( X \in R^{m \times n} \) with \( r \ rank(X) = k \), classical principal component analysis (PCA) seeks to find two smaller matrices D and A that satisfy the following formulation.

\[
X = DA, \text{with } D \in R^{m \times k}, A \in R^{k \times n} \quad (k < n)
\]

It can be rewritten as the following formulation,

\[
\min_{D \in R^{m \times k}, A \in R^{k \times n}} \|X - DA\|_F^2 \quad (2)
\]

Matrix D with lower dimension contains the latent structure of matrix X and it can be seen as a representative of X.

With potential sparsity requirements on D and A of model (2), we can get sparse solutions, which is very common in statistics and signal processing domain. Formulated as follows:

\[
\min_{D \in C, A \in R^{k \times n}} \frac{1}{2} \|X - DA\|_F^2 + \lambda \psi(A) \quad (3)
\]

where, matrix A carries the decomposition coefficients of the signals X, each signal \( x_i \) form X can be represented as \( x_i \approx DA_i = \sum_{j=1}^{k} a_i[j]d_j \). \( \psi \) is sparsity-inducing regularization function, columns of A will become sparse under its control. And C is the constraints on D, for example, C = \( \{D \in R^{m \times k} : \forall j \|d_j\|_2 \leq 1\} \). Parameter \( \lambda \) controls the trade-off between data fitting and sparsity.

The above sparse model can be seen as a matrix factorization problem with sparse regularization, which has been proven very effective for feature extraction [29]. It is also flexible enough to accommodate a large set of constraints and regularizations, and has gained significant attention in scientific domains where dimension reduction is a key aspect [30]. Two representative of it are as following.

Zou and Hastie [5] imposed a method called SPCA, formulating PCA as a regression-type optimization problem and using the elastic-net regulation to produce modified principal components with sparse loadings.

\[
\min_{A \in R^{m \times k}, B \in R^{n \times k}} \sum_{i=1}^{n} ||x_i - AB^T x_i||^2 + \lambda \sum_{j=1}^{k} ||\beta_j||^2 + \sum_{j=1}^{k} \lambda_j ||\beta_j||_1 \quad s.t. A^T A = I_{k \times k} \quad (4)
\]

An alternate optimization strategy was used in SPCA to get the solutions of sparse loading matrix B and modified principal components matrix A. Four main procedures were included in their algorithm:

1) Apply conventional PCA on X and initialize A with the first k loadings,
2) Solve an elastic-net problem to get solution B with respect to fixed A
3) Compute the SVD of \( X^T X \)
4) Repeat 2) and 3) under a pre-defined convergence criteria.

SPCA has proved to be a very efficient dimension reduction method for synthetic dataset and gene expression dataset. However, as we can find that SVD decomposition on the whole dataset was needed in every iteration, making more complexity and limited information extracted from the original dataset. This can be found in the later experiment section.

Haipeng Shen et. al [6] proposed sPCA-rSVD method as another choice for sparse dimension reduction. They used the connection of PCA with SVD to extract the PCs through solving a low rank matrix approximation problem. Their model formulation is (5).

\[
\|X - \tilde{u} \tilde{v}^T\|_F^2 + P_\lambda(\tilde{v}) \quad (5)
\]
sPCA-rSVD can accommodate different form of sparse regularizations (\( P_\lambda \)) on the targeted feature (\( \tilde{v} \)). Different from SPCA, it didn’t obtain the desired number of principal feature simultaneously, but one by one. Concretely, they get the first \( \tilde{u}_1 \) and \( \tilde{v}_1 \) via the rank-one
approximate of $X$, and then get the second from another rank-one approximate of the residual matrix $X_1 = X - \tilde{u}_1 \tilde{v}_1$.

**Proposed Method**

In this section we formulate our method also form the matrix factorization view and introduce different regularization strategies to illustrate a new way of feature extraction from HSI. Because the dimension reduction procedure doesn’t need the label information of the HSI image and can support sparsity tuning possibility on both the loadings and the features, we name it UDSDL (Unsupervised Double Sparsity Dimension Learning). Details of UDSDL are as follows.

Performing conventional PCA on the given dataset $X \in R^{m \times n}$, we can get its first $k$ largest principal components by solving the following formulation:

$$\min_{U \in R^{m \times k}, V \in R^{n \times k}} \|X - UV\|_F^2 \quad \text{s.t.} \quad U^T U = I_k, \ (k < n)$$

Reform it to the same as the sparse model:

$$\min_{D \in R^{m \times k}, A \in R^{m \times k}} \|X - DA\|_F^2, \ C = \{D \in R^{m \times k} : \forall i \neq j \ d_i^T d_j = 0\}$$

(7)

Here, $D = U$ and $A = V^T$, conventional PCA with orthogonal requirement transforms to a matrix factorization model with constraint $C$.

Sparse solutions can be obtained by enforcing sparsity-inducing regularizations on $A$ or $D$ or even on both. Mairal[7] proofed that these sparse matrix $D$ is or is equivalent to the latent features of $X$. Then, dimension reduction with sparse constraint can be cast to the following sparse model.

$$\min_{D \in C, A \in R^{k \times n}} \|X - DA\|_F^2 + \lambda \psi(A), \ C = \{D \in R^{m \times k} : \forall i \neq j \ d_i^T d_j = 0\}$$

(8)

It can be performed by minimizing a quadratic data-fitting term, with constraints and/or penalties over the loadings $\alpha$ and the features $D$.

$$\min_{A = [\alpha_1, \ldots, \alpha_n] \in R^{k \times n}} \sum_{i=1}^{n} \frac{1}{2} \|x_i - DA_i\|_2^2 + \lambda \varphi(\alpha_i)$$

(9)

where $\|\|_2$ is the Euclidean norm, $C$ is a convex set of $R^{m \times k}$, and $\varphi: R^k \rightarrow R$ is a penalty (equal to the above mentioned sparsity-inducing regularization) over the loadings to enforce sparsity.

In general, the constraint set $C$ is traditionally a technical constraint ensuring that the coefficients $\alpha_i$ do not vanish, making the effect of the penalty disappear. Choice of constraint set and penalties typically offers some flexibility for a specific problem; for instance, the most used $l_1 - \text{norm}$ ($\varphi = \|\|_1$), leading to sparse coefficients $\alpha$ and $l_2 - \text{norm}(\varphi = \|\|_2)$ constraint on $D$ can ensure the reasonability of solution $A$.

Following the above analysis, we use $l_1 - \text{norm}$ on $A$ and elastic-net (used in Zou and Hastie 2006 to get sparse loadings, with the form $\|d\|_2^2 + \gamma \|d\|_1 \leq 1$) constraints on $D$, leading to sparse coefficients $\alpha$ and sparse features $d$ as the same time. Our model can be formulated as follows:

$$\min_{D \in C, A \in R^{k \times n}} \frac{1}{2} \|X - DA\|_F^2 + \lambda \sum_{i=1}^{n} \|\alpha_i\|_1, \ C = \{D \in R^{m \times k} : \forall j \ (d_j)^2 + \gamma \|d_j\|_1 \leq 1\}, k < n$$

(10)

Obviously, this model can introduce sparsity both on $A$ and $D$ controlled by the tuning parameter $\lambda$ and $\gamma$. This model isn’t convex with both $D$ and $A$ simultaneously, but is convex with $A$ when keeping $D$ fixed and vice versa. Thus, an alternative—optimization strategy is the optimal choice to solve it, as proposed by Lee et al [8].

For fixed $D, A = [\alpha_1, \ldots, \alpha_n]$ can be obtained by solving the typical $l_1 - \text{normal}$ sparse coding problem:

$$\alpha_i = \arg \min_{\alpha \in R^{k \times 2}} \frac{1}{2} \|x_i - D\alpha\|_2^2 + \lambda \|\alpha\|_1$$

(11)

There exist three efficient methods to the above optimization problem: coordinate descent [9][10][11], proximal gradient [12][13], LARS (Least angle Regression) [14][15]. Details of each
method can be found by relevant reference, due to the limitation of space we omit the analysis of them. For computing complexity consideration we choose the LARS method.

Once a sparse solution of $A$ is found via LARS, the next step is to solve the optimization with respect to $D$. For fixed $A$, finding the optimal $D$ is a classical dictionary learning problem also with a few approaches, such as stochastic gradient descent [16][17], MOD[8], K-SVD[18], online dictionary learning [7].

Based on stochastic approximations, online dictionary learning method proposed by Mairal et al. has two obvious advantages: scaling up gracefully to large datasets and faster performance than the others. These two characters just satisfy the high performance requirement of HSI processing under the context of big data. That is the exact reason we choose it as our foundation and make adaptive improvement to find the sparse features by changing the $l_2-norm$ constraint to elastic-net with corresponding solving algorithm named LARS-EN [19]. This improvement can not only enforce sparsity, but also ensure stability, even if it seems simple. As the new coefficient matrix $A$ has been obtained, we can update each columns of matrix $D$ iteratively and get the new $D$ satisfying the following formulation.

$$D_t = \arg \min_{D \in C} \frac{1}{t} \sum_{i=1}^{t} (\frac{1}{2} \|x_i - D \alpha_i\|_2^2 + \lambda \|\alpha_i\|_1), C = \{D \in R^{m \times k} \colon \forall j \|d_j\|_2^2 + \gamma \|d_j\|_1 \leq 1\}$$

By summarizing the above analyses, we can give the UDSDL algorithm as follows.

| Step | No. |
|------|-----|
| Initialization: $D^0 \in R^{m \times k}$ with random values, $A^0 \in R^{k \times k} \leftarrow 0, B^0 \in R^{m \times k} \leftarrow 0$ | 1 |
| Input: $k, T, A^0, B^0, \lambda, X \in R^{m \times n}$ (each column with zero mean and $l_2-norm$ unit), $\gamma$ | 2 |
| for $t=1$ to $T$ do | 3 |
| Random draw a column vector $x_t$ from $X$, solve following sparse coding by LARS |
| $\alpha_t = \arg \min_{\alpha \geq 0} \frac{1}{2} \|x_t - D^{t-1} \alpha\|_2^2 + \lambda \|\alpha\|_1$ | 4 |
| Compute: | 5 |
| $A^t \leftarrow A^{t-1} + \alpha_t A_T^T$ | 6 |
| $B^t \leftarrow B^{t-1} + \alpha_t d_T^T$ | 7 |
| Compute $D^t$ by updating each column of $D^{t-1}$ as follows: | 8 |
| for $j=1$ to $k$ do | 9 |
| $u_j \leftarrow \frac{1}{A^{t-1}} (b_j - D \alpha_j) + d_j$ | 10 |
| $d_j \leftarrow \arg \min_{d \in R^n} \|u_j - d\|_2^2$ s.t. $\|d\|_2^2 + \gamma \|d\|_1 \leq 1$ (Solved by LARS-EN) | 11 |
| end for | 12 |
| Output: $D \in R^{m \times k}$ | 13 |

$X \in R^{m \times n}$ is the original HSI data with high dimensionality and each column of it is first centered and normalized to have unit $l_2-norm$, $D \in R^{m \times k}$ is the desired low dimensional matrix, scalar $T$ is the pre-defined iteration number for optimization, parameter $\lambda$ is used to tune the sparsity of matrix $A$ and $\gamma$ is used to tune the sparsity of matrix $D$. Two temporary matrices $A^t = \sum_{i=1}^{t} \alpha_i a_i^T, B^t = \sum_{i=1}^{t} x_i a_i^T$ store the vectors produced in each iteration and they are initialized to zero in the beginning. Each column of the desired matrix $D$ will be updated one by one by the new sparse coefficient vector $\alpha$ in every iteration.

**Experiments**

**Competitors and Evaluation Method**

In this section, we validate our proposed method with three typical feature extraction methods as competitors: first is PCA, the de facto technique with no sparsity regularization or constraint, the other two are sparse-oriented: SPCA[5] and sPCA-rSVD [6] respectively.

Without loss of universality, we evaluate the performance of these DR (dimension reduction) techniques by the quantized classification results conducted on the low dimensional data.
Support vector machine (SVM)[20] was selected as the classifier, as it performs well even with a limited number of training samples. Three quantitative measurements are used to evaluate the DR results: OA (overall accuracy), AA (average accuracy), and the Kappa coefficient (kappa).

**HSI DataSet**

One HSI benchmark dataset named Indian Pines(https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html) was used in our experiment. False color composite image of the Indian Pines dataset and the ground truth can be seen in Figure 3.

**Experimental Setup**

In each experiment, we first use different DR methods on the HSI data, then, the support vector machine (SVM) classifier with polynomial kernels is employed for classification on the result low dimensions of each method. We use the MATLAB SVM Toolbox, LIBSVM[20], in our experiments. The SVM parameters (the degree d and the penalty factor C) are selected by grid search method within the same given set \(\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}\). We use the implementation of LARS and LARS-EN, provided by another MATLAB Toolbox SpaSM (http://spams-devel.gforge.inria.fr/).

We use the regularization tuning parameter \(\lambda = 1.2/\sqrt{m}\) in all of our experiments. The \(1/\sqrt{m}\) is a classical normalization factor [21] and the constant 1.2 has shown to yield about 20 nonzero coefficients for Indian Pines Set. For faire comparison, we set the SPCA and the sPCA-rSVD have the same number of nonzero loadings.

The second tuning parameter is responsible for the controlling of the number of zero entries of columns in the learned feature matrix D. We will experimentally explore its effect on real HSI dataset.

Parameter \(\gamma\) is responsible for controlling the number of zero entries of the desired feature matrix D. We will experimentally explore its effect on real HSI dataset.

In each experiment, we randomly choose 10% of the related samples for training and the remaining 90% for testing. Specific classes and the selected number of training and testing data in each class are reported in Table 2.

**Experimental Results and Analysis**

First we evaluate the applicability of the parse based methods on AVIRIS Indian Pines.

As neither SPCA nor sPCA-rSVD has the sparsity tuning ability in components (or features), we first set the UDSDL parameter \(\gamma=0\) to vanish the sparse constraint on D. Different number of dimensions (range in \([5 10 15 20 25 30 35 40 45 50 55 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200]\)) was extracted from the original dataset by four DR methods: PCA, SPCA, sPCA-rSVD and UDSDL.

Changing curve of overall accuracy (in percent) corresponding to different dimensions extracted by DR method is shown in Figure 1 and the partially enlarged one is in Figure 2. Different colors are used to represent the result of each method. The curve with black color represent the classification result from the original data without dimension reduction, which is used as a baseline. Corresponding legend of color and method can be found on the right bottom of each figure.
From the above two figures we can see:

1) The overall trend of each method’s performance is upward, but the growth rate is very slow after reaching a certain dimension.

2) PCA method (blue curve with asterisk) presents optimal performance in most dimensions but without any sparse tuning strategy.

3) UDSDL (red curve with hollow circle) can obtain better OA than baseline after 20 dimensions and exceeds both SPCA (green curve with short vertical line) and sPCA-rSVD (yellow dot dash curve) under any dimension.

We tabulate the quantized AA, OA Kappa and accuracy for each classes in Table 2 according to their peak performance. Numbers in bold font represent the optimal result under different methods. Two observations can be found from the quantized results:

1) UDSDL obtains the sub-optimal OA and Kappa but the best AA among the four DR methods.

2) For the classes with very few training samples (see classes of 1, 7, 9), classification accuracy of UDSDL exceeds that of PCA.

Table 2. Classification accuracy for Indian Pines Scene (optimal dimensions in bracket).

| No | Name          | Train | Test  | original | PCA(20) | SPCA(20) | sPCA-rSVD(20) | UDSDL(20) |
|----|---------------|-------|-------|----------|---------|----------|---------------|-----------|
| 1  | Alfalfa       | 6     | 54    | 66.6667  | 68.7500 | 67.3750  | 72.9167       | 72.9167   |
| 2  | Corn-no till  | 144   | 1434  | 77.7519  | 80.0775 | 80.0713  | 76.7442       | 78.9147   |
| 3  | Corn-min till | 84    | 834   | 79.2000  | 82.6667 | 77.4853  | 68.4000       | 73.6000   |
| 4  | Corn          | 24    | 234   | 82.8571  | 70.9524 | 76.5333  | 76.6667       | 79.5238   |
| 5  | Grass/pasture | 50    | 497   | 95.0783  | 87.2483 | 94.0537  | 93.5123       | 94.4072   |
| 6  | Grass/tree    | 75    | 747   | 97.1726  | 94.1964 | 91.7292  | 89.8810       | 93.7500   |
| 7  | Grass/pasture-mowed | 3 | 26   | 95.6522  | 95.6522 | 93.7391  | 95.6522       | 95.6522   |
| 8  | Hay-windrowed | 49    | 489   | 96.3636  | 97.9545 | 95.3273  | 94.7727       | 96.5909   |
| 9  | Oats          | 2     | 20    | 50.0000  | 33.3333 | 49.0000  | 50.0000       | 61.1111   |
| 10 | Soybeans-no till | 97   | 968   | 73.1343  | 82.8932 | 76.6223  | 69.6800       | 68.4271   |
| 11 | Soybeans-min till | 247  | 2468  | 79.7839  | 84.1063 | 81.3652  | 81.9451       | 84.4665   |
| 12 | Soybeans-clean till | 62  | 614   | 84.9638  | 85.1449 | 83.2645  | 78.4420       | 87.6812   |
| 13 | Wheat         | 22    | 212   | 99.4737  | 99.4737 | 97.4842  | 99.4737       | 99.4737   |
| 14 | Woods         | 130   | 1294  | 94.5017  | 93.2990 | 93.1168  | 96.7354       | 94.2440   |
| 15 | Bldg-grass-tree-drives | 38   | 380   | 57.6023  | 50.5848 | 50.7193  | 40.6433       | 51.4620   |
| 16 | Stone-steel towers | 10  | 95    | 91.7647  | 84.7045 | 83.0945  | 81.0898       | 83.4174   |

OA

83.4281  84.7045  83.0945  81.0898  83.4174

AA

82.6229  80.6899  81.1135  79.7504  82.6756

kappa

0.8115  0.8259  0.8102  0.7837  0.8106
Alignment the above curves, quantized and visual results of the four methods we can find that although the UDSDL method proposed in this paper is slightly lower than PCA in OA and Kappa indexed, the average accuracy of all classes and accuracy of some classes, especially those with very small number of training set are higher than it. This fully proves the ability of UDSDL to extract discriminant features from small training sets, which is very important for HSI. At the same time, different form the SPCA and sPCA-rSVD, UDSDL introduce the random approximation strategy which can avoid high time-consuming operations such as eigenvalue solution and SVD decomposition and is more applicable in the context of large amounts of data.

![Image](attachment:image1.png)

Figure 3. Classification maps for Indian Pines (dimensions needs for best OA in bracket).

Further, we explore the possibility of enforcing sparsity on features (according to matrix D in model 11) and preserving the sparsity of A at the same time. Here we tune the parameter $\gamma$ range from 0 to 0.5 with 0.02 step increasement of each test, use the proportion of zero entries in feature matrix as a measure of sparsity. Figure 4 presents the curve of OA changing with the sparsity of features.

We can see that UDSDL can achieve the highest OA with no sparsity constraint and with a slightly fallen according to less sparsity and keeps a relatively stable performance in a fixed sparsity range between [0.05, 0.5]. This demonstrates that we can have the privileged choice of balancing among the accuracy and storage demanding. In some situation, on-board processing of HSI for example, fast processing requirements under limited storage space are more important than the precise classification results and UDSDL provides the possibility that none of the other method can.

Conclusions
In this paper, we present a sparse dimension reduction method UDSDL for HSI. This method can leaning the latent sparse structure from high dimension data in an unsupervised mode. Comparing to this kind of two typical sparse methods SPCA and sPCA-rSVD, our method can learn more discriminative features with the same sparsity and can enforce double sparsity when necessary. Comparing to the de facto DR technique, PCA, the experimental results on real HSI dataset indicate that UDSDL can achieve the very close performance with much more sparse and less volume of
data in terms of classification accuracy. Future work will involve exploring the parallel version of this method for further speedup.

Figure 4. SVM classification of Indian Pines (OA versus different sparsity).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61872042); the Key Project of the Beijing Education Commission (KZ201911417048). supported in part by the 2018 visiting scholar project and 2019 scientific research project of Beijing Union University (ZK50201903).

References

[1] Q. Du, J. E. Fowler, and W. Zhu “On the impact of atmospheric correction on lossy compression of multispectral and hyperspectral imagery,” IEEE Trans. Geoscience and Remote Sensing, vol. 47, no. 1, pp. 130-132, 2009.

[2] H. Chen et al., “A BOI-preserving-based compression method for hyperspectral images,” IEEE Trans. Geoscience and Remote Sensing, vol. 48, no. 11, pp. 3913-3923, Nov. 2010.

[3] C. I. Chang, “Hyperspectral Data Exploitation: Theory and Applications,” New York, USA: John Wiley Sons. 2007.

[4] v. Vu an J. Lei (2012, March). Minimax Rates of Estimation for Sparse PCA in High Dimensions. presented at AISTATS.

[5] H. Zou, T. Hastie and R. Tibshirani “Sparse Principal Component Analysis,” Journal of Computational and Graphical Statistics, vol. 15, no. 2, pp. 265-286, 2006.

[6] H. Shen and J. Z. Huang “Sparse principal component analysis via regularized low rank matrix approximation,” Journal of Multivariate Analysis, vol. 99, no. 6, pp. 1015-1034, 2008.

[7] J. Mairal et al., “Online learning for matrix factorization and sparse coding,” Journal of Machine Learning, Research, vol. 11, pp. 19-60, Jan. 2010.

[8] H. Lee et al.. (2007).Efficient sparse coding algorithms. presented at NIPS.

[9] W.J. Fu, “Penalized regressions: the bridge versus the lasso,” Journal of Computational and Graphical Statistics, vol. 7, no. 3, pp. 397-416, 1998.

[10] P. Tseng and S. Yun, “A coordinate gradient descent method for nonsmooth separable minimization,” Mathematical Programmin, vol. 117, no. 1, pp. 387-423, 2009.

[11] T. T. Wu and K. Lange, “Coordinate descent algorithms for lasso penalized regression,” Annals of Applied Statistics, vol. 2, no. 1, pp. 224-244, 2008.
[12] R. D. Nowak and M. A. T. Figueiredo. (2001, November). Fast wavelet based image deconvolution using the EM algorithm. presented at IEEE ASILOMAR.

[13] I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” Communications on Pure and Applied Mathematics, vol. 57, no. 11, pp. 1413-1457, 2004.

[14] M. R. Osborne, B. Presnell, and B. A. Turlach, “A new approach to variable selection in least squares problems,” IMA journal of numerical analysis, vol. 20, no. 3, pp. 389-403, 2000.

[15] B. Efron et al., “Least angle regression,” Annals of Statistics, vol. 32, no. 2, pp. 407-499, 2004.

[16] H. J. Kushner and G, “Stochastic approximation and recursive algorithms and applications,” Springer Science Business Media, 2003.

[17] L. Bottou et al. (2008). The tradeoffs of large scale learning. presented at NIPS.

[18] M. Aharon, M. Elad, and A. M. Bruckstein, “The K-SVD: An algorithm for designing of over complete dictionaries for sparse representations,” IEEE Trans. Signal Processing, vol. 54, no. 11, pp. 4311-4322, Nov. 2006.

[19] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” IEEE Trans. Signal Processing, vol. 58, no. 4, pp. 2121-2130, Jan. 2010.

[20] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” Data Mining Knowl. Discovery, vol. 2, no. 2, pp. 121-167, 1998.

[21] P. J. Bickel et al. (2009). Simultaneous analysis of Lasso and Dantzig selector. presented at Annals of Statistics.