計測自動制御学会東北支部 第 284 回研究集会 (2013.11.8)
資料番号 284-11

Sparse Representation by Nonnegative Matrix Factorization with the Itakura-Saito Divergence and Sparsity Constraint

Zhenni Li* Shuxue Ding* Yujie Li* and Zunyi Tang**

*School of Computer Science and Engineering, The University of Aizu,
**Faculty of Biomedical Engineering, Osaka Electro-Communication University

キーワード: Nonnegative matrix factorization, Sparse representation, Dictionary learning, Itakura-Saito divergence, Multiplicative updates

連絡先: 〒965-8580 348E Room, The University of Aizu, Tsuruga, Ikki- Machi,
Aizu-Wakamatsu City,Fukushima, Japan
Zhenni Li, Tel.: (024)237-2799, Fax.: (024)237-2799, Email:lizhenni2012@gmail.com

***Abstract*—In this paper, we propose a novel and efficient dictionary learning method for sparse representation of signals. The proposed algorithm is based on the nonnegative matrix factorization (NMF). We adopt the Itakura-Saito (IS) divergence as the error function and impose ℓ_1-norm as the sparsity constraint. The error function is quite different from conventional dictionary learning methods using Euclidean (EUC) distance as error function. Numerical experiments show that the proposed algorithm performs better than the other three compared algorithms which all use Euclidean distance as the error function.**

1. INTRODUCTION

Amazing nonnegative matrix factorization (NMF)[1, 2], which is a method for dimensionality reduction and data analysis [3], has attracted great attention over the past few years. The NMF has found a wide variety of application, including machine learning, image processing, blind source separation, *etc*. Given such a representation $\mathbf{Y} \in \mathbb{R}_+^{m\times n}$, the NMF consists in finding a factorization of the form $\mathbf{Y} \approx \mathbf{WH}$, where $\mathbf{W} \in \mathbb{R}_+^{m\times r}$ is termed as the base matrix, and $\mathbf{H} \in \mathbb{R}_+^{r\times n}$ is termed as the coefficient matrix. The factorization is obtained by minimizing error function $D(Y, WH)$ between $\mathbf{Y}$ and $\mathbf{WH}$. For $D(Y, WH)$, the Euclidean (EUC) distance and the generalized Kullbank-Leibler (KL) divergence are often chosen in the literature. People note that nonnegativity constraint could lead to sparse representation which is attracting more and more attention in recent years. The sparse representations, where most coefficients are zero, are proven to be an interesting and powerful tool for analysis and processing of signals. In the model of sparse representation of signals, using an overcomplete dictionary matrix $\mathbf{W} \in \mathbb{R}^{m\times r}$ that contains r atoms of size $m \times 1$ for columns, signals $\mathbf{Y} \in \mathbb{R}^{m\times n}$ can be described by sparse and efficient linear combinations of few atoms. An overcomplete dictionary can lead to sparse representation where overcomplete means $m < r$. $\mathbf{Y} = \mathbf{WH}$ or $\mathbf{Y} \approx \mathbf{WH}$ are two ways to represent $\mathbf{Y}$, where the $\mathbf{H} \in \mathbb{R}^{r\times n}$ is termed as the coefficient matrix which contains the coefficients for representation of signals $\mathbf{Y}$. The popular EUC distance can be chosen as the error function to measure the approximation error between $\mathbf{Y}$ and $\mathbf{WH}$, namely satisfying $\| \mathbf{Y} - \mathbf{WH} \|_2 \leq \varepsilon$.

It is obvious and interesting that dictionary learning, building a dictionary consisting of atoms or subspaces so that a class of signals can be efficiently and sparsely represented in terms of the atoms, is an important topic. In recent practices [4, 5], a learned dictionary has been proved to be critical for achieving superior results in the field of signal and image processing. On the other hand, in some applications the nonnegativity of the signals and the dictionary are required, such as the multilateral data analysis [6] and the nonnegative factorization for recognition [7, 8]. These requirements call for the dictionary learning method imposed with the nonnegativity, namely, so-called nonnegative dictionary learning. Extensive research in this field concentrates mainly on the study of pursuit dictionary learning algorithms. To some extent, sparse representation of nonnegative signals is similar to NMF in the sense that it can increase interpretability and admitting nonnegative combinations of the coefficient matrix can lead to sparse results. Unfortunately, the degree of sparseness can not be controlled and the results may not be enough if only using the nonnegativity. In order to render a sparser representation, some kinds of sparsity-constrained NMF methods, with different constraints imposed on the matrix factors, have been proposed. Expansion speaking, ℓ_0-norm, ℓ_1-norm and ℓ_2-norm constraint have been usually used as the sparsity constraints [9–11]. Since the ℓ_0-norm optimization problem is generally NP-hard, fortunately it can be replaced by ℓ_1-norm for the convenience of optimization in the real-world applications. Some authors also impose sparsity constraints by using ℓ_2-norm constraint[12], because of the particularity of the sparse NMF. The classical methods include NMFSC Hoyer proposed [10] and NMFℓ^0-H [13]. And other classical algorithms is NN-KSVD (nonnegative variant of K-SVD) algorithm [14, 17]. These algorithms all use the EUC distance as the measure of approximation error. In addition, the KL divergence is also used for the measure error[15].

In this paper, we propose using the IS divergence as the measure of approximation error. This divergence is obtained by Itakura-Satio [16] from the maximum likelihood (ML) estimation. The divergence has good properties, in particular scaleinvariant meaning that low energy components of input signal bear the same relative importance as high energy ones. And we use the ℓ_1-norm as sparsity constraint to reach the sparsest representation for the coefficient matrix under the condition of nonnegativity. Then we combine sparsity constraint with IS divergence into a novel algorithm. We used multiplicative updating to develop an algorithm termed as IS divergence with ℓ_1-norm sparsity constraint (ISSC). Gradient multiplicative updating is an efficient way in this case, since it can easily preserve nonnegativity constraint at each iteration. Numerical experiments show that the proposed algorithm can recover almost all aimed dictionary atoms from training data, which is superior over the other algorithms including NMFSC, NN-KSVD, and NMFℓ^0-H.

The remaining part of paper is organized as follows. In section II, we review the relationships between the IS divergence and other error functions used in NMF, especially the Euclidean distance and the generalized Kullback-Leibler (KL) divergence. In section III, we describe proposed algorithm ISSC. In Section IV we present the results of numerical experiments for the proposed algorithm and compare these results with those of several other algorithms. Finally, we give the conclusion and discuss the future work.

2. PROBLEM FORMULATION

The NMF problem is formulated as follows. Given an input matrix $\mathbf{Y} \in \mathbb{R}_+^{m\times n}$, where each element is nonnegative. NMF aims to find nonnegative matrices $\mathbf{W} \in \mathbb{R}_+^{m\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times n}$ satisfying the condition of $\mathbf{Y} = \mathbf{WH}$ or $\mathbf{Y} \approx \mathbf{WH}$, where $r \ll \min(m,\ n)$. This problem can be formulated as the minimization of an objective function,

$$\min \quad D(\mathbf{Y}|\mathbf{WH}) \tag{1}$$

$$\text{subject to} \quad \mathbf{W} \in \mathbb{R}_+^{m\times r},\ \mathbf{H} \in \mathbb{R}_+^{r\times n}$$

In order to complete the approximate factorization, we need to define some error functions to measure quality of the approximation error. Popular choices are the EUC distance which can be defined as,

$$\min D_{EUC}(\mathbf{Y}|\mathbf{WH}) = \frac{1}{2}\|\mathbf{Y} - \mathbf{WH}\|_F^2 \tag{2}$$

and the generalized KL divergence, also termed as I-divergence, can be defined as

$$\min D_{KL}(\mathbf{Y}|\mathbf{WH}) = \sum_{ij}\Big(\mathbf{Y}_{ij}\log\frac{\mathbf{Y}_{ij}}{(\mathbf{WH})_{ij}} - \mathbf{Y}_{ij} + (\mathbf{WH})_{ij}\Big) \tag{3}$$

$$1 \le i \le m, 1 \le j \le n$$

Above two error functions are both nonnegative, and if $\mathbf{Y} = \mathbf{WH}$ the value equal to zero.

In this paper, we consider NMF under other error function which is the IS divergence, whose expression is given by

$$\min D_{IS}(\mathbf{Y}|\mathbf{WH}) = \sum_{ij}\Big(\frac{\mathbf{Y}_{ij}}{(\mathbf{WH})_{ij}} - \log\frac{\mathbf{Y}_{ij}}{(\mathbf{WH})_{ij}} - 1\Big) \tag{4}$$

$$1 \le i \le m, 1 \le j \le n$$

The three mentioned error functions are the special cases of β-divergence [18–20], which show that $\beta = 2$ is the case of EUC distance, $\beta = 1$ is the case of KL divergence, and the IS divergence is a limiting case when $\beta = 0$.

3. THE PROPOSED ALGORITHM

Although NMF yields sparseness for nonnegativity to some extent, it is believed that much sparser representation can be learned by imposing sparsity constraints on the matrix factors. ℓ_0-norm, ℓ_1-norm and ℓ_2-norm constraint have been usually used as the sparsity constraints. The ℓ_0-norm constraint is the sparsest of the three. Unfortunately, solving the ℓ_0-norm constraint optimization problem is generally NP-hard. For solving this problem, one can replace ℓ_0-norm by ℓ_1-norm constraint. In addition, some authors also impose sparsity constraints by using ℓ_2-norm, because of the particularity of the sparse NMF. However ℓ_2-norm constraint is not sparser than ℓ_0-norm and ℓ_1-norm constraint. In this paper, we adopt ℓ_1-norm as sparsity constraint.

Combining the error function of the IS divergence with ℓ_1-norm sparsity constraint, then we propose our problem as the following optimization problem,

$$\min D_{IS}(\mathbf{Y}|\mathbf{WH}) = \sum_{ij}\Big(\frac{\mathbf{Y}_{ij}}{(\mathbf{WH})_{ij}} - \log\frac{\mathbf{Y}_{ij}}{(\mathbf{WH})_{ij}} - 1\Big) + \lambda \parallel \mathbf{H} \parallel_1 \tag{5}$$

$$\text{subject to} \quad \mathbf{W} \in \mathbb{R}_+^{m\times r},\ \mathbf{H} \in \mathbb{R}_+^{r\times n}$$

where $||\mathbf{H}||_1$ means $\sum_{i,j}|\mathbf{H}_{i,j}|$ and $\lambda \ge 0$ is a regularization parameter which can be adjusted for controlling the tradeoff between the approximation error and the sparsity of the coefficient matrix $\mathbf{H}$. One can repeat experiments with different values of λ and determine which value for λ is optimal according to the output results.

For solving the constrained NMF problem, many algorithms have been developed and most of them are structured with iterative strategy, which utilize the fact that the problem can be reduced into two sequential convex nonnegative problems about $\mathbf{W}$ or $\mathbf{H}$ whereas the other of them is regarded as fixed and known. The traditional multiplicative gradient descent approach [1, 19] consists in updating each parameter by multiplying its value at the previous iteration by a certain coefficient, which can be computed by the following formulas,

$$\mathbf{W}_{ij} = \mathbf{W}_{ij} - \varphi_{ij}\frac{\partial D(\mathbf{W},\mathbf{H})}{\partial \mathbf{W}_{ij}} \tag{6}$$

$$1 \leq i \leq m, 1 \leq j \leq r$$

$$\mathbf{H}_{ij} = \mathbf{H}_{ij} - \psi_{ij}\frac{\partial D(\mathbf{W},\mathbf{H})}{\partial \mathbf{H}_{ij}} \tag{7}$$

$$1 \leq i \leq r, 1 \leq j \leq n$$

The gradients of criterion $D_{IS}(\mathbf{Y}|\mathbf{WH})$ with regard to $\mathbf{W}$ and $\mathbf{H}$ can be written as,

$$\frac{\partial D(\mathbf{W},\mathbf{H})}{\partial \mathbf{W}_{ij}} = \Big((\mathbf{WH})^{\cdot(-1)}\mathbf{H}^T - \big((\mathbf{WH})^{\cdot(-2)} \odot \mathbf{Y}\big)\mathbf{H}^T\Big)_{ij} \tag{8}$$

$$\frac{\partial D(\mathbf{W},\mathbf{H})}{\partial \mathbf{H}_{ij}} = \Big(\mathbf{W}^T(\mathbf{WH})^{\cdot(-1)} - \mathbf{W}^T\big((\mathbf{WH})^{\cdot(-2)} \odot \mathbf{Y}\big) + \lambda\Big)_{ij} \tag{9}$$

where $\odot$ and $.(-2), .(-1)$ denote element-wise multiplication and power. If we choose

$$\varphi_{ij} = \frac{\mathbf{W}_{ij}}{\big((\mathbf{WH})^{\cdot(-1)}\mathbf{H}^T\big)_{ij}} \tag{10}$$

$$\psi_{ij} = \frac{\mathbf{H}_{ij}}{\big(\mathbf{W}^T(\mathbf{WH})^{\cdot(-1)} + \lambda\big)_{ij}} \tag{11}$$

Substituting (8),(9),(10),(11) into (6),(7), then the algorithm is obtained as the following updates,

$$\mathbf{W}_{ij} \longleftarrow \mathbf{W}_{ij} \cdot \frac{\Big(\big((\mathbf{WH})^{\cdot(-2)} \odot \mathbf{Y}\big)\mathbf{H}^T\Big)_{ij}}{\Big((\mathbf{WH})^{\cdot(-1)}\mathbf{H}^T\Big)_{ij}} \tag{12}$$

$$\mathbf{H_{ij}} \longleftarrow \mathbf{H}_{ij} \cdot \frac{\Big(\mathbf{W}^T\big((\mathbf{WH})^{\cdot(-2)}\big)\Big)_{ij}}{\Big(\mathbf{W}^T(\mathbf{WH})^{\cdot(-1)} + \lambda\Big)_{ij}} \tag{13}$$

Obviously, we can observe that the criterion is still nonincreasing under updates (12) and (13). In addition, the convergence had been theoretically proven by the paper [21] based on the expectation maximization (EM) algorithm.

According to the analysis above, the proposed IS divergence with ℓ_1-norm sparsity constraint (ISSC) is summarized in Algorithm 1.

Algorithm 1 ISSC

```
Require: Data Matrix $\mathbf{Y} \in \mathbb{R}_+^{m \times n}$
    1) Initialize $\mathbf{W} \in \mathbb{R}_+^{m \times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r \times n}$ as random nonnegative matrices;
    2) Scale rows of $\mathbf{H}$ to sum to one;
    Normalize columns of $\mathbf{W}$ to a unit $\ell^2$-norm
    3) Iterate until converge or stop;
        $\mathbf{W} \leftarrow \mathbf{W} \odot \frac{[(\mathbf{WH})^{\cdot(-2)} \odot \mathbf{Y}]\mathbf{H}^T}{(\mathbf{WH})^{\cdot(-1)}\mathbf{H}^T + \delta}$
        $\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T[(\mathbf{WH})^{\cdot(-2)}]}{\mathbf{W}^T(\mathbf{WH})^{\cdot(-1)} + \lambda + \delta}$
      $\delta = 10^{-9}$ (overcoming the numerical instabilities)
    4) Scale rows of $\mathbf{H}$ to sum to one and normalize columns of $\mathbf{W}$ to a unit $\ell^2$-norm: $\sum_{i=1}^{m} w_{ij}^2 = 1, \forall j$.
```

4. NUMERICAL EXPERIMENTS

Dictionary learning for sparse representation of signals has been successfully applied in many areas, e.g. data classification and image processing. Especially in the latter area, it has been applied for image denoising [22], image inpainting [23], image compression [14, 24], and superresolution reconstruction [25], *etc*. In this section, we made experiments by using ISSC algorithm for synthetic signals, to test whether this algorithm can recover the original dictionary and to compare its results with other algorithms.

4.1Experiment on Synthetic Signals Generated with a Dictionary

For our experiments, we begun with generating a stochastic nonnegative matrix $\mathbf{W}$ of size 20×50 with i.i.d. uniformly distributed entries. Each column was normalized to the unit ℓ^2-norm. The stochastic matrix was referred to as the true dictionary $\mathbf{W}$, which was not used in the learning but only for evaluation. Then we synthesized 1500 test signals of dimension 20, each of which was produced by a linear combination of three different atoms in the true dictionary $\mathbf{W}$. These test signals are uniformly distributed i.i.d coefficients in random and independent locations. We then synthesized 1500 test signals $\mathbf{Y}$ of dimension 20, each of which was produced by a linear combination of three different atoms in the true dictionary, with three corresponding coefficients in random and independent positions. The uniformly distributed noise of varying signal-to-noise ratio (SNR) for performance analysis of noise-robustness was added to the experiments.

4.2 Applying the ISSC

The initialized dictionary was composed of the signals selected randomly from 1500 test signals. The corresponding coefficients were initialized with random entries that were i.i.d. uniformly distributed and nonnegative. The maximum number of iterations was set to 200. For ISSC, the sparsity of the coefficient matrices was adjusted via the regularization parameters λ. The parameter λ could be determined off-line calibrating. We repeated the experiment with different λ and determined which value for λ was optimal according to the output results. In the our algorithm ISSC λ was set to 0.015. Then we executed ISSC on the test signals to estimate learned $\mathbf{W}$ and evaluate its accuracy by comparing with the true $\mathbf{W}$.

4.3 Comparison with the Other Algorithms

Since NN-KSVD, NMFSC and NMFℓ^0-H were the three state-in-art algorithms for nonnegative dictionary learning, we compared our algorithm with these algorithms. We executed NMFSC, NN-KSVD, NMFℓ^0-H using the same test signals with ISSC, respectively. Then it is easy to estimate learned $\mathbf{W}$ and evaluate every accuracy by comparing with the true $\mathbf{W}$. For the three algorithms, the initialized dictionary matrices of size 20×50 were composed of the randomly selected parts of the test signals. Note that for NMFSC the corresponding coefficient matrices were initialized with i.i.d. uniformly distributed random nonnegative entries. NN-KSVD and NMFℓ^0-H did not require a specified coefficient matrix, as they could generate the corresponding coefficient matrix by sparse coding. The implementation of NN-KSVD[1] algorithm is online available. We executed the NN-KSVD algorithm for a total number of 200 iterations. Matlab code for NMFSC [2] and NMFℓ^0-H[3] algorithms are also online available. The learning procedure with NMFSC was stopped after 3000 iterations because it converged fairly slower than the other algorithms. And the maximum number of iterations of NMFℓ^0-H algorithm was fairly set to 200. It was worth noting that, in the experiment, NN-KSVD and NMFℓ^0-H needed the specified exact number of non-zero elements in the coefficient matrix (3/50=0.06 for the case), while NMFSC was executed with a sparsity factor of 0.85 on the coefficients.

4.4 Results of Experiment

The learned dictionaries were compared with the true generated dictionary. The comparisons were done as described in [14] by sweeping through the columns of the true and the learned dictionaries and finding the closest column (in ℓ^2-norm distance) between the two dictionaries. A distance less than 0.01 was considered as a success. All trials were repeated 15 times. In the experiment, the ISSC algorithm could recovery averaged 9.07%, 69.2%, 88.13% and 93.07% atoms under the noise levels of 10 dB, 20 dB, 30 dB, and in the noiseless case, respectively. For NN-KSVD could obtain averaged 15.7%, 68.0%, 82.9% and 86.5%. While NMFℓ^0-H could recover 23.7%, 80.8%, 84.9% and 84.0% atoms under the same conditions. For NMFSC, it recovered only averaged 0.4%, 13.5%, 38.4% and 49.3% atoms. The detailed results of the experiment for these algorithms are shown in the Fig. 1. The proposed ISSC performed best on dictionary learning in levels of 30dB and noiseless case.

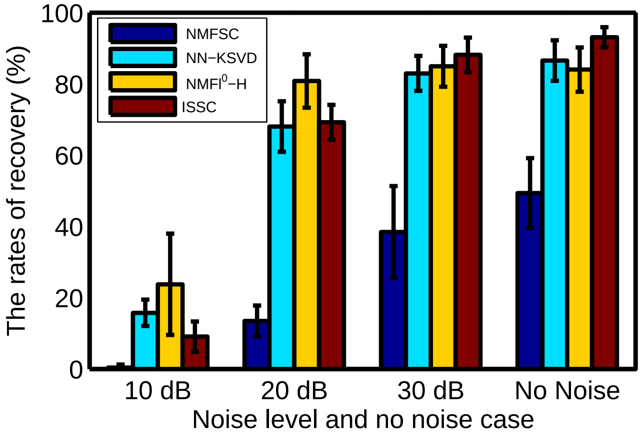

Fig. 1. Experiment result when apply to the synthetic signal: for each of the tested algorithms and for each noise level, 15 trials were performed and their results were sorted. The averaged recovery rates of learned atoms and corresponding deviation of recovery rates are displayed. $\lambda = 0.015$

5. CONCLUSIONS

In this paper we have presented a novel and efficient nonnegative dictionary learning algorithm which is obtained by combining IS divergence as the error function and with ℓ_1-norm as the sparsity constraint. Using IS divergence as error function is quite distinguished comparing with the conventional algorithms. Results of experiments on dictionary recovery show that the ISSC algorithm can correctly learn an overcomplete, nonnegative dictionary on synthetic signals and further show that the proposed algorithm performs the robustness against different level of noise in comparison with the other compared algorithms which all using the EUC distance as the error function. It implies that different error functions lead to different results. In future work, we will employ the proposed algorithm for inpainting, image denoising and other application.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] A. Cichocki and A.-H. Phan, "Fast local algorithms for large scale nonnegative matrix and tensor factorizations," *IEICE Trans. on Fundamentals of Electronics*, vol. E92-A, no. 3, pp. 708–721, 2009.

[3] V. P. Pauca, J. Piper, and R. J. Plemmons, "Nonnegative matrix factorization for spectral data analysis," *Linear Algebra and its Applications*, vol. 416, no. 1, pp. 29–47, 2006.

[4] M. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. Davies, "Sparse representations in audio and music: From coding to source separation," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 995–1005, Jun 2010.

[5] M. Elad, M. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 972–982, 2010.

[1] Online available *http://www.cs.technion.ac.il/~elad/software/*

[2] Online available *http://www.cs.helsinki.fi/u/phoyer/contact.html*

[3] Online available *http://www3.spsc.tugraz.at/people/robert-peharz*

[6] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 3, pp. 765–777, 2007.

[7] S. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, vol. 1, 2001, pp. 207–212.

[8] I. Kotsia, S. Zafeiriou, and I. Pitas, "A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 588–595, Sep 2007.

[9] R. Peharz and F. Pernkopf, "Sparse nonnegative matrix factorization using ℓ^0-constraints," *Neurocomputing*, vol. 80, pp. 38–46, Mar 2012.

[10] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[11] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons, "Text mining using non-negative matrix factorizations," in *Proc. of the Fourth SIAM International Conference on Data Mining*, 2004, pp. 452–456.

[12] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Information Processing & Management*, vol. 42, no. 2, pp. 373–386, 2006.

[13] R. Peharz, M. Stark, and F. Pernkopf, "Sparse nonnegative matrix factorization using ℓ^0-constraints," pp. 83–88, Sep. 2010.

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

[15] W. Liu, N. Zheng, and X. Lu, "Non-negative matrix factorization for visual coding," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, April 2003, pp. 293–296.

[16] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," *In Proc 6th International Congress on Acoustics*, pp. C–17 – C–20, 1968.

[17] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD and its nonnegative variant for dictionary design," vol. 5914, pp. 327–339, Jul. 2005.

[18] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," *Technical report, Institute of Statistical Mathematics*, 2001.

[19] A. Cichocki, R. Zdunek, and S. Amari, "Csisz*á*rs divergences for non-negative matrix factorization : Family of new algorithms," *Proc. Int. Conf. Independent Component Analysis and Blind Signal Separation*, p. 32C39, 2006.

[20] C. F*é*votte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[21] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," in *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*, 2010, pp. 283–288.

[22] 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.

[23] M. Fadili, J.-L. Starck, and F. Murtagh, "Inpainting and zooming using sparse representations," *The Computer Journal*, vol. 52, no. 1, pp. 64–79, 2009.

[24] J. Zepeda, C. Guillemot, and E. Kijak, "Image compression using sparse representations and the iteration-tuned and aligned dictionary," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 1061 –1073, Sept. 2011.

[25] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, June 2008, pp. 1–8.