Robust Non-Negative Dictionary Learning

Qihe Pan[1], Deguang Kong[2], Chris Ding[2] and Bin Luo[3]
[1]Beihang University, China; [2]University of Texas, Arlington, U.S.A; [3]Anhui University, China
panqihe2006@gmail.com; doogkong@gmail.com; chqding@uta.edu; luobin@ahu.edu.cn

Abstract

Dictionary learning plays an important role in machine learning, where data vectors are modeled as a sparse linear combinations of basis factors (i.e., dictionary). However, how to conduct dictionary learning in noisy environment has not been well studied. Moreover, in practice, the dictionary (i.e., the lower rank approximation of the data matrix) and the sparse representations are required to be nonnegative, such as applications for image annotation, document summarization, microarray analysis. In this paper, we propose a new formulation for non-negative dictionary learning in noisy environment, where structure sparsity is enforced on sparse representation. The proposed new formulation is also robust for data with noises and outliers, due to a robust loss function used. We derive an efficient multiplicative updating algorithm to solve the optimization problem, where dictionary and sparse representation are updated iteratively. We prove the convergence and correctness of proposed algorithm rigorously. We show the differences of dictionary at different level of sparsity constraint. The proposed algorithm can be adapted for clustering and semi-supervised learning.

Introduction

In dictionary learning, a signal is represented as a sparse representation of basis factors (called dictionary), instead of predefined wavelets (Mallat 1999). Dictionary learning has shown the state of the art performance, and has many applications for image denoising (Elad and Aharon 2006)), face recognition (Protter and Elad 2009), document clustering, microarray analysis, etc. Recent researches (Raina et al. 2007; Delgado et al. 2003; Mairal et al. 2009; Olshausen and Fieldt 1997) have shown the sparsity helps to eliminate data redundancy, and capture the correlations inherent in data. Compared with Principal Component Analysis (PCA), dictionary learning does not have a strict constraint (such as orthogonal) on the basis vector, and thus the dictionary can be learned in a more flexible way. The key to dictionary learning, at different context with different constraints, is to solve the corresponding optimization problem. For example, different objective functions (Aharon, Elad, and Bruckstein 2006; Mairal et al. 2010) have been proposed to meet the requirement of specific applications, e.g., supervised dictionary learning (Mairal et al. 2008), a joint learning using dictionary learning and clustering-based sparse representation (Dong et al. 2011), online dictionary learning (Kasiviswanathan et al. 2011), tensor decomposition for image storage (Zhang and Ding 2013), etc.

In this paper, we focus on a general non-negative dictionary learning problem in noisy environment, i.e., data could be noisy and have missing values. To summarize, the main contribution of this paper is in three-fold. **(1)** We formulate the non-negative dictionary learning problem in noisy environment through the optimization of a non-smooth loss function over non-negative set with LASSO-type regularization term. **(2)** It is challenging to solve this problem due to the non-smoothness of reconstruction error term and sparsity regularization term. Different from the recent second order iterative algorithms (e.g., (Lee et al. 2007; Aharon, Elad, and Bruckstein 2006)) used for dictionary learning, we propose an efficient multiplicative updating algorithm, where the convergence and correctness of algorithm are rigorously proved. **(3)** As shown in experiment, our algorithm converges very fast. The learned sparse coding $\mathbf{Y}$ can be used for clustering and semi-supervised learning.

Robust Dictionary Learning Objective

In standard dictionary learning, given a set of training signals $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, where $\mathbf{x}_i \in \Re^p$ represents a data of p-dimension. We use $\mathbf{A} \in \Re^{p\times k}$ to represent fixed size dictionary, where $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k], \mathbf{a}_i \in \Re^p$. For each signal $\mathbf{x} \in \Re^p$, we need to optimize the loss function $L(\mathbf{x}, \mathbf{A})$ such that the loss is small using dictionary representation. Note this dictionary representation usually needs to be sparse. Usually, we need to optimize LASSO type objective (Tibshirani 1994), i.e.,

$$\min_{\mathbf{y}} ||\mathbf{x} - \mathbf{A}\mathbf{y}||^2 + \alpha||\mathbf{y}||_1, \tag{1}$$

where $\mathbf{y} \in \Re^k$ is sparse representation of signal $\mathbf{x}$ using dictionary $\mathbf{A}$, $\alpha > 0$ is a parameter. Note standard least square loss is used in Eq.(1), which implies Gaussian noises existed in input data signals. However, in real world, data measurement could be noisy and have missing values. It is known that least square loss is prone to noises and large deviations. Replacing the least square loss of Eq.(1) with more robust ℓ_1 loss, robust dictionary learning becomes,

$$\min_{\mathbf{y}_i, \mathbf{A}\in\mathcal{C}} \sum_i ||\mathbf{x}_i - \mathbf{A}\mathbf{y}_i||_1 + \alpha||\mathbf{y}_i||_1, \tag{2}$$

where dictionary $\mathbf{A} \in \mathcal{C}$, where $\mathcal{C}$ is the feasible domain of problem, i.e., $\mathcal{C} = \{\mathbf{A}|\mathbf{A} \geq 0\}$, or $\mathcal{C} = \{\mathbf{A}|\|\mathbf{a}_j\|_2 \leq 1\}$ ($\mathbf{a}_j$ is j-th column of $\mathbf{A}$).

In real world problems (such as images features, text vector, etc), input data are non-negative values, which requires the dictionary to be non-negative, i.e., $\mathbf{A} \geq 0$. Naturally, the sparse representation $\mathbf{y}_i$ for each signal $\mathbf{x}_i$ should be non-negative.

Problem Formulation

Thus, in this paper, we focus on feasible domain of $\mathbf{A}$ to be: $\mathcal{C} = \{\mathbf{A}|\mathbf{A} \geq 0\}$. Then objective of Eq.(2) becomes,

$$\min_{\mathbf{y}_i, \mathbf{A}} \sum_i ||\mathbf{x}_i - \mathbf{A}\mathbf{y}_i||_1 + \alpha||\mathbf{y}_i||_1 + \beta\|\mathbf{A}\|_F^2, \quad s.t. \quad \mathbf{A} \geq 0, \mathbf{y}_i \geq 0. \tag{3}$$

Note smooth term $\|\mathbf{A}\|_F^2$ is added in Eq.(3) to avoid the trivial solution. In practice, we require $\beta > 0$. If $\beta = 0$, suppose $(\mathbf{A}^*, \mathbf{y}_i^*)$ is the current optimal solution for Eq.(3), then we can always get a better solution $(\mathbf{A}^*, \mathbf{y}_i^*)$ with smaller objective function value of Eq.(3), where $\mathbf{A}^{**} = \theta\mathbf{A}^*, \mathbf{y}_i^{**} = \frac{1}{\theta}\mathbf{y}_i^*$ and $\theta > 1$.

Using matrix formulation, let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n]$, then Eq.(3) becomes,

$$\min_{\mathbf{Y}, \mathbf{A}} ||\mathbf{X} - \mathbf{A}\mathbf{Y}||_1 + \alpha||\mathbf{Y}||_1 + \beta\|\mathbf{A}\|_F^2, \quad s.t. \quad \mathbf{A} \geq 0, \mathbf{Y} \geq 0, \tag{4}$$

where $||\mathbf{Y}||_1 = \sum_{ki} |Y_{ki}|$. By introducing Lagrangian multiplier to enforce the constraint, Eq.(4) can be equivalently expressed as,

$$\min_{\mathbf{Y}, \mathbf{A}} ||\mathbf{X} - \mathbf{A}\mathbf{Y}||_1, \quad s.t. \quad \mathbf{A} \geq 0, \mathbf{Y} \geq 0, ||\mathbf{Y}||_1 \leq q, \|\mathbf{A}\|_F^2 \leq p. \tag{5}$$

The optimization of Eq.(4) is general non-convex. But if one of the variables ($\mathbf{A}$ or $\mathbf{Y}$) is known, we can find the global optimal solution w.r.t the other variable. Note two non-smooth ℓ_1 terms are involved in Eq.(4), and thus it is a bit challenging to solve Eq.(4). However, it does not add any difficulty, because in Eq.(4), ℓ_1 term appeared together with non-negative constraint. Thus ℓ_1 term w.r.t sparse coding can be rewritten as, $||\mathbf{Y}||_1 = \text{Tr}(\mathbf{E}\mathbf{Y})$, where $\mathbf{E} \in \Re^{k \times n}$ is a matrix with all ones.

Algorithm

A main contribution of this paper is to derive the following multiplicative updating algorithms for problem of Eq.(4), i.e.,

$$A_{jk} \Leftarrow A_{jk} \frac{[\mathbf{X} \odot \mathbf{W}\mathbf{Y}^T]_{jk}}{[(\mathbf{A}\mathbf{Y}) \odot \mathbf{W}\mathbf{Y}^T + 2\beta\mathbf{A}]_{jk}}, \tag{6}$$

$$Y_{ki} \Leftarrow Y_{ki} \frac{[\mathbf{A}^T\mathbf{X} \odot \mathbf{W}]_{ki}}{[\mathbf{A}^T(\mathbf{A}\mathbf{Y}) \odot \mathbf{W} + \alpha\mathbf{E}]_{ki}} \tag{7}$$

where $\mathbf{E}$ is a all-ones matrix, $\mathbf{W}$ is a matrix given by $W_{ij} = \left((\mathbf{X}-\mathbf{A}\mathbf{Y})_{ij}^2 + \epsilon^2\right)^{-1/2}$ and $\odot$ is the Hadamard product, i.e., elementwise product between two matrices. Here we assume Hadamard product has higher operator precedence over regular matrix product, i.e., $\mathbf{A}\mathbf{B} \odot \mathcal{C}\mathbf{D} = \mathbf{A}(\mathbf{B} \odot \mathcal{C})\mathbf{D}$. Note that as $\epsilon \to 0$, $\left((\mathbf{X} - \mathbf{A}\mathbf{Y})_{ij}^2 + \epsilon^2\right)^{1/2} \to |(\mathbf{X} - \mathbf{A}\mathbf{Y})_{ij}|$. We add a small number ϵ here to prevent overflow of $\mathbf{W}_{ij}$ in the case $(\mathbf{X} - \mathbf{A}\mathbf{Y})_{ji}^2 \simeq 0$. [1] Because ϵ is not zero, the algorithm updating rules of Eqs.(6,7) actually minimize the objective function

$$\min_{\mathbf{A} \geq 0, \mathbf{Y} \geq 0} \sum_{i=1}^{n}\sum_{j=1}^{p}((\mathbf{X}-\mathbf{A}\mathbf{Y})_{ji}^2+\epsilon^2)^{1/2}+\alpha\sum_{i=1}^{k}\sum_{j=1}^{p}|\mathbf{Y}_{ij}|+\beta\|\mathbf{A}\|_F^2. \tag{8}$$

Figure 1: Computed $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k)$ on YaleB dataset (K = 31) shown as 3 rows using Eq.(4) at $\alpha = 0.5$.

Figure 2: Computed $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k)$ on YaleB dataset (K = 31) shown as 3 rows using Eq.(4) at $\alpha = 1$.

Illustration of dictionary at different α

To simply the problem, we fix $\beta = 0.1$. On YaleB face data set, each image $\mathbf{x}_i \in \Re^p$ is linearized into a vector, thus $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \Re^{p \times n}$ is used to compute the dictionary $\mathbf{A} \in \Re^{p \times k}$ and sparse coding $\mathbf{Y} \in \Re^{k \times n}$. Each dictionary $\mathbf{a}_i \in \Re^p$ in the computed $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k]$ is corresponding to each category ($K = 31$), and thus shown as an image. Dictionary results $\mathbf{A}$, at $\alpha = 0.5$, $\alpha = 1$ of Eq.(4), are shown in Fig.1 and Fig.2, respectively. Clearly, the dictionary changes slightly at different sparsity constraint (say, different α values). Generally, $\alpha = 1$ gives slightly better visual results as compared to $\alpha = 0.5$, due to larger sparsity enforcement.

Connections to Related Works

Connection to Sparse Coding Our model has also some connections to sparse coding (Olshausen and Fieldt 1997), lasso (Tibshirani 1994) and elastic net (Zou and Hastie 2005). The basic idea of sparse coding is to represent a feature vector as linear combination of few bases from a predefined dictionary, hence inducing a concept of sparsity. Given dictionary $\mathbf{A}$, our model is to findsparse representation $\mathbf{y}$ for each signal $\mathbf{x}$, i.e.,

$$\min_{\mathbf{y}} ||\mathbf{x} - \mathbf{A}\mathbf{y}||_1 + \alpha||\mathbf{y}||_1. \tag{9}$$

If we replace the ℓ_1 norm on the error term (1st term) by ℓ_2 norm, this is exactly the LASSO. If we add the smooth

[1] ϵ is set to machine precision in our experiments.

term of $||\mathbf{y}||_2^2$ to Eq.(9), this is identical to the elastic net, which improves the smoothness of the process. Using matrix format, Eq.(4) becomes,

$$\min_{\mathbf{Y},\mathbf{A}} ||\mathbf{X} - \mathbf{AY}||_1 + \alpha||\mathbf{Y}||_1 + \gamma||\mathbf{Y}||_F^2, \quad s.t. \quad \mathbf{A} \geq 0, \mathbf{Y} \geq 0, \tag{10}$$

because $||\mathbf{Y}||_F^2 = \sum_i ||\mathbf{y}_i||_2^2$. Note the multiplicative rule of Eq.(6) for dictionary $\mathbf{A}$ will not change, we only need to change multiplicative rule for $\mathbf{Y}$ of Eq.(7) to,

$$Y_{ki} \Leftarrow Y_{ki} \frac{[\mathbf{A}^T\mathbf{X} \odot \mathbf{W}]_{ki}}{[\mathbf{A}^T(\mathbf{AY}) \odot \mathbf{W} + \alpha\mathbf{E} + 2\gamma\mathbf{Y}]_{ki}} \tag{11}$$

If we use original data $\mathbf{X}$ as dictionary, Eq.(4) becomes,

$$\min_{\mathbf{S}} ||\mathbf{X} - \mathbf{XS}||_1 + \alpha||\mathbf{S}||_1, \quad s.t. \quad \mathbf{A} \geq 0, \mathbf{S} \geq 0, \tag{12}$$

where $\mathbf{S} \in \Re^{n\times n}$ acts as pairwise similarity between data points, and can be updated with the following rule,

$$\mathbf{S}_{ij} \Leftarrow \mathbf{S}_{ij} \frac{[\mathbf{X}^T(\mathbf{X} \odot \hat{\mathbf{W}})]_{ij}}{[\mathbf{X}^T(\mathbf{XS}) \odot \hat{\mathbf{W}} + \alpha\mathbf{E}]_{ij}}, \tag{13}$$

where $\hat{\mathbf{W}}_{ij} = [(\mathbf{X} - \mathbf{XS})_{ij}]^{-\frac{1}{2}}$, and $\mathbf{E} \in \Re^{n\times n}$ is a matrix with all ones. The correctness of Eq.(11) and Eq.(13) can be similarly proved as that of Eq.(4), which has been sharply observed in (Kong and Ding 2012a) .

Connection to Non-negative Matrix Factorization It has been shown non-negative matrix factorization (Lee and Seung 2000) has close relations with dictionary learning. In our model of Eq.(4), if we set $\lambda = 0$, this is exactly robust non-negative matrix factorization using ℓ_1 norm, where dictionary $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k]$ plays the role of basis vector in NMF, and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n]$ is the cluster indicator, k is the dimension of subspace. Sparse term $||\mathbf{Y}||_1$ enforces the cluster indicator of NMF solution to be sparse. It also has clear differences with NMF model using $\ell_{2,1}$ error function (Kong, Ding, and Huang 2011), (Ding and Kong 2012) (using our notation), i.e.,

$$\|\mathbf{X} - \mathbf{AY}\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{p} (\mathbf{X} - \mathbf{AY})_{ji}^2} = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{y}_i\|, \tag{14}$$

where index i (number of data), j (dimension of features) are differently treated. Back to our model, Eq.(4) can be rewritten as,

$$\min_{\mathbf{Y},\mathbf{A}} ||\mathbf{X} - \mathbf{AY}||_{2,1} + \alpha||\mathbf{Y}||_1 + \beta\|\mathbf{A}\|_F^2, \quad s.t. \quad \mathbf{A} \geq 0, \mathbf{Y} \geq 0, \tag{15}$$

where dictionary $\mathbf{A}$ is learned with a robust $\ell_{2,1}$ function.

Connection to k-means Clustering The objective function of the K-means clustering (MacQueen 1967) is $J_{K2} = \sum_{k=1}^{K} \sum_{i\in C_k} \|\mathbf{x}_i - \mathbf{f}_k\|^2$, where $\mathbf{f}_k$ is the centroid of the k-th cluster C_k. If we use a more robust error function of L_1 norm, we have $J_{K1} = \sum_{k=1}^{K} \sum_{i\in C_k} \|\mathbf{x}_i - \mathbf{f}_k\|_1$. In our formulation of Eq.(4), set $\alpha = 0, \beta = 0$, let sparse representation $\mathbf{Y}$ be the solution of the clustering: $\mathbf{Y}_{ki} = 1$ if $\mathbf{x}_i$ belongs to cluster C_k; otherwise, $\mathbf{Y}_{ki} = 0$ Thus we have

$$J_{K1} = \sum_{k=1}^{K} \sum_{i\in C_k} \|\mathbf{x}_i - \sum_{k=1}^{K} \mathbf{a}_k \mathbf{Y}_{ki}\|_1$$

$$= \sum_{k=1}^{K} \sum_{i\in C_k} \|\mathbf{x}_i - \mathbf{A}\mathbf{y}_i\|_1 = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{y}_i\|_1 = \|\mathbf{X} - \mathbf{AY}\|_1.$$

Thus our model of Eq.(4) implicitly performs a ℓ_1 K-means clustering. If $\alpha \neq 0, \beta \neq 0$, it performs a constraint k-means clustering by imposing constraint $||\mathbf{Y}||_1 < q$ on cluster indicators.

Convergence of the Algorithm

We give the convergence of algorithm in **Theorem** 1.

Theorem 1. *(A) Updating $\mathbf{Y}$ using the rule of Eq.(7) while fixing $\mathbf{A}$, the objective function of Eq.(4) monotonically decreases. (B) Updating $\mathbf{A}$ using the rule of Eq.(6) while fixing $\mathbf{Y}$, the objective function of Eq.(4) monotonically decreases.*

To prove **Theorem** 1, we need the following definition: $W_{ij} = \left((\mathbf{X} - \mathbf{AY})_{ij}^2 + \epsilon^2\right)^{-1/2}$, $\|\mathbf{B}\|_{\mathbf{W}}^2 = \sum_{ij} B_{ij}^2 W_{ij}$, which are used in the following Lemmas. Note $\mathbf{W}$ is a constant given current $\mathbf{A}, \mathbf{Y}$.

Updating $\mathbf{Y}$ We focus on updating $\mathbf{Y}$ while fixing $\mathbf{A}$. Let LHS be left-hand-side of an equation, and RHS be right-hand-side of an equation. The proof of Theorem 1(A) requires the following two lemmas. Note in the updating process of $\mathbf{Y}$ from $\mathbf{Y}^t$ to $\mathbf{Y}^{t+1}$, $\mathbf{A}, \mathbf{W}$ remain the same. Thus,

$$\|\mathbf{X} - \mathbf{AY}^{t+1}\|_{\mathbf{W}}^2 = \sum_{ji} (\mathbf{X} - \mathbf{AY}^{t+1})_{ji}^2 \mathbf{W}_{ji},$$

$$\|\mathbf{X} - \mathbf{AY}^{t}\|_{\mathbf{W}}^2 = \sum_{ji} (\mathbf{X} - \mathbf{AY}^{t})_{ji}^2 \mathbf{W}_{ji},$$

where $\mathbf{W}_{ji} = [(\mathbf{X} - \mathbf{AY}^t)_{ji}^2 + \epsilon^2]^{-1/2}$.

Lemma 2. *Let $\mathbf{Y}^t$ be the old $\mathbf{Y}$ [on the RHS of Eq.(7)] and $\mathbf{Y}^{t+1}$ be the new $\mathbf{Y}$ [on the LHS of Eq.(7)]. Under the updating rule of Eq.(7), the following holds*

$$\frac{1}{2}\|\mathbf{X} - \mathbf{AY}^{t+1}\|_{\mathbf{W}}^2 + \alpha Tr(\mathbf{EY}^{t+1})$$
$$\leq \frac{1}{2}\|\mathbf{X} - \mathbf{AY}^{t}\|_{\mathbf{W}}^2 + \alpha Tr(\mathbf{EY}^{t}). \tag{16}$$

Lemma 3. *Under the updating rule of Eq.(7), the following holds*

$$\|\mathbf{X} - \mathbf{AY}^{t+1}\|_1 - \|\mathbf{X} - \mathbf{AY}^{t}\|_1 \leq \tag{17}$$
$$\frac{1}{2}\Big(\|\mathbf{X} - \mathbf{AY}^{t+1}\|_{\mathbf{W}}^2 - \|\mathbf{X} - \mathbf{AY}^{t}\|_{\mathbf{W}}^2\Big),$$

The key idea of proof of Lemma 2 is to construct an auxiliary function to show the convergence of the objective function. The key idea of proof of Lemma 3 is to compute the difference between LHS and RHS of Eq.(17).

Proof of Theorem 1. From Lemma 3, the RHS of Eq.(17) is negative or zero. Therefore

$$[\|\mathbf{X} - \mathbf{AY}^{t+1}\|_1 + \alpha \mathrm{Tr}(\mathbf{EY}^{t+1})]$$
$$-[\|\mathbf{X} - \mathbf{AY}^{t}\|_1 + \alpha \mathrm{Tr}(\mathbf{EY}^{t})]$$
$$\leq \frac{1}{2}\|\mathbf{X} - \mathbf{AY}^{t+1}\|_{\mathbf{W}}^2 - \frac{1}{2}\|\mathbf{X} - \mathbf{AY}^{t}\|_{\mathbf{W}}^2$$
$$+\alpha \mathrm{Tr}(\mathbf{EY}^{t+1}) - \alpha \mathrm{Tr}(\mathbf{EY}^{t}) \leq 0. \tag{18}$$

This proves that the objective decreases monotonically. □

Updating $\mathbf{A}$ We now focus on updating $\mathbf{A}$ while fixing $\mathbf{Y}$. The proof of Theorem 1(B) requires the following two lemmas:

Lemma 4. *Let $\mathbf{A}^t$ be the old $\mathbf{A}$ [on the RHS of Eq.(6)] and $\mathbf{A}^{t+1}$ be the new $\mathbf{A}$ [on the LHS of Eq.(6)]. Under the updating rule of Eq.(6), the following holds*

$$\frac{1}{2}\|\mathbf{X} - \mathbf{A}^{t+1}\mathbf{Y}\|_{\mathbf{W}}^2 + \beta\|\mathbf{A}^{t+1}\|_F^2 \leq \frac{1}{2}\|\mathbf{X} - \mathbf{A}^{t}\mathbf{Y}\|_{\mathbf{W}}^2 + \beta\|\mathbf{A}^{t}\|_F^2,$$

Lemma 5. *Under the updating rule of Eq.(7), the following holds*

$$\|\mathbf{X}-\mathbf{A}^{t+1}\mathbf{Y}\|_1-\|\mathbf{X}-\mathbf{A}^{t}\mathbf{Y}\|_1\leq \frac{1}{2}\Big[\|\mathbf{X}-\mathbf{A}^{t+1}\mathbf{Y}\|^2_{\mathbf{W}}-\|\mathbf{X}-\mathbf{A}^{t}\mathbf{Y}\|^2_{\mathbf{W}}\Big], \quad (19)$$

The proofs of Lemmas 4, 5 are similar to the proofs of Lemmas 2, 3 and thus are skipped due to space limitation.

Poof of Theorem 1(B). From Lemma 5, the RHS value of Eq.(19) is negative or zero. Therefore

$$\begin{aligned}&[\|\mathbf{X}-\mathbf{A}^{t+1}\mathbf{Y}\|_1+\alpha||\mathbf{Y}||_1+\beta\|\mathbf{A}^{t+1}\|_F^2]\\&-[\|\mathbf{X}-\mathbf{A}^{t}\mathbf{Y}\|_1+\alpha||\mathbf{Y}||_1+\beta\|\mathbf{A}^{t}\|_F^2]\\&\leq[\frac{1}{2}\|\mathbf{X}-\mathbf{A}^{t+1}\mathbf{Y}\|^2_{\mathbf{W}}+\beta\|\mathbf{A}^{t+1}\|_F^2]\\&-[\frac{1}{2}\|\mathbf{X}-\mathbf{A}^{t}\mathbf{Y}\|^2_{\mathbf{W}}+\beta\|\mathbf{A}^{t}\|_F^2]\leq 0. \quad (20)\end{aligned}$$

This proves that the objective decreases monotonically. □

Remark Proposed multiplicative update algorithm converges to a local optimum due to the non-convexity of $f(\mathbf{A},\mathbf{Y})$ w.r.t both $\mathbf{A}$ and $\mathbf{Y}$. However, even local minima still provides very desirable properties for dictionary learning tasks. It is usually very difficult to choose step size to guarantee the convergence of general gradient descent method. The proposed multiplicative method provides a smart choice for step size, and thus for a better dictionary.

Due to space limit, the proof of Lemma 2 is omitted here.

Proof of Lemma 3 The left-hand-side (LHS) of Eq.(17) is

$$\sum_{j=1}^{p}\sum_{i=1}^{n}\Big[\sqrt{(\mathbf{X}-\mathbf{A}\mathbf{Y}^{t+1})^2_{ji}+\epsilon^2}-\sqrt{(\mathbf{X}-\mathbf{A}\mathbf{Y}^{t})^2_{ji}+\epsilon^2}\Big]$$
$$=\sum_{j=1}^{p}\sum_{i=1}^{n}[\sqrt{(\mathbf{X}-\mathbf{A}\mathbf{Y}^{t+1})^2_{ji}+\epsilon^2}-1/W_{ji}]$$

using the definition of $W_{ji}=[(\mathbf{X}-\mathbf{A}\mathbf{Y}^t)^2_{ji}+\epsilon^2]^{-1/2}$. The right-hand-side (RHS) of Eq.(17) is

$$RHS=\frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{n}\Big[(\mathbf{X}-\mathbf{A}\mathbf{Y}^{t+1})^2_{ji}W_{ji}-(\mathbf{X}-\mathbf{A}\mathbf{Y}^{t})^2_{ji}W_{ji}\Big]$$
$$=\frac{1}{2}\sum_{ji}\Big[[(\mathbf{X}-\mathbf{A}\mathbf{Y}^{t+1})^2_{ji}+\epsilon^2]W_{ji}-[(\mathbf{X}-\mathbf{A}\mathbf{Y}^{t})^2_{ji}+\epsilon^2]W_{ji}\Big]$$
$$=\frac{1}{2}\sum_{ji}\Big[[(\mathbf{X}-\mathbf{A}\mathbf{Y}^{t+1})^2_{ji}+\epsilon^2]W_{ji}-1/W_{ji}\Big]$$

Therefore,

$$LHS-RHS=-\sum_{ji}\frac{W_{ji}}{2}\Big[-2\frac{1}{W_{ji}}\sqrt{(\mathbf{X}-\mathbf{A}\mathbf{Y}^{t+1})^2_{ji}+\epsilon^2}$$
$$+\frac{1}{W^2_{ji}}+[(\mathbf{X}-\mathbf{A}\mathbf{Y}^{t+1})^2_{ji}+\epsilon^2]\Big]$$
$$=-\sum_{ji}\frac{W_{ji}}{2}\Big(\sqrt{(\mathbf{X}-\mathbf{A}\mathbf{Y}^{t+1})^2_{ji}+\epsilon^2}-\frac{1}{W_{ji}}\Big)^2\leq 0$$

□

Correctness of the Algorithm

We prove that the converged solution satisfies the Karush-Kuhn-Tucker condition of the constrained optimization theory. We prove the correctness of the algorithm w.r.t. $\mathbf{A}$ and $\mathbf{Y}$, respectively.

Figure 3: umist data, half of the images from each category are occluded. Occlusion size: 7 x 7.

Figure 4: Caltech data. Shown images are from "face" category.

Table 1: Descriptions of occluded datasets

Dataset	#Size	#Dimension	#Class	occluded size
AT&T	400	644	40	10 x 10
Mnist	150	784	10	8 x 8
Umist	360	644	20	7 x 7
YaleB	1984	504	31	N/A
Caltech	600	432	20	N/A

Theorem 6. *The converged solution $\mathbf{Y}^*$ of the updating rule of Eq.(7) satisfies the KKT condition of optimization theory.*

Theorem 7. *The converged solution $\mathbf{A}^*$ of the updating rule of Eq.(6) satisfies the KKT condition of optimization theory.*

Proof of Theorem 6. Let $J(\mathbf{Y})=\|\mathbf{X}-\mathbf{A}\mathbf{Y}\|_1+\alpha\|\mathbf{Y}\|_1+\beta\|\mathbf{A}\|_F^2$ of Eq.(10). The KKT condition for $\mathbf{Y}$ with the constraints $Y_{ki}\geq 0,\ i=1\cdots n,\ k=1\cdots K$ is $\frac{\partial J(\mathbf{Y})}{\partial Y_{ki}}Y_{ki}=0,\ \forall\, i,k$. The derivative is

$$\frac{\partial J(\mathbf{Y})}{\partial Y_{ki}}=\sum_{i'=1}^{n}\sum_{j=1}^{p}\frac{(\mathbf{X}-\mathbf{A}\mathbf{Y})_{ji'}}{\sqrt{(\mathbf{X}-\mathbf{A}\mathbf{Y})^2_{ji'}+\epsilon^2}}\frac{\partial(\mathbf{X}-\mathbf{A}\mathbf{Y})_{ji'}}{\partial Y_{ki}}$$
$$+\frac{\lambda\mathrm{Tr}(\mathbf{E}\mathbf{Y})}{Y_{ki}}=\sum_{j=1}^{p}-W_{ji}(\mathbf{X}-\mathbf{A}\mathbf{Y})_{ji}\mathbf{A}_{jk}+\alpha E_{ki} \quad (21)$$
$$=-(\mathbf{A}^T\mathbf{X}\odot\mathbf{W})_{ki}+[\mathbf{A}^T(\mathbf{A}\mathbf{Y})\odot\mathbf{W}+\alpha\mathbf{E}]_{ki}.$$

Thus the KKT condition for $\mathbf{Y}$ is

$$[-(\mathbf{A}^T\mathbf{X}\odot\mathbf{W})_{ki}+(\mathbf{A}^T(\mathbf{A}\mathbf{Y})\odot\mathbf{W}+\alpha\mathbf{E})_{ki}]Y_{ki}=0, \quad (22)$$

On the other hand, once $\mathbf{Y}$ converges, according to the updating rule of Eq.(7), the converged solution $\mathbf{Y}^*$ satisfies

$$Y^*_{ki}=Y^*_{ki}\frac{(\mathbf{A}^T\mathbf{X}\odot\mathbf{W})_{ki}}{(\mathbf{A}^T(\mathbf{A}\mathbf{Y})\odot\mathbf{W}+\alpha\mathbf{E})_{ki}}, \quad (23)$$

which can be written as $[-(\mathbf{A}^T\mathbf{X}\odot\mathbf{W})_{ki}+(\mathbf{A}^T(\mathbf{A}\mathbf{Y})\odot\mathbf{W}+\alpha\mathbf{E})_{ki}]Y^*_{ki}=0$. This is identical to Eq.(22). Thus the converged solution satisfies the KKT condition. □

The proof of Theorem 7 is similar to that of Theorem 6, and thus is skipped due to space limit.

Experiment

In this section, we empirically evaluate the proposed approach, where our goal is to examine the convergence of the

proposed algorithm, and also compare against other robust dictionary learning methods in noisy environment.

We do experiment on 5 data sets in our experiments, including two face datasets AT&T [1] and Umist, YaleB, one digit datasets mnist (Lecun et al. 1998) and one image scene datasets Caltech101 (Dueck and Frey 2007). Table 1 summarizes the characteristics of the datasets.

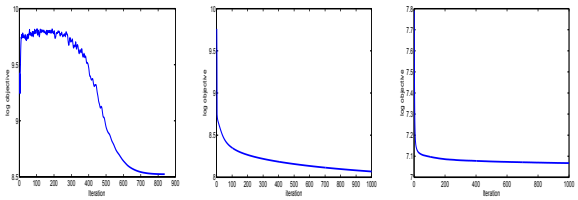
We generate occluded image datasets corresponding to above 3 original data sets (except YaleB and Caltech). For YaleB dataset, the images are taken under different poses with different illumination conditions. The shading parts of the images play the similar role of occlusion (noises). Thus we use the original YaleB data. For Caltech dataset, the natural scenes images are polluted by noises when pictures are taken. For the other 4 datasets, half of the images are selected from each category for occlusion with block size of wxw pixels (e.g., w = 10). The locations of occlusions are random generated without overlaps among the images from the same category. A demonstration of images are shown in Figs.3,4.

Convergence of the algorithm We show the convergence of our algorithm for Eq.(4) in first 1000 iterations on dataset AT&T and Umist in Fig.5(b) and Fig.5(c), respectively. "x-axis" is the number of iteration, "y-axis" is the value of log function of Eq.(4) at $\lambda = 2$. We use results $\mathbf{G} \in \{0,1\}^{k\times n}$ computed from standard k-means clustering, to initialize $\mathbf{Y} = \mathbf{G} + 0.3$, and then dictionary $\mathbf{A} \in \Re^{p\times k}$ is computed from the centroid of each category. Experiment results indicate our algorithm of Eqs.(6, 7) converges very fast. We note Alternating direction method (ADM) (Bertsekas 1996) can be used to solve Eq.(4). We show the convergence of ADM on dataset AT&T at $\lambda = 2$ in Fig.5(a), where objective function of Eq.(4) decreases from $1.2426e+4$ to $5.032e+3$ in 838 iterations. As compared to ADM, our algorithm decreases very fast at first, and guarantees monotonically decreaseing in each step.

Data Clustering experiment As is shown before, the obtained sparse coding $\mathbf{Y}$ can be used as "cluster indicator" to do clustering tasks, where each data $\mathbf{x}_i$ is attributed to category k, such that $k = \arg\max_{k'} \mathbf{Y}_{k'i}$. The evaluation metrics (Bühler and Hein 2009) we used here are clustering accuracy, normalized mutual information and purity. These measurement are widely used in the evaluation of different clustering approaches. The larger values of these metrics indicate the better performance of clustering methods.

Compared Methods We compare the proposed method with the following related methods: (1) k-means clustering (k-means); (2) standard non-negative matrix factorization using least square error function (L2NMF); (3) non-negative matrix factorization with ℓ_1 sparse constraint on cluster indicator (L2NMFs) (Kim et al. 2011), which optimizes: $\min_{\mathbf{A}\geq 0,\mathbf{Y}\geq 0} \|\mathbf{X} - \mathbf{AY}\|_F^2 + \alpha\|\mathbf{Y}\|_1 + \beta\|\mathbf{A}\|_F^2$; (4) sparse non-negative matrix factorization with group sparse constraint (L2NMFgs) on cluster indicator (Kim, Monteiro, and Park 2012), which optimizes: $\min_{\mathbf{A}\geq 0,\mathbf{Y}\geq 0} \|\mathbf{X} - \mathbf{AY}\|_F^2 + \alpha\sum_{j=1}^{n}(\sum_{i=1}^{k}|\mathbf{Y}_{ij}|)^2 + \beta\|\mathbf{A}\|_F^2$; (5) robust non-negative matrix factorization using $\ell_{2,1}$ error function (L21NMF) (Ding et al. 2006), which optimizes: $\min_{\mathbf{A}\geq 0,\mathbf{Y}\geq 0} \|\mathbf{X} - \mathbf{AY}\|_{2,1}$; (6) robust non-negative matrix factorization using $\ell_{2,1}$ error function and ℓ_1 sparse constraint on cluster indicator (L21NMFs) (Kong, Ding, and Huang 2011), which optimizes: $\min_{\mathbf{A}\geq 0,\mathbf{Y}\geq 0} \|\mathbf{X} - \mathbf{AY}\|_{2,1} + \alpha\|\mathbf{Y}\|_1 + \beta\|\mathbf{A}\|_F^2$; (7) robust non-negative matrix factorization using ℓ_1 error function (L1NMF) (Ke and Kanade 2005), which optimizes: $\min_{\mathbf{A}\geq 0,\mathbf{Y}\geq 0} \|\mathbf{X} - \mathbf{AY}\|_1$.

[1] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html



(a) ADM algorithm on AT&T dataset (b) Multiplicative algorithm on AT&T dataset (c) Multiplicative algorithm on Umist dataset

Figure 5: Convergence of proposed algorithm for solving Eq.(4) at $\alpha = 2, \beta = 0.1$. x-axis: # of iteration; y-axis: log objective function value of Eq.(4). (a) ADM algorithm on AT&T dataset; (b) Multiplicative algorithm on AT&T dataset; (c) Multiplicative algorithm on Umist dataset.

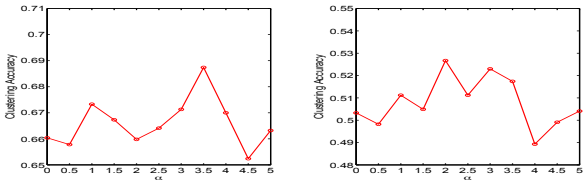
Experiment Settings In all above methods and our method, $\beta = 0.1$ if there is β. α is searched in the following set: $\{0, 0.5, 1, \cdots, 4.5, 5\}$ if there is α. We show the computed accuracy, normalized mutual information, purity results in Table 2.

Results Analysis We make several important observations from experiment results. **(1)** The proposed method is generally better than the other methods, which validates the effectiveness of proposed method for data clustering tasks, in terms of accuracy, normalized mutual information and purity. **(2)** In our method, there are two factors contributing to the performance improvement: (a) robust loss function; (b) sparsity constraint enforced on $\mathbf{Y}$. As compared to the the other loss functions used in non-negative matrix factorization, ℓ_1 loss is more robust for the noises both in data sample space and feature dimension, and thus gives better performance when data are polluted with noises. For the sparsity constraint, it generally promotes the sparsity of cluster indicator, which slightly improves the performance. This generally holds for different versions of NMF with different loss functions, such as least square loss and $\ell_{2,1}$ loss. **(3)** Group sparsity has been widely used in feature learning and variable selection. Our results, however, indicate that the group sparsity does not help much for the improvement of clustering performance, as compared to more general flat sparsity using LASSO. The reason is that, the goal of group sparsity, which is to select the most discriminant features, is inconsistent with the goal of clustering tasks, which is to find the most probable class that a data point is assigned to.

Influence of parameter α In all of our experiments, we fix β values. Only one parameter α is needed to be tuned. We study the the influence of parameter α for the perfor-

Table 2: Accuracy (ACC), Normalized Mutual information (NMI), Purity (PUR) comparisons of different algorithms: k-means clustering, least square NMF (L2NMF), NMF with sparsity constraint (L2NMFs), NMF with group sparsity constraint (L2NMFsg), NMF using $\ell_{2,1}$ error function (L21NMF), NMF using $\ell_{2,1}$ error function with sparsity constraint (L21NMFs), NMF using ℓ_1 error function (L1NMF); our methods of using ℓ_1 error function with sparsity constraint (L1NMFs) on five datasets.

Dataset	Metric	Clustering Methods							
		kmeans	L2NMF	L2NMFs	L2NMFsg	L1NMF	**L1NMFs (ours)**	L21NMF	L21NMFs
AT&T	ACC	0.5700	0.5875	0.5975	0.5925	0.6203	**0.6310**	0.6075	0.6175
	NMI	0.7544	0.7575	0.7589	0.7546	0.7944	**0.8123**	0.7638	0.7737
	PUR	0.6025	0.6175	0.5930	0.6250	0.6525	**0.6673**	0.6425	0.6475
MNIST	ACC	0.5800	0.5733	0.6133	0.6066	0.6604	**0.6733**	0.6333	0.6466
	NMI	0.5717	0.5451	0.5931	0.5984	0.6097	**0.6208**	0.5693	0.5937
	PUR	0.6234	0.6265	0.6479	0.6333	0.6846	**0.6935**	0.6466	0.6666
UMIST	ACC	0.4372	0.4611	0.4616	0.4527	0.4672	**0.4872**	0.4500	0.4711
	NMI	0.6190	0.6005	0.6158	0.6138	0.6490	**0.6690**	0.6323	0.6752
	PUR	0.4872	0.4833	0.4777	0.4933	0.4972	**0.5172**	0.5033	0.4976
YALEB	ACC	0.0866	0.1683	0.1577	0.1673	0.1882	**0.2148**	0.1956	0.2021
	NMI	0.0851	0.2864	0.2418	0.2674	0.2864	**0.3323**	0.3154	0.3214
	PUR	0.0957	0.1769	0.1678	0.1769	0.1983	**0.2343**	0.2046	0.2102
Caltech	ACC	0.4033	0.4167	0.4433	0.4450	0.5033	**0.5267**	0.4500	0.4917
	NMI	0.4272	0.4364	0.4729	0.4737	0.5272	**0.5396**	0.4611	0.5025
	PUR	0.4300	0.4333	0.4700	0.4733	0.5300	**0.5517**	0.4817	0.5200



(a) Mnist (b) Caltech

Figure 6: Clustering Accuracy w.r.t different parameter α on datasets mnist and Caltech.

mance of our algorithm. We show the clustering accuracy w.r.t different α on data set mnist and Caltech in Fig.6. An interesting observation is that, our method is not very sensitive to α. Our method also gives better clustering results at different α values.

Semi-supervised learning experiment

Another interesting application of non-negative dictionary learning is to learn the "self-representation" (i.e., $\mathbf{S}$ computed from Eq.(13)). This can be used to construct a symmetric pairwise similarity $\mathbf{S} = \frac{1}{2}(\mathbf{S} + \mathbf{S}^T)$, because it captures the relations between different data points using sparse representation. Then $\mathbf{S}$ are fed into semi-supervised learning methods, for classification purpose. The goal of this group of experiment is to test the effectiveness of $\mathbf{S}$ used for semi-supervised learning tasks. We adopt three most widely used semi-learning methods: (1) harmonic function (Zhu, Ghahramani, and Lafferty 2003); (2) local and global consistency (Zhou et al. 2004); (3) Green's function (Ding et al. 2007). We note there are other label propagation methods, *e.g.*, (Kong and Ding 2012b), due to space limit, we do not compare against them here.

We compare the classification accuracy using 10%, 20% labeled data against the pairwise similarities computed from another two methods: (1) Gaussian kernel (shown as $\mathbf{W}$ in Tables3), where $\mathbf{W}_{ij} = e^{-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2}$, and bandwidth $\gamma = 0.7/\delta^2$, where δ is the average distance of kNN (k=3) neighbors of all data points; (2) Bi-Stochastication result (shown as $\mathbf{P}$ in Tables 3) (Wang, Li, and König 2010). The experiment results indicate that, generally, Eq.(13) results are better than $\mathbf{W}$ and $\mathbf{P}$ results, except one case on dataset Mnist with LG-consistency method.

Table 3: Accuracy comparisons of semi-supervised learning with 10% labeled data. Learning algorithms used: Harmonic function, Green's function and Local and global consistency (LGC). W: results obtained from standard Gaussian kernel; P: results computed using Bi-Stochastication method (Wang, Li, and König 2010); S: results obtained from Eq.(13). Dataset: AT&T (A), Mnist (M), Umist (U), YaleB (Y), Caltech (C).

data	Harmonic			Green's			LGC		
	W	P	S	W	P	S	W	P	S
A	65.34	66.02	**67.23**	67.13	69.21	**69.23**	68.25	69.23	**69.34**
M	64.53	61.23	**66.91**	62.17	65.23	**65.31**	63.72	**64.76**	63.19
U	44.14	45.98	**46.39**	45.91	45.80	**46.38**	47.87	48.12	**48.21**
Y	26.24	27.43	**28.14**	25.13	23.98	**26.94**	32.02	31.79	**34.23**
C	43.28	42.87	**45.19**	46.47	48.23	**48.24**	42.17	43.01	**43.04**

Conclusion

We present a non-negative dictionary learning method for noisy data, where an efficient multiplicative updating algorithm is derived. We prove the convergence and correctness of the algorithm, and demonstrate its good performance in data clustering and semi-supervised learning tasks. In future,

we will explore how to effectively incorporate the group structure (or hierarchical structure) into the dictionary learning process, e.g., non-convex regularization.

Acknowledgement. This research is partially supported by NSF-CCF-0917274 and NSF-DMS-0915228 grants.

References

Aharon, M.; Elad, M.; and Bruckstein, A. 2006. k-svd: An algorithm for designing overcomplete dictionaries for sparse representation. In *IEEE Transactions on Signal Processing*, volume 54, 4311– 4322.

Bertsekas, D. P. 1996. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific.

Bühler, T., and Hein, M. 2009. Spectral clustering based on the graph p-laplacian. In *ICML*, 11.

Delgado, K. K.; Murray, J. F.; Rao, B. D.; Engan, K.; Lee, T. W.; and Sejnowski, T. J. 2003. Dictionary learning algorithms for sparse representation. *Neural Computation* 15(2):349–396.

Ding, C. H. Q., and Kong, D. 2012. Nonnegative matrix factorization using a robust error function. In *ICASSP*, 2033–2036.

Ding, C.; Zhou, D.; He, X.; and Zha, H. 2006. R_1-pca: rotational invariant l_1-norm principal component analysis for robust subspace factorization. In *ICML*, 281–288.

Ding, C.; Jin, R.; Li, T.; and Simon, H. D. 2007. A learning framework using green's function and kernel regularization with application to recommender system. In *KDD*, 260–269.

Dong, W.; Li, X.; Zhang, L.; and Shi, G. 2011. Sparsity-based image denoising via dictionary learning and structural clustering. In *CVPR*, 457–464.

Dueck, D., and Frey, B. J. 2007. Non-metric affinity propagation for unsupervised image categorization. In *ICCV*.

Elad, M., and Aharon, M. 2006. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing* 15(12):3736–3745.

Kasiviswanathan, S. P.; Melville, P.; Banerjee, A.; and Sindhwani, V. 2011. Emerging topic detection using dictionary learning. In *CIKM*, 745–754.

Ke, Q., and Kanade, T. 2005. Robust l_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *CVPR (1)*, 739–746.

Kim, W.; Chen, B.; Kim, J.; Pan, Y.; and Park, H. 2011. Sparse nonnegative matrix factorization for protein sequence motif discovery. *Expert Syst. Appl.* 38(10):13198–13207.

Kim, J.; Monteiro, R.; and Park, H. 2012. Group sparsity in nonnegative matrix factorization. In *SDM*, 851–862.

Kong, D., and Ding, C. H. Q. 2012a. An iterative locally linear embedding algorithm. In *ICML*.

Kong, D., and Ding, C. H. Q. 2012b. Maximum consistency preferential random walks. In *ECML/PKDD (2)*, 339–354.

Kong, D.; Ding, C. H. Q.; and Huang, H. 2011. Robust nonnegative matrix factorization using l21-norm. In *CIKM*, 673–682.

Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 2278–2324.

Lee, D. D., and Seung, H. S. 2000. Algorithms for non-negative matrix factorization. In *NIPS*.

Lee, H.; Battle, A.; Raina, R.; and Ng, A. 2007. Efficient sparse coding algorithms. In *In NIPS*, 801–808. NIPS.

MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 281–297. University of California Press.

Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; and Zisserman, A. 2008. Supervised dictionary learning. In *NIPS*, 1033–1040.

Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2009. Online dictionary learning for sparse coding. In *ICML*, 87.

Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2010. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research* 11:19–60.

Mallat, S. 1999. *A wavelet tour of signal processing (2. ed.)*. Academic Press.

Olshausen, B. A., and Fieldt, D. J. 1997. Sparse coding with an overcomplete basis set: a strategy employed by v1. *Vision Research* 37:3311–3325.

Protter, M., and Elad, M. 2009. Image sequence denoising via sparse and redundant representations. *IEEE Transactions on Image Processing* 18(1):27–35.

Raina, R.; Battle, A.; Lee, H.; Packer, B.; and Ng, A. Y. 2007. Self-taught learning: transfer learning from unlabeled data. In *ICML*, 759–766.

Tibshirani, R. 1994. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58:267–288.

Wang, F.; Li, P.; and König, A. C. 2010. Learning a bi-stochastic data similarity matrix. In *ICDM*, 551–560.

Zhang, M., and Ding, C. H. Q. 2013. Robust tucker tensor decomposition for effective image representation. In *ICCV*, 2448–2455.

Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; and Scholkopf, B. 2004. Learning with local and global consistency. In *NIPS*, 321–328.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. 2003. Semi-supervised learning using gaussian fields and harmonic functions. *Proc. Int'l Conf. Machine Learning*.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.