Robust Subspace Discovery via Relaxed Rank Minimization

Citation: Wang, Xinggang, Zhengdong Zhang, Yi Ma, Xiang Bai, Wenyu Liu, and Zhuowen Tu. "Robust Subspace Discovery via Relaxed Rank Minimization." Neural Computation 26, no. 3 (March 2014): 611–635. © 2014 Massachusetts Institute of Technology

As Published: http://dx.doi.org/10.1162/NECO_a_00555

Publisher: MIT Press

Persistent URL: http://hdl.handle.net/1721.1/87586

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

LETTER Communicated by John Wright

Robust Subspace Discovery via Relaxed Rank Minimization

Xinggang Wang
xgwang@hust.edu.cn
Huazhong University of Science and Technology, Wuhan, Hubel Province 43007, China

Zhengdong Zhang
zhangzd@mit.edu
MIT, Cambridge, MA 02139

Yi Ma
mayi@microsoft.com
Microsoft Research Asia, Beijing 100080, China

Xiang Bai
xbai@hust.edu.cn
Wenyu Liu
liuwy@hust.edu.cn
Huazhong University of Science and Technology, Wuhan, Hubel Province 43007, China

Zhuowen Tu
ztu@ucsd.edu
Department of Cognitive Science, University of California, San Diego, La Jolla, CA 92093, U.S.A.

This letter examines the problem of robust subspace discovery from input data samples (instances) in the presence of overwhelming outliers and corruptions. A typical example is the case where we are given a set of images; each image contains, for example, a face at an unknown location of an unknown size; our goal is to identify or detect the face in the image and simultaneously learn its model. We employ a simple generative subspace model and propose a new formulation to simultaneously infer the label information and learn the model using low-rank optimization. Solving this problem enables us to simultaneously identify the ownership of instances to the subspace and learn the corresponding subspace model. We give an efficient and effective algorithm based on the alternating direction method of multipliers and provide extensive simulations and experiments to verify the effectiveness of our method. The proposed scheme can also be used to tackle many related high-dimensional combinatorial selection problems.

Neural Computation **26, 611–635** (2014)
doi:10.1162/NECO_a_00555

1 Introduction

Subspace learning algorithms have recently been adopted for analyzing high-dimensional data in various problems (Jenatton, Obozinski, & Bach, 2010; Wright, Yang, Ganesh, Sastry, & Ma, 2009; Wagner, Wright, Ganesh, Zhou, & Ma, 2009). Assuming the data are well aligned and lie in a low-dimensional linear subspace, these methods can deal with large sparse errors and learn the low-rank subspace of data. Other approaches (e.g., Elhamifar & Vidal, 2009; Liu, Lin, & Yu, 2010; Luo, Nie, Ding, & Huang, 2011; Favaro, Vidal, & Ravichandran, 2011) have been proposed to cluster data into different subspaces. However, these methods may have difficulty in dealing with a class of unsupervised learning scenarios in which a large number of outliers exist. In this letter, we propose a method to discover low-dimensional linear subspace from a set of data containing both inliers and a significant number of outliers. Figure 1 gives a typical problem setting of this letter, as well as the pipeline of our proposed solution. Here, we are given a set of images and each image contains a common object (pattern). Our goal is to automatically identify the object and learn its subspace model.

In an abstract sense, we are given a set of data containing both inliers lying in a relative low-dimensional linear subspace and overwhelming outliers; in addition, the inliers may be corrupted by sparse errors. We make use of two constraints that have been adopted in the multiple instance learning (MIL) literature that data are divided into different bags and at least one inlier exists in each bag. These two constraints usually coexist, as is shown in Figures 1a and 1b. We may turn each image into a bag, consider image patches containing objects of the same category as inliers, and treat image patches from background or other categories as outliers. We aim to find the low-dimensional subspace and identify which data belong to the subspace. Obviously this problem is highly combinatorial and high dimensional. Here we borrow the MIL concept but assume no given negative bags in the training process, as in Zhu, Wu, Wei, Chang, and Tu (2012). The original problem becomes a weakly supervised subspace (pattern) discovery problem. We then transfer this problem into a convex optimization formulation, which can be effectively solved by the alternating direction method of multipliers (ADMM) (Gabay & Mercier, 1976; Boyd, Parikh, Chu, Peleato, & Eckstein, 2011) method. In the proposed formulation, each instance is associated with an indicator indicating whether the instance is an inlier or an outlier; this is illustrated in Figure 1b. The indicators of instances are treated as latent variables, and our objective function is to minimize both the rank of the subspace spanned by the selected instance and the ℓ_1 norm of the error in the selected instance. Thus, by solving this optimization problem, we achieve the goal of discovering the low-dimensional subspace and identifying the instances belonging to the subspace. In Figure 1c, we show the discovered face subspace and errors of each face image. We deal with various

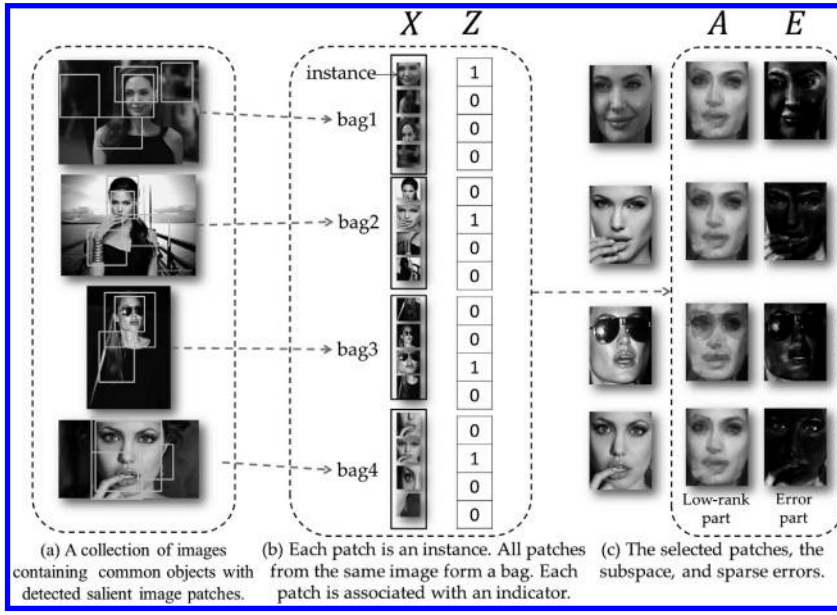

Figure 1: Pipeline of the object discovery task for subspace learning. Given a set of images, we first detect salient image patches (windows). All the image patches from the same image, considered as instances, form a bag. We assume the common object to appear as one instance in each bag. Our algorithm then detects and learns the subspace model for the common object while computing its residue. The symbols X, Z, A, and E at the top of the figure correspond to the notations in our formulation given in section 2.

object discovery tasks to demonstrate the advantage of our algorithm in the experiments. In the remainder of this section, we give the related work of our method.

1.1 Relations to Existing Work. In a nutshell, we are addressing a very challenging subspace learning problem. The existing scalable robust subspace learning methods such as robust principal component analysis (RPCA) (Candes, Li, Ma, & Wright, 2011; Xu, Sanghavi, & Caramanis, 2012) can handle a sparse number of errors or outliers, while the dense error correction method in Wright and Ma (2010) can deal with dense corruptions under some restricted conditions. These do not, however, apply to our case since here the inlying instances are very few compared to the outliers and the inliers might even be partially corrupted. Nevertheless, our problem assumes an important additional structure: we know that there is at least one inlier in each set of samples. We will demonstrate that this extra information assists us in solving a seemingly impossible subspace discovery problem.

Robust principal component analysis (Candes et al., 2011) has been successfully applied in background modeling (Candes et al., 2011), texture analysis (Zhang, Ganesh, Liang, & Ma, 2012), and face recognition (Jia, Chan, & Ma, 2012). It requires input data to be well aligned, a prohibitive requirement in many real-world situations. To overcome this limitation, robust alignment by sparse and low rank (RASL) (Peng, Ganesh, Wright, Xu, & Ma, 2012) was proposed to automatically refine the alignment of input data (e.g., a set of images with a common pattern). However, RASL demands a good initialization of the common pattern on the same scale, whereas here we are dealing with a much less constrained problem in which the common pattern (object) observes large-scale differences at unknown locations in the images.

Robustly learning a model from noisy input data is a central problem in machine learning. In the multiple instance learning (MIL) literature (Dietterich, Lathrop, & Lozano-Pérez, 1997), the input data are given in the form of bags. Each positive bag contains at least one positive instance and each negative bag consists of all negative instances. MIL falls into the class of weakly supervised learning problems. In the MIL setting, the two central subtasks are to infer the missing label information and learn the correct model for the instances. The EM algorithm (Dempster, Laird, & Rubin, 1977) has been widely adopted for inferring missing labels for such MIL problems (Zhang & Goldman, 2001), and likewise for the latent SVM (Yu & Joachims, 2009). One could modify these methods; however, as we will see in our comparison, they lead to greedy iterative optimization that often produces suboptimal solutions. Recently, Lerman, McCoy, Tropp, and Zhang (2012) proposed a convex optimization method, called REAPER, to learn subspace structure in data sets with large fractions of outliers. Compared to an approach like RPCA, this multiplicative approach has better thresholds for recovery in gaussian outlier clouds. However, it is not robust to additional sparse corruption to the data points. The bMCL algorithm (Zhu et al., 2012) deals with cases of both one-class and multiclass object assumptions in object discovery with weak supervision. In previous work, Sankaranarayanan and Davis (2012) proposed one-class multiple instance learning (MIL) along the idea of discriminative models for target tracking. Wang et al. (2012) proposed an EM approach of learning a low-rank subspace for MIL with the one-class assumption. In this letter, we emphasize the task of robust subspace discovery with an explicit generative model for global optimization.

In the rest of the letter, we refer to the common pattern as an object and focus on the problem of object discovery. Given a set of images, our goal is to automatically discover the common object across all the images, which might appear at an unknown location with an unknown size.

Along the line of object discovery, many methods have also been proposed. Russell, Freeman, Efros, Sivic, and Zisserman (2006) treated each image as a bag of visual words in which the common topics are discovered

by the latent Dirichlet allocation. Other systems, such as those of Grauman and Darrell (2006) and Lee and Grauman (2009), perform clustering on the affinity or correspondence matrix established by different approaches using different cues in the images. Although these existing methods achieve promising results on several benchmark data sets, they have notable limitations from several perspectives: there often lacks a clear generative formulation and performance guarantee; some systems are quite complicated, with many cues, classifiers, and components involved; and they have a strong dependence on the discriminative models.

In contrast, this letter explores a different direction by employing a simple generative subspace model and proposes a new formulation to simultaneously infer the label information and learn the model using low-rank optimization. Unlike EM-like approaches, our method is not sensitive to the initial conditions and is robust to severe corruption. Although different from the classical robust PCA methods, our method inherits the same kind of robustness and efficiency from its convex formation and solution. Extensive simulations and experiments demonstrate the advantages of our method.

2 Formulation of Subspace Discovery

Given K bags of candidate object instances, we denote the number of instances in the kth bag as n_k. The total number of instances is $N = n_1 + \cdots + n_K$. Each instance is represented by a d-dimensional vector $x_i^{(k)} \in \mathbf{R}^d$. We may represent all the instances from one bag as columns of a matrix $X^{(k)} = [x_1^{(k)}, \ldots, x_{n_k}^{(k)}] \in \mathbf{R}^{d \times n_k}$. Furthermore, we define $X = [X^{(1)}, \ldots, X^{(K)}] \in \mathbf{R}^{d \times N}$. By default, we assume that each bag contains at least one common object, and the rest are unrelated. To be concrete, we associate each object $x_i^{(k)}$ with a binary label $z_i^{(k)} \in \{0, 1\}$. $z_i^k = 1$ indicates that x_i^k is the common object. Similarly, we define $Z^{(k)} = [z_1^{(k)}, \ldots, z_{n_k}^{(k)}] \in \{0, 1\}^{n_k}$ and $Z = [Z^{(1)}, \ldots, Z^{(K)}]$. We assume that each bag contains at least one common object. So we have $\bigvee_{i=1}^{n_k} z_i^{(k)} = 1, \forall k \in [K]$, where $\bigvee$ is an "or" operator and $[K] = \{1, 2, \ldots, K\}$ is the set of positive integers less than or equal to K.

In general, different instances of the same object are highly correlated. It is reasonable to assume that such instances lie on a low-dimensional subspace $\Omega \subset \mathbf{R}^d$. This assumption can be verified empirically for real data. Figure 4c shows a comparison of the spectrum of a number of instances that are from the same object or from random image patches. Even if one applies a robust dimensionality reduction to the set of random image patches, their spectrum is still much higher than those for a common object.

However, due to many practical nuisance factors in real images, such as variation of pose, cast shadow, and occlusion, the observed instances of the common objects may no longer lie in a low-dimensional subspace. We may model all of these contaminations as sparse errors added to the instances.

So we could model each instance as $x = a + e$, where $a \in \Omega$ and e is a sparse vector. The observed instances of the common objects may no longer lie in a low-dimensional subspace. We may model the contamination as sparse errors added to the instances. So we could model each instance as $x = a + e$, where $a \in \Omega$ and e is a sparse vector.

From the given K bags of instances $X = [X^{(1)}, \ldots, X^{(K)}]$, our goal is to find one or more instances from each bag so that all the selected instances form a low-rank matrix A, subject to some sparse errors E. Or equivalently, we need to solve the following problem,

$$\begin{aligned} &\min_{A,E,Z} \operatorname{rank}(A) + \gamma \|E\|_0 \\ &\text{s.t.} \quad X\operatorname{diag}(Z) = A + E, \ \forall k \in [K] \ \textstyle\bigvee_{i=1}^{n_k} z_i^k = 1, \end{aligned} \tag{2.1}$$

where $\operatorname{diag}(Z)$ is an $N \times N$ block-diagonal matrix with K blocks $\left\{\operatorname{diag}\left(Z^{(k)}\right)\right\}$. To distinguish with the conventional (robust) subspace learning problems, we could refer to this problem as subspace discovery.

3 Solution via Convex Relaxation

The problem in equation 2.1 is a highly combinatorial optimization problem that involves both continuous and integer variables. It is generally intractable when the dimensions d and N are large. The recent theory of RPCA (Candes et al., 2011) suggests that rank and sparsity can be effectively minimized via their convex surrogates. Therefore, we could replace the above objective function $\operatorname{rank}(\cdot)$ with the nuclear norm $\|\cdot\|_*$ and ℓ_0 norm with ℓ_1 norm. Thus, equation 2.1 is replaced with the following program:

$$\begin{aligned} &\min_{A,E,Z} \|A\|_* + \lambda \|E\|_1 \\ &\text{s.t.} \quad X\operatorname{diag}(Z) = A + E, \ \forall k \in [K] \ \textstyle\bigvee_{i=1}^{n_k} z_i^{(k)} = 1. \end{aligned} \tag{3.1}$$

Notice that although the objective function is now convex, the constraints on all the binary variables $z_i^{(k)}$ make this program intractable.

3.1 A Naive Iterative Solution. We can use a naive way (Wang et al., 2012) to tackle the problem in equation 3.1 by alternating between estimating Z and minimizing the objective with respect to the low-rank A and sparse E in a spirit similar to the EM algorithm. With Z fixed, equation 3.1 becomes a convex optimization problem and can be solved by the RPCA method (Candes et al., 2011). Once the low-rank matrix A is known, one could perform ℓ_1-regression to evaluate the distance between each point and the subspace:

$$e_i^{(k)} = \min_w \|Aw - x_i^{(k)}\|_1. \tag{3.2}$$

Then within each bag, we reassign 1 to a number of instances with errors below a certain threshold and mark the rest as 0. One can iterate this process until convergence. Because there are many outliers, this naive iterative method is very sensitive to initialization, so we have to run this naive method many times with random initializations and pick the best solution. This is similar to the popular RANSAC scheme for robust model estimation. Suppose there are m_k positive instances within the kth bag; then the probability that RANSAC would succeed in selecting only the common objects is $\Pi_{k=1}^{K}(\frac{m_k}{n_k})$. Typically $\forall k, m_k/n_k \leq \frac{1}{5}$, so the probability that RANSAC succeeds vanishes exponentially as the number of objects increases. Even if the correct instances are selected, the above ℓ_1 regression is not always guaranteed to work well when A contains errors. Nevertheless, with careful initialization and tuning, this method can be made to work for some relatively easy cases and data sets. It can be used as a baseline method to evaluate the improved effectiveness of any new algorithm.

3.2 Relaxing Z. Instead of enforcing the variable Z to be binary $\{0, 1\}$, we relax it to have real value in $\mathbf{R}$. Also, the constraint $\bigvee_{i=1}^{n_k} z_i^{(k)} = 1$ can be relaxed with its continuous version $\mathbf{1}^T Z^{(k)} = 1$, which is linear. So the optimization problem becomes

$$\begin{aligned} &\min_{A,E,Z} \|A\|_* + \lambda \|E\|_1, \\ &\text{s.t.} \quad X\text{diag}(Z) = A + E, \quad \forall k \in [K], \mathbf{1}^T Z^{(k)} = 1. \end{aligned} \tag{3.3}$$

Although we do not explicitly require Z to be nonnegative, it turns out that the optimal solution to the above program always ensures $Z^* \geq 0$, as theorem 1 shows. This is due to some special properties of the nuclear norm and ℓ_1 norm. For our problem, this is incredibly helpful since the efficiency of the proposed algorithm based on the augmented Lagrangian method decreases quickly as the number of constraints increases. This fact saves us from imposing N extra inequality constraints on the convex program.

Theorem 1. *If none of the columns of X is zero, the optimal solution Z^* of equation 3.3 is always nonnegative.*

Proof. Suppose we are given an optimal solution (A, E, Z) where Z have negative entries. Let us consider the triple $(\hat{A}, \hat{E}, \hat{Z})$ constructed in the following way:

$$\begin{aligned} \hat{Z}^{(k)} &= \frac{1}{1^T|Z^{(k)}|}|Z^{(k)}|, \\ \hat{A}^{(k)} &= \frac{1}{1^T|Z^{(k)}|}A^{(k)}\text{diag}\big(\text{sign}\big(Z^{(k)}\big)\big), \\ \hat{E}^{(k)} &= \frac{1}{1^T|Z^{(k)}|}E^{(k)}\text{diag}\big(\text{sign}\big(Z^{(k)}\big)\big). \end{aligned} \tag{3.4}$$

Since $X\text{diag}(Z) = A + E$, obviously $X\text{diag}(\hat{Z}) = \hat{A} + \hat{E}$; thus, $(\hat{A}, \hat{E}, \hat{Z})$ is a feasible solution, and $\hat{Z}$ is nonnegative. We will show that $\|\hat{A}\|_* + \lambda\|\hat{E}\|_1 < \|A\|_* + \lambda\|E\|_1$, contradicting the fact that (A, E, Z) is optimal. Note that flipping the sign of any column of the matrix will not change the singular value of a matrix and thus has no effect on the nuclear norm of it (if the SVD of $W = U\Sigma V^*$, $\text{diag}(\pm 1, \ldots, \pm 1)V$ is still orthogonal matrix). So if we construct another matrix A' such that $A'^{(k)} = A^{(k)}\text{diag}(\text{sign}(Z^{(k)}))$, $\|A'\|_* = \|A\|_*$. Similarly we construct an E' and $\|E'\|_1 = \|E\|_1$. So $\hat{A}$ and $\hat{E}$ are just column-wise downscaled version of A' and E'. Since for the kth bag $1^T Z^{(k)} = 1, 1^T|Z^{(k)}| > 1$ if and only if any entry of $Z^{(k)}$ is negative; otherwise, $1^T|Z^{(k)}| = 1$. The columns of A' and E' in the bags with negative $Z^{(k)}$ are downscaled by a scalar $\alpha^k \in (0, 1)$. It can be proved that any downscaling of a nonzero column of a matrix will decrease the nuclear norm.

Lemma 1. *Given any matrix $Q \in \mathbf{R}^{m\times n}$, if $\tilde{Q}$ is Q with some column scaled by some scalar $\alpha \in (0, 1)$, then $\|\tilde{Q}\|_* < \|Q\|_*$.*

Proof. Without loss of generality, we assume that the last column q_n gets scaled. Let $Q = [Q_{n-1},\ q_n]$, and let $Q' = [Q_{n-1},\ 0]$ be the matrix by setting the last column to 0. The singular values of Q' are just the union of singular values of Q_{n-1} and an additional 0. Let $t = \min\{m, n\}$. According to Horn and Johnson (2012, theorem 7.3.9), $\sigma_1(Q) \geq \sigma_1(Q') \geq \sigma_2(Q) \geq \sigma_2(Q') \geq \ldots \geq \sigma_t(Q) \geq \sigma_t(Q') \geq 0$. So naturally $\|Q\|_* \geq \|Q'\|_*$, and the equality holds only if $\sigma_i(Q) = \sigma_i(Q'), \forall i \in [t]$. This is impossible since $\|Q\|_F^2 = \sum_i \sigma_i(Q)^2 > \|Q'\|_F^2 = \sum_i \sigma_i(Q')^2$. We must have $\|Q\|_* > \|Q'\|_*$.

Note that $\tilde{Q} = \alpha Q + (1 - \alpha)Q'$ and the nuclear norm $\|\cdot\|_*$ is convex. By applying Jensen's inequality, we have

$$\|\tilde{Q}\|_* \leq \alpha\|Q\|_* + (1 - \alpha)\|Q'\|_* < \alpha\|Q\|_* + (1 - \alpha)\|Q\|_* = \|Q\|_*. \tag{3.5}$$

$\hat{A}$ can be viewed as a sequence of downscaling on different columns of A, and each downscaling will decreases the nuclear norm. The same goes for the ℓ_1 norm of the sparse error E. This shows that $\|\hat{A}\|_* + \lambda\|\hat{E}\|_1 < \|A\|_* + \lambda\|E\|_1$, which contradicts the assumption that (A, E, Z) is optimal.

3.3 Solving equation 3.3 via Alternating Direction Method of Multipliers. We apply the alternating direction method of multipliers (ADMM) to solve equation 3.3. First, write the augmented Lagrangian function:

$$\begin{aligned} L(A, E, Z, Y_0, Y_1, \ldots, Y_K) &\doteq \|A\|_* + \lambda\|E\|_1 \\ &+ \langle Y_0, X\text{diag}\,(Z) - A - E\rangle + \frac{\mu}{2}\|X\text{diag}\,(Z) - A - E\|_F^2 \\ &+ \textstyle\sum_{k=1}^{K}\left(\langle Y_k, \mathbf{1}^T Z^{(k)} - 1\rangle + \frac{\mu}{2}\|\mathbf{1}^T Z^{(k)} - 1\|_F^2\right). \end{aligned} \tag{3.6}$$

Instead of following the exact ALM procedure, we adopt the approximation scheme in Boyd et al. (2011) and Lin, Chen, and Ma (2010), which basically alternates the minimization with respect to the three sets of variables in each iteration t:

$$\begin{cases} A_{t+1} = \underset{A}{\operatorname{argmin}} L(A, E_t, Z_t, Y_t, \mu_t) \\ \qquad = \underset{A}{\operatorname{argmin}} \|A\|_* + \dfrac{\mu_t}{2} \left\| X\operatorname{diag}(Z_t) - A - E_t + \dfrac{Y_{0,t}}{\mu_t} \right\|_F^2, \\ E_{t+1} = \underset{E}{\operatorname{argmin}} L(A_{t+1}, E, Z_t, Y_t, \mu_t) \\ \qquad = \underset{E}{\operatorname{argmin}} \|E\|_1 + \dfrac{\mu_t}{2} \left\| X\operatorname{diag}(Z_t) - A_{t+1} - E + \dfrac{Y_{0,t}}{\mu_t} \right\|_F^2, \\ Z_{t+1} = \underset{Z}{\operatorname{argmin}} L(A_{t+1}, E_{t+1}, Z, Y_t, \mu_t) \\ \qquad = \underset{Z}{\operatorname{argmin}} \left\| X\operatorname{diag}(Z) - A_{t+1} - E_{t+1} + \dfrac{Y_{0,t}}{\mu_t} \right\|_F^2 \\ \qquad\quad + \ldots \sum_{k=1}^{K} \left\| \mathbf{1}^T Z^{(k)} - 1 + \dfrac{Y_{k,t}}{\mu_t} \right\|_F^2. \end{cases} \tag{3.7}$$

Fortunately, the three minimization problems all have closed-form solutions. We next provide the details.

Let $\mathcal{S}_\epsilon(\cdot)$ be the following shrinkage operator:

$$\mathcal{S}_\epsilon(x) = \begin{cases} x - \varepsilon, & \text{if } x > \varepsilon \\ x + \varepsilon, & \text{if } x < -\varepsilon \\ 0, & \text{otherwise} \end{cases}. \tag{3.8}$$

If the SVD of $X\operatorname{diag}(Z_t) - E_t + \frac{Y_{0,t}}{\mu_t} = U\Sigma V^*$, then the optimal A_{t+1} is given as $A_{t+1} = U\mathcal{S}_{\frac{1}{\mu}}(\Sigma)V^*$. For E_{t+1}, the optimal solution is $\mathcal{S}_{\frac{\lambda}{\mu_t}}(X\operatorname{diag}(Z_t) - A_{t+1} + \frac{Y_{0,t}}{\mu})$. For Z, we can solve the original optimization using K independent ones for $Z^{(k)}$. Each suboptimization is a typical least square problem for $Z^{(k)}$:

$$\begin{aligned} Z_{t+1}^{(k)} = \underset{Z^{(k)}}{\operatorname{argmin}} & \left\| X^{(k)}\operatorname{diag}(Z^{(k)}) - A_{t+1}^{(k)} - E_{t+1}^{(k)} + \frac{Y_{0,t}^{(k)}}{\mu_t} \right\|_F^2 \\ & \ldots + \left\| \mathbf{1}^T Z^{(k)} - 1 + \frac{Y_{k,t}}{\mu_t} \right\|_F^2. \end{aligned} \tag{3.9}$$

To be brief, let us denote $P^{(k)} = A_{t+1}^{(k)} + E_{t+1}^{(k)} - \mu_t^{-1} Y_{0,t}^{(k)} \in \mathbf{R}^{d \times n_k}$; we mark the ith column of $P^{(k)}$ as $P_i^{(k)}$ and $Q^{(k)} = 1 - \mu_t^{-1} Y_{k,t} \in \mathbf{R}^1$. Furthermore, we define

$$X_R^{(k)} = \begin{bmatrix} x_1^{(k)} & & \\ & \ddots & \\ & & x_{n_k}^{(k)} \end{bmatrix}$$

and $P_R^{(k)} = \mathrm{vec}(P^{(k)})$. Thus, equation 3.9 can be rewritten as

$$\begin{aligned} Z_{t+1}^{(k)} &= \underset{Z^{(k)}}{\mathrm{argmin}} \left\| X_R^{(k)} Z^{(k)} - P_R^{(k)} \right\|_F^2 + \left\| \mathbf{1}^T Z^{(k)} - Q^{(k)} \right\|_F^2 \\ &= \left\| \begin{bmatrix} X_R^{(k)} \\ \mathbf{1}^T \end{bmatrix} Z_{t+1}^{(k)} - \begin{bmatrix} P_R^{(k)} \\ Q^{(k)} \end{bmatrix} \right\|_F^2 . \end{aligned} \tag{3.10}$$

Directly applying the standard least square technique would require us to compute the pseudo-inverse of $X_R^{(k)} \in \mathbf{R}^{(dn_k+1) \times n_k}$, which is high dimensional, so we perform a trick so that pseudo-inverse is calculated only for a matrix in $\mathbf{R}^{n_k \times n_k}$:

$$\begin{aligned} Z_{t+1}^{(k)} &= \left(\begin{bmatrix} X_R^{(k)T} & \mathbf{1} \end{bmatrix} \begin{bmatrix} X_R^{(k)} \\ \mathbf{1}^T \end{bmatrix} \right)^{\dagger} \begin{bmatrix} X_R^{(k)T} & \mathbf{1} \end{bmatrix} \begin{bmatrix} P_R^{(k)} \\ Q \end{bmatrix} \\ &= ((X_R^{(k)T}) X_R^{(k)} + \mathbf{1} \cdot \mathbf{1}^T)^{\dagger} (X_R^{(k)T} P_R^{(k)} + \mathbf{1} \cdot Q^{(k)}) \\ &= \begin{bmatrix} (x_1^{(k)})^T x_1^{(k)} + 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & (x_{n_k}^{(k)})^T x_{n_k}^{(k)} + 1 \end{bmatrix}^{\dagger} \\ &\quad \cdots \begin{bmatrix} (x_1^{(k)})^T P_1 + Q^{(k)} \\ \vdots \\ (x_{n_k}^{(k)})^T P_{n_k} + Q^{(k)} \end{bmatrix} . \end{aligned} \tag{3.11}$$

Algorithm 1 : ADMM for Robust Subspace Discovery.

Input: Bags X and λ.

$Z_0 = 0, Y_0 = 0; \forall k \in [K], Y_{k,0}^{(k)} = 0; E_0 = 0; \mu_0 > 0; \rho > 1; t = 0$
while not converged **do**
  // Line 4–5 solve $A_{t+1} = \text{argmin}_A L(A, E_t, Z_t, Y_t, \mu_t)$.
  $[U, \Sigma, V^*] = \text{svd}(X\text{diag}(Z_t) - E_t + \frac{Y_{0,t}}{\mu_t})$;
  $A_{t+1} = U\mathcal{S}_{\frac{1}{\mu}}(\Sigma)V^*$.
  // Line 7 solves $E_{t+1} = \text{argmin}_E L(A_{t+1}, E, Z_t, Y_t, \mu_t)$.
  $E_{t+1} = \mathcal{S}_{\frac{\lambda}{\mu_t}}\left(X\text{diag}(Z_t) - A_{t+1} + \frac{Y_{0,t}}{\mu}\right)$.
  // Line 9–12 solve $Z_{t+1} = \text{argmin}_Z L(A_{t+1}, E_{t+1}, Z, Y_t, \mu_t)$.
  for $k = 1 \to K$ **do**
    Obtain $Z_{t+1}^{(k)}$ via equation 3.11.
  end for
  $Z_{t+1} = [Z_{t+1}^{(1)}, \ldots, Z_{t+1}^{(K)}]$.
  // Line 14–16 update Y_{k+1} and μ_{k+1}.
  $Y_{0,t+1} = Y_{0,t} + \mu_t (X\text{diag}(Z_{t+1}) - A_{t+1} - E_{t+1})$.
  $Y_{k,t+1} = Y_{k,t} + \mu_t \left(\mathbf{1}^T Z_{t+1}^{(k)} - 1\right), \forall k \in [K]$.
  $\mu_{t+1} = \rho\mu_t$.
  $t \leftarrow t + 1$.
end while

Output: the converged values for (A, E, Z).

After A, E, and Z are updated, we need only to perform a gradient ascent on the dual variable Y_t:

$$Y_{0,t+1} = Y_{0,t} + \mu_t \left(X\text{diag}(Z_{t+1}) - A_{t+1} - E_{t+1}\right),$$
$$Y_{k,t+1} = Y_{k,t} + \mu_t\left(\mathbf{1}^T Z_{t+1}^{(k)} - 1\right).$$

And μ is also updated by $\mu_{k+1} = \rho\mu_k$, $\rho > 1$. The complete algorithm is summarized as algorithm 1.

The alternating minimization process in equation 3.7 is known as the alternating direction method of multipliers (ADMM) (Gabay & Mercier, 1976). A comprehensive survey of ADMM is given in Boyd et al. (2011); Lin et al. (2010) introduced it to the field of low-rank optimization. ADMM is not always guaranteed to converge to the optimal solution. If there are only two alternating terms, its convergence has been well studied and established in Gabay and Mercier (1976). However, less is known for the convergence of cases where there are more than two alternating terms, despite the strong empirical observations (Zhang et al., 2012). Tao and Yuan (2011) obtained convergence for a certain family of three-term alternation functions (applied to the noisy principal component pursuit problem). However, the scheme they proposed is different from the direct ADMM in equation 3.7, and it is also computationally heavy in practice. The convergence of the

general ADMM remains an open problem, although in practice, there is a simple and fast implementation. Nevertheless, during the submission of this letter, there was some development in the study of ADMM (Shiqian Ma & Zou, 2013) suggesting that one can design a convergent ADMM algorithm for the problem studied here. For instance, we could simply group the variables E and Z together and apply the proximal ADMM algorithm suggested in Shiqian Ma and Zou (2013), which results in a slight modification to the proposed algorithm. Such proximal ADMM is guaranteed to converge. However, in practice, it might not converge faster than the proposed algorithm, which exploits the natural separable structures in the augmented Lagrangian function among the three sets of variables A, E, and Z. In our experience, the proposed algorithm works extremely well in practice and meets our application goals.

4 Simulations and Experiments

In this section, we conduct simulations and experiments on both synthetic and real data for different applications for object discovery to verify the effectiveness of our method. We name the method described in section 3.1 the naive iterative method (NIM) and call the relaxed method ADMM. In all our experiments, we set $\lambda = 1/\sqrt{d}$, where d is the dimension of instance feature.

4.1 Robust Subspace Learning Simulation. In order to investigate the ability of the proposed ADMM method to recover the indicators of inlier instances, in this experiment, we generate synthetic data with 50 bags; in each bag, there are 10 instances—1 positive instance and 9 negative instances. The dimension of instance is $d = 500$. First, the positive instances are generated by linearly combining r randomly generated $d \times 1$ vector whose entries are independent and identically distributed (i.i.d.). Standard gaussian, and the negative instances are independently randomly generated $d \times 1$ vector following i.i.d. normal distribution. Then, for every instance (no matter whether it is positive or negative), we normalize it to make sure its ℓ_2-norm is 1. Finally, large sparse errors are added to all instances; the sparsity ratio of the error is s, and the values of the error are uniformly distributed in the range of $[-1, 1]$.

We investigate the performance when r (the rank of subspace) and s (the sparsity level the error) vary. r ranges from 1 to 31, and s ranges from 0 to 0.3. For each test, we denote the ground-truth indicator vector as Z^*, the recovered indicator vector as $\hat{Z}$, the set of indexes whose corresponding values in Z^* are 1 as I^*, and the set of indexes whose corresponding values in $\hat{Z}$ are larger than a threshold $\tau \in [0, 1]$ as $\hat{I}$. The accuracy of the recovered indicators is defined as accuracy $= \frac{\#(I^* \cap \hat{I})}{\#(I^*)}$. Given the ratio of sparsity and the rank of subspace, we run five random tests and report the average accuracy of the recovered indicators for ADMM (under different τ) and

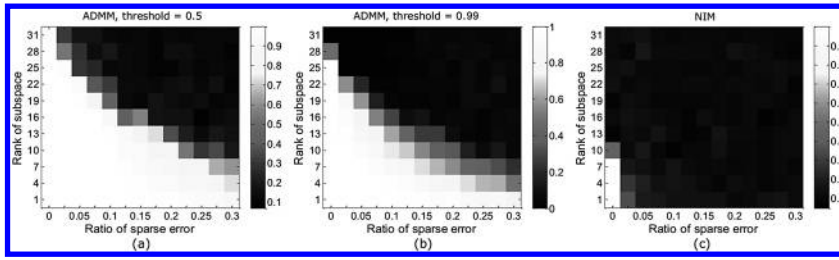

Figure 2: Accuracy of the recovered indicators when the sparsity level of error and the rank of subspace vary for ADMM at different thresholds and NIM. (a) The accuracy of ADMM at $\tau = 0.5$. (b) The accuracy of ADMM at $\tau = 0.99$. (c) The accuracy of NIM.

NIM (randomly initialized) in Figure 2. In Figure 2a, $\tau = 0.5$; 0.5 is a fair value, since there is only one recovered indicator with a value larger than 0.5. In Figure 2b, $\tau = 0.99$, a very strict value; the accuracy matrix of the recovered indicators under $\tau = 0.99$ shows how exact our relaxation is in section 3.2. The solution of NIM in Figure 2c is discrete, and it is not necessary to set a threshold for the solution. From the results in Figure 2, we see that NIM can work only when the positive instances are in a very low-rank subspace in the situation of no error. No matter whether the threshold is 0.5 or 0.99, the working ranges of ADMM are strikingly larger than NIM. Comparing the results in Figures 2a and 2b, we find that it requires positive instances to be in a lower-dimensional subspace and contain less error if we want to exactly recover the indicators of them—say, the indicator values of recovered instances are larger than 0.99.

4.1.1 Multiple Positive Instances in One Bag. The simulations are focused on the situation where only one positive instance exists in each bag. Now we study how ADMM can deal with the situation when multiple instances exist in each bag. We put three positive instances in each bag; they are randomly drawn from the same subspace and corrupted with large sparse errors. Thus, the three positive instances in each bag are not identical. For different values of r and s, we run ADMM five times. The values of the recovered indicators are used for plotting the precision-recall curve. Results are shown in Figure 3. Given a threshold τ, precision and recall are calculated by $\text{precision} = \frac{\#(I^* \cap \hat{I})}{\#(\hat{I})}$ and $\text{recall} = \frac{\#(I^* \cap \hat{I})}{\#(\text{all positive instances})}$. As shown in Figure 3(a), the performance of ADMM increases as the error becomes sparser; when there is no error, ADMM is able to perfectly identify all positive instances. Figure 3b shows that it requires the subspace to have higher rank if more positive instances exist. When the rank of subspace is 15 and the sparsity level of error is 10%, ADMM is able to recover the indicators of 99% positive instances with 100% precision. It is observed that the current formulation

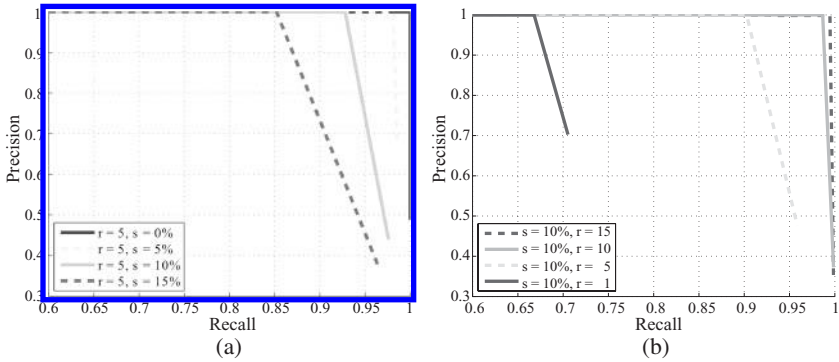

Figure 3: Precision-recall curves for the recovered indicator values by ADMM. (a) The rank of subspace is fixed at $r = 5$; the sparsity level of error $s =$ 0%, 5%, 10%, 15%. (b) The sparsity level of error is fixed at $s = 10\%$; the rank of subspace $r = 1, 5, 10, 15$.

for the subspace discovery problem in equation 3.3 has difficulty in dealing with multiple positive instances in some other settings.

4.2 Aligned Face Discovery Among Random Image Patches. We illustrate the effect of ADMM for object discovery by finding well-aligned face images among lots of randomly selected image patches. Face images are from the Yale face data set (Georghiades, Belhumeur, & Kriegman, 2001), which consists of 165 frontal faces of 15 individuals. Other image patches are randomly selected from the PASCAL image data set (Everingham, Van Gool, Williams, Winn, & Zisserman, 2011). We design bags and instances as follows: the 165 face images are in 165 bags; other than the face image, in each bag, there are 9 image patches from the PASCAL data set; every image or patch is normalized to 64×64 pixels and then vectorized to be a 4096-dimensional feature. Some of images in bags are shown in Figure 4a.

To evaluate the performance of this face recovery task, we get the images with the maximum indicator value in each bag and then calculate the percentage of Yale faces among these images as the accuracy of face discovery. Because negative instances are randomly selected, we run the experiments five times. The average accuracy and the standard deviation of ADMM and NIM (randomly initialized) are $99.5 \pm 0.5\%$ and $77.8 \pm 3.5\%$, respectively. Some of the discovered faces by ADMM are shown in Figure 4b. As it shows, facial expression and glasses are removed from the original images so that the repaired faces are better approximated by a low-dimensional subspace.

4.3 Object Discovery on Real-World Images. The task of object discovery has become a major topic to reduce manual labeling effort to learn object

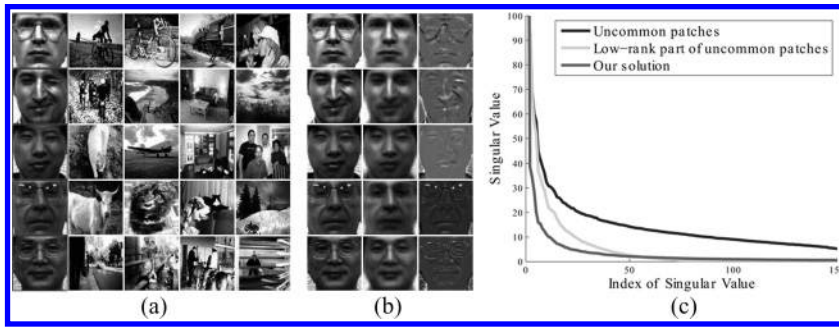

Figure 4: (a) Each row shows some sampled images in one bag. (b) Face discovery results by our algorithm. The first column shows the original patches in the bag; the second and third columns show recovered low-rank and sparse components, respectively. (c) The distributions of singular values of our solution, the low-rank component of uncommon patches via RPCA (Candes et al., 2011), and the original uncommon patches in the experiment.

class; it is a challenging task. In this task, we are given a set of images, each containing one or more instances of the same object class. In contrast to the fully supervised scenario, the locations of objects are not given. Different from subspace learning with simulated data, the appearance of an object varies a lot in real-world images, and thus is different from subspace learning with simulated data; instead, it requires using image descriptors that are somewhat robust to substantial pose variations, for example, HoG and LBP. Moreover, the location and scale of the objects are unknown, which means the number of instances can rise to the millions. To address this problem, we use an existing unsupervised salient object detection algorithm (e.g., Feng, Wei, Tao, Zhang, & Sun, 2011), to reduce the number of instances per bag or image. The reason for us to choose the HoG and LBP descriptors for characterizing object is due to the observation that objects from the same category with the same view may not have similar color or texture. However, they often have similar shapes. Both HoG and LBP show good performance in supervised object detection (Felzenszwalb, Girshick, McAllester, & Ramanan, 2010; Ahonen, Hadid, & Pietikainen, 2006). The common shape structures of objects are the subspaces we want to discover.

In the experiments of object discovery on real-world images, we evaluate the proposed ADMM algorithm on four diverse data sets: the PASCAL 2006 data set (Everingham, Zisserman, Williams, & Van Gool, 2006), the PASCAL 2007 data set (Everingham, Van Gool, Williams, Winn, & Zisserman, 2007), the face detection data set and benchmark (FDDB) subset (Jain & Learned-Miller, 2010), and the ETHZ Apple logo class (Ferrari, Tuytelaars, & Van Gool, 2006), and compare ADMM with the state-of-the-art object discovery methods. Because different performance evaluation protocols are used, we

Table 1: Object Discovery Performance Evaluated by CorLoc on PASCAL 2006 and 2007 Data Sets.

	PASCAL06-		PASCAL07-	
Method	6×2	all	6×2	all
ESS (Lampert et al., 2009)	24	21	27	14
Russell et al. (2006)	28	27	22	14
Chum and Zisserman (2007)	45	34	33	19
ADMM (our method)	57	43	40	27
Deselaers et al. (2012)	**64**	**49**	50	28
Pandey and Lazebnik (2011)	NA	NA	**61**	**30**

Note: The numbers in bold indicate the best results in each column.

give the experimental results for the PASCAL 2006 and 2007 data sets and for the FDDB subset and ETHZ Apple logo class in two different parts.

4.3.1 PASCAL 2006 and 2007 Data Sets. The PASCAL 2006 and 2007 data sets are challenging and have been widely used as benchmarks for evaluating supervised object detection and image classification systems. For the object discovery task, we follow the protocol of Deselaers, Alexe, and Ferrari (2012). The performance is evaluated by the CorLoc measure, which is the percentage of correctly localized objects, according to the PASCAL criterion (window intersection-over-union > 0.5). Two subsets are taken from the PASCAL 2006 and 2007 data sets: PASCAL06-6×2, PASCAL06-all, PASCAL07-6×2, and PASCAL07-all. PASCAL06-6×2 contains 779 images from 12 classes or views; PASCAL06-all contains 2184 images from 33 classes or views; PASCAL07-6×2 contains 463 images from 12 classes or views; and PASCAL07-all contains 2047 images from 45 classes or views. (For more details about the data sets, as well as the evaluation protocol, refer to Deselaers et al., 2012.)

Each image is considered as a bag, and a patch in the image detected by the salient object detector in Feng et al. (2011) is considered an instance. The parameter of score threshold in Feng et al. (2011) is denoted as τ_s, which controls the number of salient objects detected. Standard HoG and LBP features are then extracted for each image patch. We let $\tau_s = 0.22$ for the PASCAL06-6×2 and PASCAL06-all data sets and use $\tau_s = 0.165$ for the PASCAL07-6×2 and PASCAL07-all data sets. We run the proposed ADMM method on these images and report the image patch with the maximum indicator value as the detected object. The results of ADMM are reported in Table 1 and compared with the results of other methods (Pandey & Lazebnik, 2011; Deselaers et al., 2012; Chum & Zisserman, 2007; Russell, Freeman, Efros, Sivic, & Zisserman, 2006; Lampert, Blaschko, & Hofmann, 2009).

Table 1 shows favorable results by our method compared with those by Chum and Zisserman (2007), Russell et al. (2006), and Lampert et al. (2009).

Figure 5: Red rectangles: object discovery results of ADMM on the challenging PASCAL 2007. Green rectangles: annotated object ground-truth. From top to bottom: airplane, bicycle, bus, motorbike, plotted plants and TV monitors. (To view this figure in color see the online supplement, available at http://www.mitpressjournals.org/doi/suppl/10.1162/NECO_a_00555.)

The state-of-the-art performances are reported in Pandey and Lazebnik (2011) and Deselaers et al. (2012), which either uses extra bounding-box annotations or adopts complicated object models (Felzenszwalb et al., 2010). Here we study a generative model of subspace learning with a clean and effective solution. Figure 5 shows some discovered objects on the PASCAL-all data set.

4.3.2 FDDB Subset and ETHZ Apple Logo Class. The FDDB subset contains 440 face images; the ETHZ Apple logo class contains 36 images with Apple logos. The appearance of objects and the background of the two data sets are quite diverse. In these two data sets, we use HoG only as the descriptor. Coordinating with the formulation in this letter, the low-rank term corresponds to the common shape structures of faces or apple logos, since we use the HoG as the descriptor; the sparse error term corresponds to the occlusions and the appearance variations in faces or apple logos. We run

Table 2: Performance Comparison with APs for SD (Feng et al., 2011), bMCL (Zhu et al., 2012), NIM-SD, NIM-Rand, and ADMM on FDDB Subset.

Method	FDDB Subset	ETHZ Apple Logo
SD	0.148	0.532
bMCL	0.619	0.697
NIM-SD	0.671	0.826
NIM-Rand	0.669	0.726
ADMM (our method)	**0.745**	**0.836**

Note: The numbers in bold indicate the best results in each column.

ADMM and get the indicator value of each instance. For each image, the indicator value is normalized by dividing the maximum indicator value in the bag; the normalized indicator value is used as the score of each patch.

A selected patch is correct if it intersects with the ground truth object by more than half of their union (PASCAL criteria). Object discovery performance is evaluated by precision-recall curves (Everingham et al., 2011), generated by varying the score threshold, and average precision (AP) (Everingham et al., 2011), computed by averaging multiple precisions corresponding to different recalls at regular intervals.

We compare ADMM with four methods: the baseline saliency detection method (SD) in Feng et al. (2011), the state-of-the-art discriminative object discovery approach named bMCL in Zhu et al. (2012), the naive iterative method initialized with the saliency score (NIM-SD), and the naive iterative method with random initialization (NIM-Rand). The parameters of the four methods are adjusted to make sure they achieve their best performances. The AP of NIM-Rand is the average value of three rounds. APs of all four methods are compared with ADMM in Table 2 on both data sets. As we can see, ADMM significantly improves the results from the saliency detection and outperforms all the other competing methods. The precision-recall curves of the four methods in Figure 6 confirm this as well. The SD method is a purely bottom-up approach. The other three methods make the assumption that all of the input images contain a common object class of interest. The bMCL method (Zhu et al., 2012) is a discriminative method; it obtains state-of-the-art performances on image data sets with a simple background, such as the SIVAL data set (Rahmani, Goldman, Zhang, Krettek, & Fritts, 2005). The images in the FDDB data set are more cluttered, posing additional difficulty. Our methods, both ADMM and NIM-SD, are able to deal with a cluttered background since they do not seek to discriminate the object from the background, an important property in tackling the problem object discovery and subspace learning. The patches with maximum scores by SD, bMCL, NIM-SD, and ADMM are shown in Figure 7.

In the experiments, we observe that there are situations in which ADMM might fail: the objects are not contained in the detected salient image

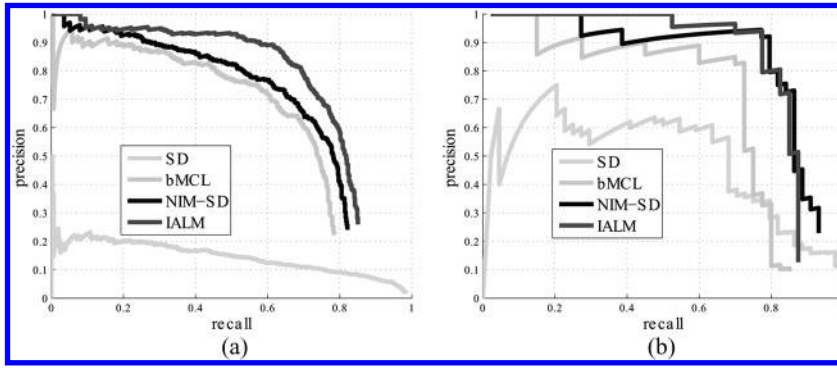

Figure 6: (a) Precision-recall curves of SD (Feng et al., 2011) (in cyan), bMCL (Zhu et al., 2012) (in green), NIM-SD (in blue) and ADMM (in red) in the task of object discovery in FDDB subset (Jain & Learned-Miller, 2010) and (b) ETHZ Apple logo class (Ferrari et al., 2006). (To view this figure in color see the online supplement, available at http://www.mitpressjournals.org/doi/suppl/10.1162/NECO_a_00555.)

Figure 7: Face discovery results on the FDDB subset (Jain & Learned-Miller, 2010). The patches with the maximum score given by SD (Feng et al., 2011), bMCL (Zhu et al., 2012), NIM-SD (in blue) and ADMM are plotted in cyan, green, blue, and red, respectively. (To view this figure in color see the online supplement, available at http://www.mitpressjournals.org/doi/suppl/10.1162/NECO_a_00555.)

windows, and the objects observe a large variation due to articulation or nonrigid transformation, which do not reside in a common low-rank space. Note that in this letter, we focus on the problem of subspace learning and make the assumption of the common pattern spanning a low-rank subspace.

4.4 Instance Selection for Multiple Instance Learning. In this experiment, we show how to apply the proposed ADMM to the traditional MIL problem (Dietterich et al., 1997). Our basic idea is to use ADMM to directly

distinguish the positive instances from the negative instances in positive bags; the positive instances together with all the negative instances from negative bags are used to train an instance-level classifier (e.g. SVM with RBF kernel for the MIL task). In the testing stage, we use the learned instance-level SVM classifier for bag classification, based on a noisy-or model: if there exists any positive instance in a bag, the bag is identified as being positive, and otherwise as negative.

To use ADMM to distinguish positive from negative instances, we follow the assumption that has been previously made in this letter: positive instances lie in a low-dimensional subspace. In practice, we collect all positive bags as input of ADMM algorithm in algorithm 1 and obtain the indicator value of each instance. For each bag, the indicator value is normalized by dividing the maximum indicator value in the bag. Then the instances whose normalized indicator values are larger than an upper threshold τ_u are labeled as positive instances; the instances whose normalized indicator values are less than a lower threshold τ_l are labeled as negative instances. In this experiment, we fix $\tau_u = 0.7$ and $\tau_l = 0.3$. The instances with normalized indicator values between 0.3 and 0.7 are omitted and not used for training the instance SVM classifier. When training the RBF kernel SVM, we adopt the LibSVM (Chang & Lin, 2011).

We evaluate the proposed method on five popular bench mark datasets, including Musk1, Musk2, Elephant, Fox, and Tiger. Detailed descriptions of the data sets can be found in Dietterich and Lathrop (1997) and Andrews et al. (2003). We compare our method with MI-SVM and mi-SVM (Andrews et al., 2003), MILES (Chen et al., 2006), EM-DD (Zhang & Goldman, 2001), PPMM Kernel (Wang et al., 2008), MIGraph and miGraph (Zhou et al., 2009), and MI-CRF (Deselaers & Ferrari, 2010) via 10 times 10-fold cross-validation and report the average accuracy and the standard deviation in Table 3. Some of them were obtained in different studies, and the standard deviations were not available. The average accuracy over the five tested data sets is reported in the for right column. The best performance on each compared item is noted in bold.

As shown in Table 3, the best results are reported by MIGraph and miGraph, which exploit graph structure based on the affinities. We focus on comparing with mi-SVM, which learns to weigh instances by maximizing the margin between the positive and negative instances under MIL conditions via iterative SVM. This problem is nonconvex, and the optimization method of mi-SVM does not guarantee a local optimum. Here, our method selects the instance of a common subspace with a convex formulation and obtains promising results.

5 Conclusion

In this letter, we have proposed a robust formulation for unsupervised subspace discovery. We relax the highly combinatorial high-dimensional

Table 3: Performance Comparison with Per Class and Average Bag Classification Accuracies (%).

Data Sets	Musk1	Musk2	Elephant	Fox	Tiger	Average
MI-SVM	77.9	84.3	81.4	59.4	84.0	77.4
mi-SVM	87.4	83.6	82.0	58.2	78.9	78.0
MILES	86.3	87.7	–	–	–	–
EM-DD	84.8	84.9	78.3	56.1	72.1	75.2
PPMM kernel	**95.6**	81.2	82.4	60.3	80.2	79.9
MI-CRF	87.0	78.4	85.0	65.0	79.5	79.0
ADMM (our method)	89.9±0.7	85.0±1.6	79.6±0.9	**65.4±1.2**	81.5±1.0	80.3
MIGraph	90.0±3.8	90.0±2.7	85.1±2.8	61.2±1.7	81.9±1.5	81.6
miGraph	88.9±3.3	**90.3±2.6**	**86.8±0.7**	61.6±2.8	**86.0±1.6**	**82.7**

Notes: Comparisons on five benchmark data sets. MI-SVM and mi-SVM in Andrews, Tsochantaridis, and Hofmann (2003); MILES in Chen, Bi, and Wang, 2006; EM-DD in Zhang and Goldman, 2001; PPMM Kernel in Wang, Yang, and Zha, 2008; MIGraph and miGraph in Zhou, Sun, and Li, 2009; and MI-CRF in Deselaers and Ferrari, 2010. ADMM (our method) refers to the one proposed in this letter. The numbers in bold indicate the best results in each column.

problem into a convex program and solve it efficiently with the augmented Lagrangian multiplier method. Unlike other approaches based on discriminative training, our proposed method can discover objects of interest by using common patterns across input data. We demonstrate the evident advantage of our method over the competing algorithms in a variety of benchmark data sets. Our method suggests that an explicit low-rank subspace assumption with a robust formulation naturally deals with a subspace discovery problem in the presence of overwhelming outliers, which allows a rich emerging family of subspace learning methods to have a wider scope of applications. It enlarges the application range of the RPCA-based methods.

Acknowledgments

This work was supported by Microsoft Research Asia, NSF IIS-1216528 (IIS-1360566), NSF CAREER award IIS-0844566 (IIS-1360568), NSFC 61173120, NSFC 61222308, and the Chinese Program for New Century Excellent Talents in University. X.W. was supported by Microsoft Research Asia Fellowship 2012. We thank John Wright for encouraging discussions, David Wipf for valuable comments, and Jun Sun for his helpful discussion on the proof of theorem 1.

References

Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(12), 2037–2041.

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2003). Support vector machines for multiple-instance learning. In S. Becker, S. Thrün, & K. Obermayer (Eds.), *Advances in neural information processing systems, 15* (pp. 561–568). Cambridge, MA: MIT Press.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122.

Candes, E., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM*, *58*(3), 1–37.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27.

Chen, Y., Bi, J., & Wang, J. Z. (2006). Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(12), 1931–1947.

Chum, O., & Zisserman, A., 2007. An exemplar model for learning object classes. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). Piscataway, NJ: IEEE.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Series B*, *39*(1), 1–38.

Deselaers, T., Alexe, B., & Ferrari, V. (2012). Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, *100*(3), 275–293.

Deselaers, T., & Ferrari, V. (2010). A conditional random field for multiple-instance learning. In *Proceedings of the 26th International Conference on Machine Learning* (pp. 287–294). Madison, WI: Omnipress.

Dietterich, T. G., & Lathrop, R. H. (1997). Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, *89*, 31–71.

Dietterich, T., Lathrop, R., & Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, *89*(1–2), 31–71.

Elhamifar, E., & Vidal, R. (2009). Sparse subspace clustering. In *IEEE Conference on Computer Vision and pattern recognition* (pp. 2790–2797). Piscataway, NJ: IEEE.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., & Zisserman, A. (2007). *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) results.* Retrieved from http://www.pascal-network.org /challenges/ VOC/voc2007/workshop/index .html

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., & Zisserman, A. (2011). *The PASCAL Visual Object Classes Challenge (VOC) results.* Retrieved from http://www.pascal-network.org /challenges/ VOC/voc2011/workshop/index .html

Everingham, M., Zisserman, A., Williams, C.K.I., & Van Gool, L. (2006). The PASCAL Visual Object Classes Challenge 2006 (VOC2006) results. Retrieved from http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf

Favaro, P., Vidal, R., & Ravichandran, A. (2011). A closed form solution to robust subspace estimation and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1801–1807). Piscataway, NJ: IEEE.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1627–1645.

Feng, J., Wei, Y., Tao, L., Zhang, C., & Sun, J. (2011). Salient object detection by composition. In *Proceedings of the International Conference on Computer Vision* (pp. 1028–1035).

Ferrari, V., Tuytelaars, T., & Van Gool, L. (2006). Object detection by contour segment networks. In *Proceedings of the European Conference on Computer Vision* (pp. 14–28). New York: Springer.

Gabay, D., & Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers and Mathematics with Applications*, *2*(1), 17–40.

Georghiades, A., Belhumeur, P., & Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions Pattern Analysis and Machine Intelligence*, *23*(6), 643–660.

Grauman, K., & Darrell, T. (2006). Unsupervised learning of categories from sets of partially matching image features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 19–25). Piscataway, NJ: IEEE.

Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge: Cambridge University Press.

Jain, V., & Learned-Miller, E. (2010). *FDDB: A benchmark for face detection in unconstrained settings* (Tech. Rep. No. UM-CS-2010-009). Amherst: University of Massachusetts.

Jenatton, R., Obozinski, G., & Bach, F. (2010). Structured sparse principal component analysis. In *Proceedings of the International Conference on Artificial Intelligence and Statistics* (pp. 366–373). Cambridge, MA: MIT Press.

Jia, K., Chan, T.-H., & Ma, Y. (2012). Robust and practical face recognition via structured sparsity. In *Proceedings of the 12th European Conference on Computer Vision* (pp. 331–344). New York: Springer.

Lampert, C., Blaschko, M., & Hofmann, T. (2009). Efficient subwindow search: A branch and bound framework for object localization. *IEEE Transactions Pattern Analysis and Machine Intelligence*, *31*, 2129–2142.

Lee, Y., & Grauman, K. (2009). Foreground focus: Unsupervised learning from partially matching images. *International Journal of Computer Vision*, *85*(2), 143–166.

Lerman, G., McCoy, M. B., Tropp, J. A., & Zhang, T. (2012). *Robust computation of linear models, or how to find a needle in a haystack*. CoRR, abs/1202.4044.

Lin, Z., Chen, M., & Ma, Y. (2010). *The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices*. Arxiv preprint arXiv:1009.5055.

Liu, G., Lin, Z., & Yu, Y. (2010). Robust subspace segmentation by low-rank representation. In *Proceedings of the 26th International Conference on Machine Learning* (pp. 663–670). Madison, WI: Omnipress.

Luo, D., Nie, F., Ding, C., & Huang, H. (2011). Multi-subspace representation and discovery. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 405–420). New York: Springer.

Pandey, M., & Lazebnik, S. (2011). Scene recognition and weakly supervised object localization with deformable part-based models. In *IEEE International Conference on Computer Vision* (pp. 1307–1314). Piscataway, NJ: IEEE.

Peng, Y., Ganesh, A., Wright, J., Xu, W., & Ma, Y. (2012). RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*, 2233–2246.

Rahmani, R., Goldman, S. A., Zhang, H., Krettek, J., & Fritts, J. E. (2005). Localized content based image retrieval. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval* (pp. 227–236). New York: ACM.

Russell, B., Freeman, W., Efros, A., Sivic, J., & Zisserman, A. (2006). Using multiple segmentations to discover objects and their extent in image collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1605–1614).

Sankaranarayanan, K., & Davis, J. (2012). One-class multiple instance learning and applications to target tracking. In *Proceedings of the Asian Conference on Computer Vision* (pp. 126–139). New York: Springer.

Shiqian Ma, L. X., & Zou, H. (2013). Alternating direction methods for latent variable gaussian graphical model selection. *Neural Computation*, *25*(8), 2172–2198.

Tao, M., & Yuan, X. (2011). Recovering low-rank and sparse components of matrices from incomplete and noisy observations. *SIAM Journal on Optimization*, *21*(1), 57–81.

Wagner, A., Wright, J., Ganesh, A., Zhou, Z., & Ma, Y. (2009). Towards a practical face recognition system: Robust registration and illumination via sparse representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 597–604). Piscataway, NJ: IEEE.

Wang, H.-Y., Yang, Q., & Zha, H. (2008). Adaptive p-posterior mixture-model kernels for multiple instance learning. In *Proceedings of the 25th Annual International Conference on Machine Learning* (pp. 1136–1143). New York: ACM.

Wang, X., Zhang, Z., Ma, Y., Bai, X., Liu, W., & Tu, Z. (2012). One-class multiple instance learning via robust PCA for common object discovery. In *Proceedings of the Asian Conference on Computer Vision* (pp. 246–258). New York: Springer.

Wright, J., & Ma, Y. (2010). Dense error correction via ℓ^1-minimization. *IEEE Transactions on Information Theory*, *56*(7), 3540–3560.

Wright, J., Yang, A., Ganesh, A., Sastry, S., & Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions Pattern Analysis and Machine Intelligence*, *31*(2), 210–227.

Xu, H., Sanghavi, S., & Caramanis, C. (2012). Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, *58*(5), 3047–3064.

Yu, C., & Joachims, T. (2009). Learning structural SVMS with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1169–1176). New York: ACM.

Zhang, Q., & Goldman, S. A. (2001). EM-DD: An improved multiple-instance learning technique. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* (pp. 1073–1080). Cambridge, MA: MIT Press.

Zhang, Z., Ganesh, A., Liang, X., & Ma, Y. (2012). Tilt: Transform invariant low-rank textures. *International Journal of Computer Vision*, *99*(1), 1–24.

Zhou, Z.-H., Sun, Y.-Y., & Li, Y-F. (2009). Multi-instance learning by treating instances as non-IID samples. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1249–1256). New York: ACM.

Zhu, J., Wu, J., Wei, Y., Chang, E., & Tu, Z. (2012). Unsupervised object class discovery via saliency-guided multiple class learning. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3218–3225). Piscataway, NJ: IEEE.

Received March 3, 2013; accepted September 24, 2013.