Elastic-Net Regularization of Singular Values for Robust Subspace Learning

Eunwoo Kim[†] Minsik Lee[‡] Songhwai Oh[†]
[†]Department of ECE, ASRI, Seoul National University, Seoul, Korea
[‡]Division of EE, Hanyang University, Korea
kewoo15@snu.ac.kr mleepaper@hanyang.ac.kr songhwai@snu.ac.kr

Abstract

Learning a low-dimensional structure plays an important role in computer vision. Recently, a new family of methods, such as l_1 minimization and robust principal component analysis, has been proposed for low-rank matrix approximation problems and shown to be robust against outliers and missing data. But these methods often require heavy computational load and can fail to find a solution when highly corrupted data are presented. In this paper, an elastic-net regularization based low-rank matrix factorization method for subspace learning is proposed. The proposed method finds a robust solution efficiently by enforcing a strong convex constraint to improve the algorithm's stability while maintaining the low-rank property of the solution. It is shown that any stationary point of the proposed algorithm satisfies the Karush-Kuhn-Tucker optimality conditions. The proposed method is applied to a number of low-rank matrix approximation problems to demonstrate its efficiency in the presence of heavy corruptions and to show its effectiveness and robustness compared to the existing methods.

1. Introduction

Low-rank matrix approximation has attracted much attention in the areas of data reconstruction [10, 12, 32], image denoising [7, 13, 23], collaborative filtering [25], background modeling [8, 24, 35], structure from motion [3, 13, 31, 34], and photometric stereo [4, 20, 33], to name a few. It is sometimes assumed that the rank of a matrix is fixed or known beforehand and it is also known as a subspace learning problem.

Although real-world data are usually high dimensional, they can be well-represented with fewer parameters in many cases. Hence, reducing the data dimension to a dominating principal structure is desirable to reduce the computation time and also to remove unwanted noisy components. A popular method for addressing this issue is principal component analysis (PCA) [11]. PCA transforms data to a low-dimensional subspace which maximizes the variance of the data based on the l_2-norm. To handle missing data, [25] solved a weighted low-rank matrix approximation problem using the expectation-maximization algorithm. Lin [16] proposed projected gradient methods to solve nonnegative matrix factorization problems for image and text data. Mitra *et al*. [20] presented a matrix factorization technique which adds regularization terms to prevent data overfitting and solves the problem using semi-definite programming. These conventional l_2-norm based approximation methods have been utilized in many problems but it is known that they are sensitive to outliers and corruptions because the l_2-norm amplifies the negative effects of corrupted data.

As an alternative, low-rank matrix approximation methods using the l_1-norm have been proposed for robustness against outliers and missing data [7, 12–14, 21]. In addition, there have been several probabilistic extensions of low-rank matrix factorization for robust approximation [6, 19, 32]. Ke and Kanade [12] presented a low-rank matrix approximation method by alternatively minimizing an l_1-norm based cost function using convex programming. Eriksson and Hengel [7] proposed the l_1-Wiberg method for weighted low-rank matrix approximation in the presence of missing data. Kwak [14] proposed an l_1 maximization approach to find successive principal components using a greedy approach in the feature space. Kim *et al*. [13] proposed two gradient-based approaches for an l_1 minimization problem using a rectified representation.

Recently, many efficient approaches using augmented Lagrangian method have been proposed to solve the l_1 minimization problem [4, 23, 34]. Shen *et al*. [23] proposed a low-rank matrix approximation method using the l_1-norm based on the augmented Lagrangian alternating direction method (ALADM). Zheng *et al*. [34] proposed a practical weighted low-rank approximation method using nuclear-norm regularized l_1 cost function and orthogonality constraint (Regl_1-ALM). Cabral *et al*. [4] proposed a unifying approach, which combines nuclear-norm minimization and bilinear factorization using an alternative definition of the nuclear-norm [22]. Their approach leads to an efficient

algorithm without high computational complexity. These methods have been successfully applied to low-rank factorization problems in the presence of missing data and outliers, outperforming other rank minimization methods [17, 33]. But, it is difficult for factorization methods to find the global optimal solution since the problem is non-convex.

There is another family of approaches based on the recent advances in nuclear-norm minimization, which is called robust principal component analysis (RPCA), and it has been successfully applied to a number of problems [5, 17, 29]. RPCA tries to find a solution for a non-fixed rank matrix approximation problem using the l_1-norm regularized nuclear-norm cost function and solves using various approaches such as augmented Lagrangian method (ALM) [5, 17] and accelerated proximal gradient (APG) [29]. It has been shown that RPCA is suitable for problems, such as shadow removing and background modeling [5]. However, algorithms proposed for RPCA have high computational complexity, especially for large-scale problems, because they perform singular value decomposition (SVD) at each iteration. Recently, Shu *et al*. [24] proposed efficient low-rank recovery methods using a new rank measure. But, the above methods sometimes find a suboptimal solution under heavy corruptions, which remains a difficult problem in practice.

In this paper, we present an efficient low-rank matrix factorization method based on elastic-net regularization for robust subspace learning problems in the presence of heavy corruptions, including both outliers and missing entries. Our method is a holistic approach which utilizes both nuclear-norm minimization and bilinear factorization. To prevent the instability of the algorithm, which may arise from highly corrupted data, we introduce elastic-net regularization for singular values to introduce strong convexity to a lasso-type nuclear-norm minimization problem. The strong convexity of the proposed method alleviates the instability problem by shrinking and correcting inaccurate singular values in the presence of unwanted noises. We also show that any limit point of the proposed algorithm satisfies necessary conditions to be a local optimal solution. We demonstrate the performance of the proposed method in terms of the reconstruction error and computational speed using well-known benchmark datasets including non-rigid motion estimation, photometric stereo, and background modeling.

2. The Proposed Method

2.1. Problem formulation

In this paper, we consider the low-rank matrix and sparse matrix separation problem [4, 34] based on convex envelopes of rank and sparsity functions as follows:

$$\min_{P,X} f_1(P,X) + \lambda \|PX\|_*, \tag{1}$$

where $f_1(P,X) = \|W \odot (Y - PX)\|_1$, Y is an observation matrix, and λ is a pre-defined weighting parameter. $\|\cdot\|_1$ and $\|\cdot\|_*$ denote the entry-wise l_1-norm and the nuclear-norm, which are convex relaxation of the l_0-norm and the rank function, respectively. Here, $\odot$ is the component-wise multiplication or the Hadamard product and W is a weighting matrix, whose element w_{ij} is 1 if y_{ij} is known, and 0 if y_{ij} is unknown. The problem is similar to RPCA [5, 17] if a low-rank matrix D and a sparse error matrix E replace PX and $Y - PX$, respectively. Generally, (1) is a non-convex and non-smooth problem, making it difficult to find a solution efficiently and exactly. To solve the problem efficiently, a common strategy is to use an alternating minimization approach which solves for one variable while other variables are fixed [10, 12, 35].

Notice that the regularization term in (1), $\|PX\|_*$, can be interpreted as a sum of singular values, $\sum_i^r |\sigma_i|$, where σ_i is the ith singular value of a low-rank matrix PX and r is the rank of PX. It leads to a lasso problem [15, 28], which has a thresholding effect over singular values. But, lasso-based approaches lack a shrinkage effect due to their weak convexity, which makes the algorithm unstable when highly corrupted data are presented. To improve the stability of the algorithm, we introduce a strong convex regularizer over singular values with the l_2-norm penalty of singular values, $\lambda_1 \sum_i^r |\sigma_i| + \frac{\lambda_2}{2} \sum_i^r |\sigma_i|^2$. Based on the fact that $\|D\|_F^2 = \mathrm{tr}(V\Sigma U^T U \Sigma V^T) = \mathrm{tr}(\Sigma^2) = \sum_i |\sigma_i|^2$, where $D = U\Sigma V^T$ is SVD of D, we introduce a new penalized optimization problem as follows:

$$\min_{P,X} f_1(P,X) + \lambda_1 \|PX\|_* + \frac{\lambda_2}{2} \|PX\|_F^2. \tag{2}$$

In (2), we have elastic-net regularization of singular values, which has shown its superiority compared to lasso [15, 28] in many applications [9, 15, 36]. It is capable of stabilizing a lasso-type method due to its strong convexity, owing to the Frobenius norm [9, 26, 36]. In addition, we have both a thresholding effect over singular values from the l_1 regularizer and a shrinkage effect from the l_2 regularizer to make a parsimonious and stable model.

Note that, without these regularization terms, the problem (2) can be solved using the augmented Lagrangian alternating direction method (ALADM) [23]. There is another approach using a nuclear-norm regularized l_1-norm cost function [34]. It is extended using an alternative definition of the nuclear-norm (Unifying[1]) [4], which does not contain the smoothness term given in (2). However, these methods can find a suboptimal solution since these alternating minimization based approaches without a proper cor-

[1] We call the method in [4] as Unifying for simplicity.

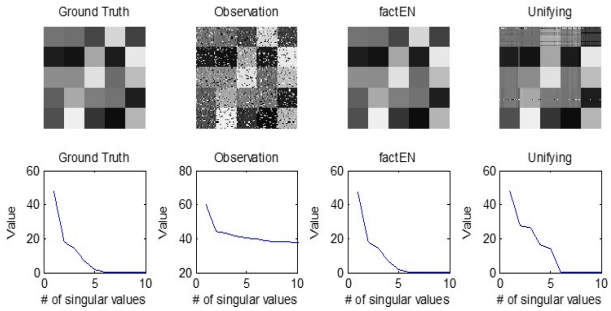


Figure 1. Evaluation of the proposed method (factEN) and a lasso-based method (Unifying [4]) for a toy example.

rection term may lead to a poor solution in the presence of highly corrupted data (see Section 3.1).

Figure 1 shows results of the proposed method compared to Unifying [4], a lasso-based method, and ground-truth on a simple example (100×100) with 20% outliers. The rank of the ground-truth is five. From the figure, the proposed method gives a stable result against outliers and eliminates noises by suppressing the singular values, whereas Unifying finds relatively inaccurate and higher singular values and shows a poor reconstruction result compared to the proposed method and ground-truth.

Unfortunately, (2) can suffer from heavy computational complexity for large-scale problems because the problem is solved by performing SVD at each iteration which is used for solving a nuclear-norm based cost function. To solve (2) in practice, the following property of the nuclear-norm can be utilized [18].

Lemma 1 ([18]). *For any matrix $D \in \mathbb{R}^{m\times n}$, the following holds:*

$$\|D\|_* = \min_{P,X} \frac{1}{2}\left(\|P\|_F^2 + \|X\|_F^2\right) \;\; s.t. \;\; D = PX. \quad (3)$$

If the rank of D is $r \leq \min(m,n)$, then the minimum solution above is attained at a factor decomposition $D = P_{m\times r}X_{r\times n}$.

Using Lemma 1, we make an equivalent form of (2) as follows:

$$\min_{P,X,D} f_2(D) + \frac{\lambda_1}{2}\left(\|P\|_F^2 + \|X\|_F^2\right) + \frac{\lambda_2}{2}\|D\|_F^2, \quad (4)$$

such that $D = PX$, where $f_2(D) = \|W \odot (Y - D)\|_1$. Due to the difficulty of solving the problem (4) in practice, we introduce an auxiliary variable $\widehat{D}$ and solve the following problem instead.

$$\begin{aligned}\min_{P,X,D,\widehat{D}} f_2(\widehat{D}) + \frac{\lambda_1}{2}\left(\|P\|_F^2 + \|X\|_F^2\right) + \frac{\lambda_2}{2}\|D\|_F^2 \\ \text{s.t.} \;\; D = PX, \;\; \widehat{D} = D.\end{aligned} \quad (5)$$

To solve (5), we utilize the augmented Lagrangian framework which converts (5) into the following unconstrained problem:

$$\begin{aligned}\mathcal{L}(P,X,D,\widehat{D},\Lambda_1,\Lambda_2) &= f_2(\widehat{D}) + \frac{\lambda_1}{2}\left(\|P\|_F^2 + \|X\|_F^2\right) \\ &+ \frac{\lambda_2}{2}\|D\|_F^2 + \text{tr}\left(\Lambda_1^T(D - PX)\right) + \text{tr}\left(\Lambda_2^T(\widehat{D} - D)\right) \\ &+ \frac{\beta}{2}\left(\|D - PX\|_F^2 + \|\widehat{D} - D\|_F^2\right),\end{aligned} \quad (6)$$

where $\Lambda_1, \Lambda_2 \in \mathbb{R}^{m\times n}$ are Lagrange multipliers and $\beta > 0$ is a small penalty parameter.

2.2. Algorithm

Based on the previous formulation, we develop a method based on the augmented Lagrangian framework and solve it using an alternating minimization technique [10,23,35]. To solve for P, we fix the other variables and solve the following optimization problem:

$$\begin{aligned}P_+ = \arg\min_P \; &\frac{\lambda_1}{2}\|P\|_F^2 + \text{tr}\left(\Lambda_1^T(D - PX)\right) \\ &+ \frac{\beta}{2}\|D - PX\|_F^2.\end{aligned} \quad (7)$$

This optimization problem is a least square problem and the solution is

$$P_+ = (\Lambda_1 + \beta D)X^T(\lambda_1 I + \beta XX^T)^{-1}, \quad (8)$$

where I denotes an identity matrix. For X, we solve the following optimization problem:

$$\begin{aligned}X_+ = \arg\min_X \; &\frac{\lambda_1}{2}\|X\|_F^2 + \text{tr}\left(\Lambda_1^T(D - PX)\right) \\ &+ \frac{\beta}{2}\|D - PX\|_F^2,\end{aligned} \quad (9)$$

which can be solved similar to (7) and its solution is

$$X_+ = (\lambda_1 I + \beta P^T P)^{-1}P^T(\Lambda_1 + \beta D). \quad (10)$$

For finding D, we consider the following optimization problem:

$$\begin{aligned}D_+ = \arg\min_D \; &\frac{\lambda_2}{2}||D||_F^2 \\ &+ \text{tr}\left(\Lambda_1^T(D - PX)\right) + \text{tr}\left(\Lambda_2^T(\widehat{D} - D)\right) \\ &+ \frac{\beta}{2}\left(\|D - PX\|_F^2 + \|\widehat{D} - D\|_F^2\right),\end{aligned} \quad (11)$$

and its solution is

$$D_+ = \frac{\beta PX + \beta\widehat{D} + \Lambda_2 - \Lambda_1}{\lambda_2 + 2\beta}. \quad (12)$$

Algorithm 1 factEN by ALM for optimizing (5)

1: **Input:** $Y \in \mathbb{R}^{m\times n}$, r, β, ρ, and $\lambda_1, \lambda_2 = 10^{-3}$
2: **Output:** $P \in \mathbb{R}^{m\times r}$, $X \in \mathbb{R}^{r\times n}$, and $D \in \mathbb{R}^{m\times n}$
3: **while** not converged **do**
4: **while** not converged **do**
5: Update P using (8)
6: Update X using (10)
7: Update D using (12)
8: Update $\widehat{D}$ using (14)
9: **end while**
10: Update the Lagrange multipliers Λ_1, Λ_2 using (15)
11: $\beta = \min(\rho\beta, 10^{20})$
12: **end while**

We obtain the following equation to solve for $\widehat{D}$,

$$\widehat{D} = \arg\min_{\widehat{D}} f_2(\widehat{D}) + \mathrm{tr}\left(\Lambda_2^T(\widehat{D} - D)\right) + \frac{\beta}{2}\|\widehat{D} - D\|_F^2 \tag{13}$$

and the solution can be computed using the absolute value thresholding operator [5, 17, 34]:

$$\begin{cases} W \odot \widehat{D}_+ \leftarrow W \odot \left(Y - \mathcal{S}\left(Y - D + \frac{\Lambda_2}{\beta}, \frac{1}{\beta}\right)\right), \\ \overline{W} \odot \widehat{D}_+ \leftarrow \overline{W} \odot \left(D - \frac{\Lambda_2}{\beta}\right), \end{cases} \tag{14}$$

where $\mathcal{S}(x, \tau) = \mathrm{sign}(x)\max(|x| - \tau, 0)$ for a variable x and $\overline{W} \in \mathbb{R}^{m\times n}$ is a complementary matrix of W whose element $\overline{w}_{ij}$ is 0 if y_{ij} is known, and is 1 if y_{ij} is unknown.

Finally, we update the Lagrange multipliers as

$$\begin{aligned} \Lambda_1 &= \Lambda_1 + \beta(D - PX), \\ \Lambda_2 &= \Lambda_2 + \beta(\widehat{D} - D). \end{aligned} \tag{15}$$

Based on the previous analysis, we derive a robust elastic-net regularized low-rank matrix factorization algorithm and it is summarized in Algorithm 1. Since the algorithm is constructed based on elastic-net regularization and solved using a matrix factorization approach, the proposed method is named as *factEN*. In the algorithm, we have assumed a normalized observation matrix. Hence, the output matrices P and X can be obtained by re-scaling them using the scaling factor. We initialize the optimization variables with the Gaussian distribution $\mathcal{N}(0, 10^{-3})$.[2]

The computational complexity of the inner loop (line 4–9 in Algorithm 1) is $O(mnr)$ for the proposed method, which is the same as Unifying [4] and ALADM [23]. Since IALM [17] and Regl_1-ALM [34] perform an SVD operation at each iteration, their computational complexities are $O(\min(m, n)\max(m, n)^2)$ and $O(r\max(m, n)^2)$, respectively, requiring more computational efforts than factEN. Note that the proposed method can be easily extended to speed up the algorithm with linear complexity at each iteration by sampling sub-matrices from a measurement matrix as described in [24, 27].

[2]Note that we have empirically found that our algorithm is not sensitive to initial values and finds similar solutions with different initial values.

2.3. Convergence analysis

In this section, we analyze the convergence property of the proposed method. Although it is difficult to guarantee its convergence to a local minimum, an empirical evidence suggests that the proposed algorithm has a strong convergence behavior (see Figure 2). Nevertheless, we provide a proof of weak convergence of factEN by showing that under mild conditions any limit point of the iteration sequence generated by the algorithm is a stationary point that satisfies the Karush-Kuhn-Tucker (KKT) conditions [2]. It is worth proving that any converging point must be a point that satisfies the KKT conditions because they are necessary conditions to be a local optimal solution. This result provides an assurance about the behavior of the proposed algorithm.

We rewrite the cost function of factEN by assuming the fully-observed data model of (5), i.e., $W_{ij} = 1$ for all i, j, as follows:

$$\begin{aligned} \min_{P,X,D,\widehat{D}} & f_3(\widehat{D}) + \frac{\lambda_1}{2}\left(\|P\|_F^2 + \|X\|_F^2\right) + \lambda_2\|D\|_F^2 \\ \text{s.t.} \quad & D = PX,\ \widehat{D} = D. \end{aligned} \tag{16}$$

where $f_3(\widehat{D}) = \|Y - \widehat{D}\|_1$. However, a similar result can be derived for the partially-observed data model.

Let us assume that the proposed algorithm reaches a stationary point. The KKT conditions for (16) are derived as follows:

$$\begin{aligned} & D - PX = 0,\ \widehat{D} - D = 0,\ \frac{\partial\mathcal{L}}{\partial P} = \lambda_1 P - \Lambda_1 X^T = 0, \\ & \frac{\partial\mathcal{L}}{\partial X} = \lambda_1 X - P^T\Lambda_1 = 0,\ \frac{\partial\mathcal{L}}{\partial D} = \lambda_2 D + \Lambda_1 - \Lambda_2 = 0, \\ & \Lambda_2 \in -\partial_{\widehat{D}}(||Y - \widehat{D}||_1). \end{aligned} \tag{17}$$

Here, we can obtain the following equation from the the last relationship in (17):

$$\begin{aligned} Y - D + \frac{\Lambda_2}{\beta} &\in Y - D - \frac{1}{\beta}\partial_{\widehat{D}}(||Y - \widehat{D}||_1) \\ &= Y - \widehat{D} - \frac{1}{\beta}\partial_{\widehat{D}}(||Y - \widehat{D}||_1) \triangleq Q_\beta(Y - \widehat{D}), \end{aligned} \tag{18}$$

where scalar function $Q_\beta(t) \triangleq t - \frac{1}{\beta}\partial|t|$ is applied element-wise to $Y - \widehat{D}$. From [23], we can obtain the following

relation:

$$Y - \widehat{D} = Q_\beta^{-1}\left(Y - D + \frac{\Lambda_2}{\beta}\right) \equiv \mathcal{S}\left(Y - D + \frac{\Lambda_2}{\beta}, \frac{1}{\beta}\right), \tag{19}$$

where $\mathcal{S}(x,\tau) = \text{sign}(x)\max(|x| - \tau, 0)$. Therefore, the KKT conditions can be rewritten as follows:

$$\begin{aligned} &D - PX = 0, \;\; \widehat{D} - D = 0, \;\; \lambda_1 P - \Lambda_1 X^T = 0, \\ &\lambda_1 X - P^T\Lambda_1 = 0, \;\; \lambda_2 D + \Lambda_1 - \Lambda_2 = 0, \\ &Y - \widehat{D} = \mathcal{S}\left(Y - D + \frac{\Lambda_2}{\beta}, \frac{1}{\beta}\right). \end{aligned} \tag{20}$$

Based on these conditions, we prove the convergence to a point which satisfies the KKT conditions.

Theorem 1. *Let $G \triangleq (P, X, D, \widehat{D}, \Lambda_1, \Lambda_2)$ and $\{G^j\}_{j=1}^\infty$ be generated by factEN. Assume that $\{G^j\}_{j=1}^\infty$ is bounded and $\lim_{j\to\infty}\{G^{j+1} - G^j\} = 0$. Then, any accumulation point of $\{G^j\}_{j=1}^\infty$ satisfies the KKT conditions. In particular, whenever $\{G^j\}_{j=1}^\infty$ converges, it converges to a KKT point.*

Proof. First, we get the Lagrange multipliers Λ_1, Λ_2 from the algorithm as

$$\begin{aligned} \Lambda_{1+} &= \Lambda_1 + \beta(D - PX) \\ \Lambda_{2+} &= \Lambda_2 + \beta(\widehat{D} - D), \end{aligned} \tag{21}$$

where Λ_{i+} is a next point of Λ_i in a sequence $\{\Lambda_i^j\}_{j=1}^\infty$. If sequences of variables $\{\Lambda_1^j\}_{j=1}^\infty$ and $\{\Lambda_2^j\}_{j=1}^\infty$ converge to a stationary point, i.e., $(\Lambda_{1+} - \Lambda_1) \to 0$ and $(\Lambda_{2+} - \Lambda_2) \to 0$, then $(D - PX) \to 0$ and $(\widehat{D} - D) \to 0$, respectively. This satisfies the first two conditions of the KKT conditions.

Second, from P_+ derived in the algorithm, we get

$$P_+ - P = (\Lambda_1 + \beta D)X^T(\lambda_1 I + \beta XX^T)^{-1} - P, \tag{22}$$

where I denotes an identity matrix and it can be rewritten by multiplying $(\lambda_1 I + \beta XX^T)$ to both sides in (22) as

$$\begin{aligned} &(P_+ - P)(\lambda_1 I + \beta XX^T) \\ &\qquad = \Lambda_1 X^T - \lambda_1 P + \beta(D - PX)X^T. \end{aligned} \tag{23}$$

From the first condition, we can derive $\Lambda_1 X^T - \lambda_1 P \to 0$ when $(P_+ - P) \to 0$.

Third, using $X_+ = (\lambda_1 I + \beta P^T P)^{-1} P^T(\Lambda_1 + \beta D)$ derived from the algorithm, we can obtain the following:

$$\begin{aligned} &(\lambda_1 I + \beta P^T P)(X_+ - X) \\ &\qquad = P^T\Lambda_1 - \lambda_1 X + \beta P^T(D - PX). \end{aligned} \tag{24}$$

If $(X_+ - X) \to 0$, then $(P^T\Lambda_1 - \lambda_1 X) \to 0$ as well.

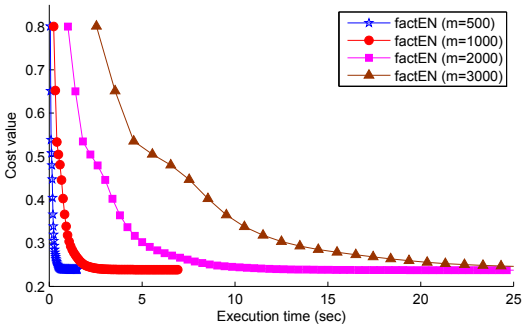


Figure 2. Scaled cost values of the proposed algorithm at each iteration for four synthetic examples.

Likewise, we can get the following equation using D_+ from the proposed algorithm,

$$\begin{aligned} &(\lambda_2 + 2\beta)(D_+ - D) \\ &\quad = \beta(PX - D + \widehat{D} - D) - \Lambda_1 + \Lambda_2 - \lambda_2 D. \end{aligned} \tag{25}$$

Since $PX - D$ and $\widehat{D} - D$ converge to zero from the previous analysis, we obtain $\Lambda_1 - \Lambda_2 + \lambda_2 D = 0$ whenever $D_+ - D \to 0$.

Lastly, from (19), we obtain the following equation:

$$\widehat{D}_+ - \widehat{D} = Y - \mathcal{S}\left(Y - D + \frac{\Lambda_2}{\beta}, \beta\right) - D. \tag{26}$$

Since $\{G^j\}_{j=1}^\infty$ is bounded by our assumption, $\{X_+X_+^T\}_{j=1}^\infty$ and $\{P_+^T P_+\}_{j=1}^\infty$ in (23) and (25) are bounded. Hence, $\lim_{j\to\infty}(G^{j+1} - G^j) = 0$ implies that both side of the above equations (21), (23), (24), (25), and (26) tend to zero as $j \to \infty$. Therefore, the sequence $\{G^j\}_{j=1}^\infty$ asymptotically satisfies the KKT condition for (16). This completes the proof. □

In our algorithm, we set the stopping criterion as

$$\frac{||D^{(t)} - P^{(t)}X^{(t)}||_1}{||Y||_1} < \theta, \tag{27}$$

where t is the number of iterations and θ is a small positive number. Here, we compute the whole elements of D including elements corresponding to the unknown entries. Since it is enough for the algorithm to achieve a nearly stationary point when the difference between the terminating cost of adjacent iterations becomes small, we set the stopping condition as $\theta = 10^{-5}$ in our experiments in Section 3. Figure 2 shows scaled cost values[3] of the proposed method at each iteration for four examples from 500 × 500 to 3,000 × 3,000 with outliers as described in Section 3.1. Each point

[3]We have scaled cost values as $(f_2(\widehat{D}) + \frac{\lambda_1}{2}(\|P\|_F^2 + \|X\|_F^2) + \frac{\lambda_2}{2}\|D\|_F^2)/\|W \odot Y\|_1$ in order to display four cases under the same scale.

denotes a cost value at each iteration. As shown in the figure, the cost value of factEN decreases fast and converges to a stationary point in a small number of iterations.

3. Experimental Results

We evaluated the performance of the proposed method, factEN, by experimenting with various synthetic and real-world problems, such as non-rigid motion estimation [30, 34], photometric stereo [4, 33], and background modeling [24, 32]. We compared factEN to the state-of-the-art low-rank approximation methods, ALADM[4] [23], Regl_1-ALM[5] [34], and Unifying [4], and rank estimation methods, IALM[6] [17] and ROSL[7] [24]. We set the parameters of factEN as follows: $\rho = 1.2$ for all cases, except for Giraffe and Static Face data sets, in which $\rho = 1.05$; and $\beta_0 = 0.5$ for all cases, except for non-rigid motion estimation problems, in which $\beta_0 = 10^{-2}$. Note that $\beta = \beta_0 / \|Y\|_\infty$.

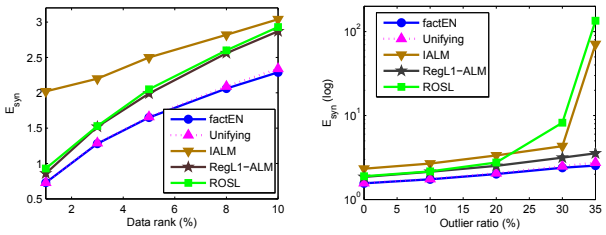
3.1. Synthetic data

First, we applied the proposed method to synthetic examples. We generated six test sets from 500×500 to $10,000 \times 10,000$ with Gaussian noises which were sampled from $\mathcal{N}(0, 10^{-2})$. In the experiment, the average reconstruction error E_{Syn} is calculated as $E_{Syn} = \frac{1}{n}||M^{gt} - \widehat{M}||_1$, where M^{gt} is the ground truth and $\widehat{M}$ is the low-rank matrix approximated by the applied algorithm.

Figure 3 shows average performances on a synthetic example (500×500) with various data ranks[8] and various outliers ratios to verify the robustness under various conditions. Overall, the proposed method and Unifying give the best average performance with respect to the reconstruction error for both cases. From Figure 3(b), most methods are robust when the outlier ratio is small, but ROSL and IALM give poor performance when the number of outliers increases, restricting their applications in practice.

To verify the ability of the proposed method compared to Unifying with respect to the rank and sparsity, we conducted an experiment for a 1,000×1,000 synthetic example. Figure 4 plots the fraction of correct recoveries at different rank and sparsity ratios. The region which is correctly recovered by the proposed method appears to be broader than that of Unifying. From the figure, the proposed method is more capable of handling corruptions than Unifying.

Figure 5(a) and 5(b) show average reconstruction errors and execution times of different algorithms, respectively, for various matrix sizes with 8% fixed data rank and



(a) (b)

Figure 3. Average performances on a synthetic example (500×500) with various conditions. (a) Average reconstruction errors for different observation data rank ratios (5% outliers). (b) Average reconstruction errors for different outlier ratios (5% data rank).

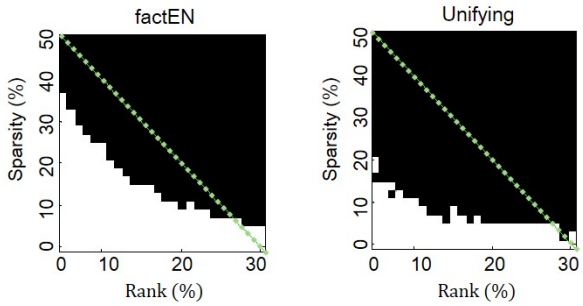


Figure 4. Phase transition in rank and sparsity for a synthetic example (1,000×1,000) using the proposed method and Unifying. Correct recovery (white region) is achieved when a recovered low-rank matrix $\widehat{M}$ satisfies $\|M^{gt} - \widehat{M}\|_1 / \|M^{gt}\|_1 \leq 5 \times 10^{-4}$.

4% outliers which were uniformly distributed in the range of $[-20, 20]$. We could not evaluate IALM and Regl_1-ALM for a large-scale problem ($10,000 \times 10,000$) because of their heavy computational complexity. The proposed method outperforms the other methods with respect to the reconstruction error in all cases. Although Regl_1-ALM shows the similar performance compared with the proposed method for small-scale data sets, it takes a longer computation time to get a good solution and shows poor performance for large-scale problems. The computing time of ALADM is faster than factEN, but it performs poorer than factEN.

To compare the proposed algorithm in realistic conditions, we changed the outliers to block corruptions with missing entries in a synthetic example. For a similarly constructed 300×300 example, we added occlusions with various sizes with 20% missing data. Figure 5(c) shows reconstruction errors of different methods. As shown in the figure, the proposed method robustly reconstructs corruptions while other methods except ALADM give poor reconstruction results when there are large-sized block corruptions.

3.2. Real-world problems

We evaluated the proposed method for real-world problems, which are summarized in Table 1. For these problems,

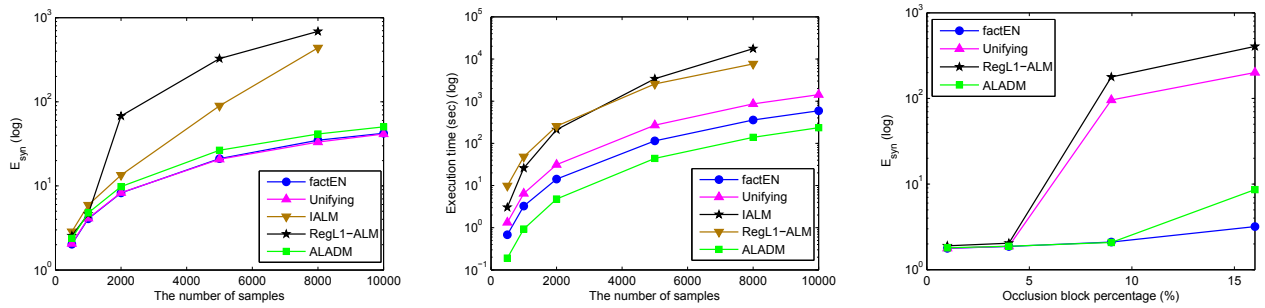
[4] http://lmafit.blogs.rice.edu/

[5] https://sites.google.com/site/yinqiangzheng/

[6] http://perception.csl.illinois.edu/matrix-rank/sample_code.html/

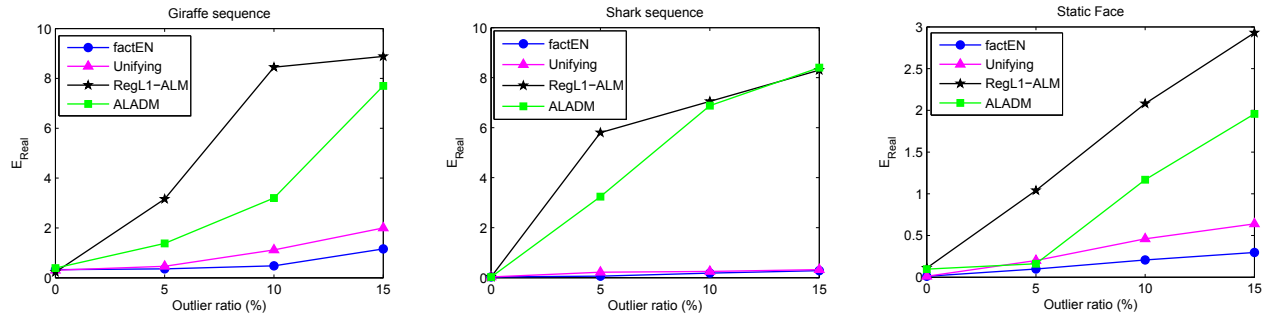
[7] https://sites.google.com/site/xianbiaoshu/

[8] Note that the data rank means the percentage of the true rank over the maximum possible rank of the data matrix.



(a) (b) (c)

Figure 5. Average performances for synthetic problems in the presence of corruptions. (a) Average reconstruction errors with random outliers for various data sizes. (b) Average execution times for various data sizes. (c) Average reconstruction errors with various block corruption sizes and 20% missing for an example of 300×300 in size.



(a) (b) (c)

Figure 6. Average performances on real-world problems (non-rigid motion estimation, photometric stereo) in the presence of outliers and missing data. (a) Giraffe sequence. (b) Shark sequence. (c) Static face.

we computed the mean absolute error (MAE) over the observed entries as $E_{Real} = \frac{||W \odot (M^{gt} - \widehat{M})||_1}{||W||_1}$.

First, we conducted a non-rigid motion estimation experiment using Giraffe sequence [3]. To demonstrate the robustness of the proposed method, we replaced 5% of the randomly selected points in a frame by outliers in the range of $[0, 100]$ whereas the data points are in the range of $[127, 523]$. In this setting, we performed several experiments by changing outlier ratio in the data. The result for the Giraffe sequence in the presence of various outlier levels is shown in Figure 6(a). The figure also includes the case when no outliers are added. As shown in the figure, factEN gives the best performance regardless of the outlier ratio. Although Unifying gives similar reconstruction performance when the outlier ratio is small, the performance gets worse as the outlier ratio increases. Regl_1-ALM and ALADM show worse performance compared to other state-of-the-art methods. Figure 7 shows how the average reconstruction error is affected by the choice of λ_1 for factEN and Unifying [4]. The proposed method shows more stable results under different values of λ_1 and λ_2, whereas Unifying is sensitive to the choice of λ_1.

Table 1. Summary of real-world problems with known rank r.

Datasets	Size	Rank r	Missing
Giraffe [3]	91×240	6	30 %
Shark [31]	240×167	6	10 %
Static Face [3]	$4{,}096 \times 20$	4	42 %
PETS 2009 [1]	$110{,}592 \times 221$	2	0 %

We also performed the motion estimation problem using the Shark sequence [31]. In this data, we randomly dropped 10% of points in each frame as missing data. We set from 0% to 15% of tracked points as outliers in each frame in the range of $[-1000, 1000]$, whereas the data points were located in the range of $[-105, 105]$. Average reconstruction errors at various outlier ratios by different methods are shown in Figure 6(b). As shown in the figure, factEN and Unifying both give outstanding reconstruction results. However, the proposed method gives the better reconstruction results than Unifying on average. The reconstruction results of the three selected algorithms are shown in Figure 8. From the figure, we can observe excellent reconstruction results by the proposed method against missing data and outliers compared to the other approaches.

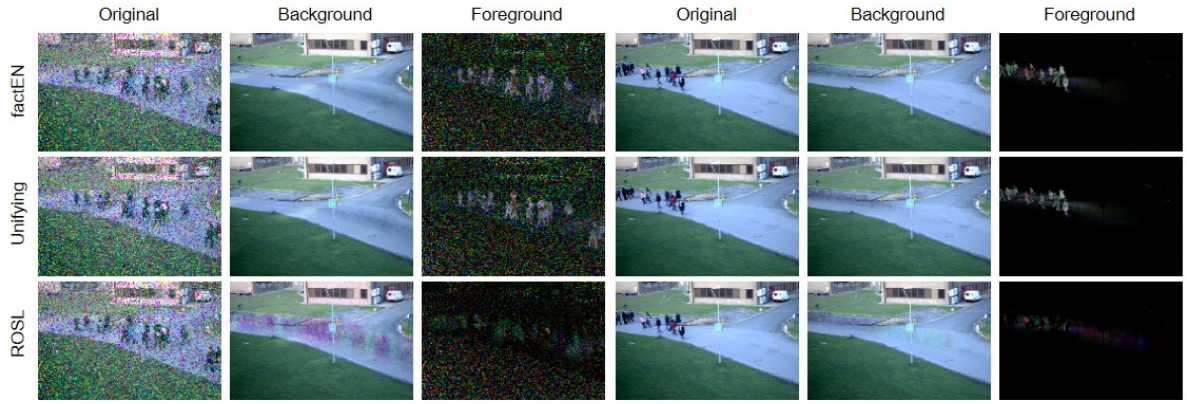


Figure 9. Background modeling results of the methods for two selected frames in the PETS2009 dataset. Each algorithm decomposes the original image into background and foreground images.

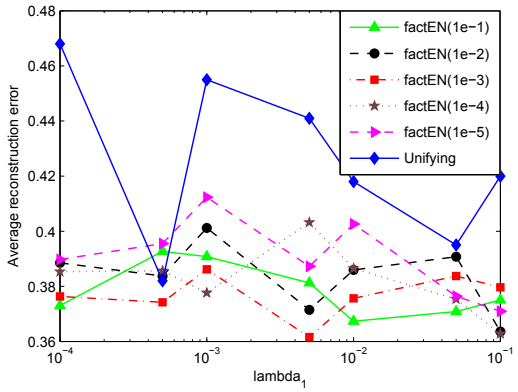


Figure 7. Comparison between the proposed method and Unifying [4] at different values of λ_1 for the Giraffe sequence. $(\cdot)$ denotes a value of λ_2.

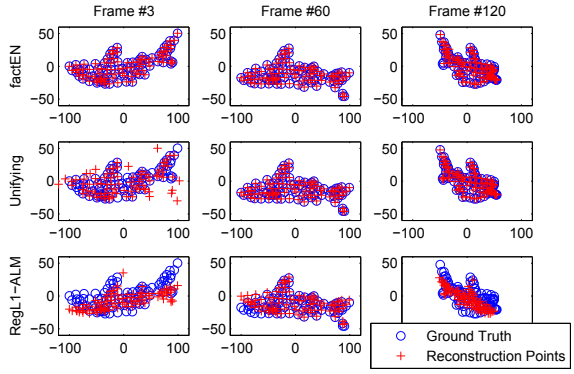


Figure 8. Reconstruction results from the shark sequence by three methods: factEN, Unifying [4], and RegL1-ALM [34].

For the photometric stereo problem, we used the Static Face sequence [4]. We examine how robust the proposed method is for various outlier ratios in the presence of missing data. We set from 0% to 15% of tracked points as outliers in each frame in the range of $[0, 100]$. The overall results are represented in Figure 6(c). From the figure, the proposed method gives the obvious distinction compared to other methods regardless of the outlier ratio.

For background modeling task, we used PETS2009 [1] and resized each frame to 288×384. We performed the proposed method compared with state-of-the-art methods: Unifying [4] and ROSL [24]. We added 30% random noises in randomly selected frames. Figure 9 shows the background modeling results on two selected frames. As shown in the figure, factEN and Unifying correctly separated foreground from background. The rank estimation method, ROSL, fails to find a good solution in the presence of heavy corruptions. The computation times are 186.37 sec for the proposed method, 497.46 sec for Unifying, and 145.93 sec for ROSL. Although ROSL gives the slightly faster computation time than factEN, it did not provide satisfying results.

4. Conclusions

In this paper, we have proposed a novel method, factEN, for practical subspace learning based on elastic-net regularization of singular values. The proposed method can handle missing or unknown entries as well as outliers. With the introduction of the proposed elastic-net regularization scheme, the proposed method can find a robust solution more efficiently and is stable against missing data, outliers, and different parameter values. The proposed method has been applied to various problems including non-rigid motion estimation, photometric stereo, and background modeling problems. The experimental results show that the proposed method outperforms other existing methods in terms of the approximation error and execution time. It will be interesting to investigate the competitiveness of the proposed method for large-scale and more challenging problems, including automatic rank estimation.

Acknowledgements

This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2013R1A1A2065551) and by the ICT R&D program of MSIP/IITP (B0101-15-0307, Basic Software Research in Human-Level Lifelong Machine Learning).

References

[1] PETS 2009 dataset. http://www.cvg.rdg.ac.uk/PETS2009.

[2] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] A. M. Buchanan and A. W. Fitzgibbon. Damped newton algorithms for matrix factorization with missing data. In *CVPR*, 2005.

[4] R. Cabral, F. D. la Torre, J. P. Costeira, and A. Bernardino. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *ICCV*, 2013.

[5] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58:11:1–11:37, 2011.

[6] X. Ding, L. He, and L. Carin. Bayesian robust principal component analysis. *IEEE Trans. on Image Processing*, 20:3419–3430, 2011.

[7] A. Eriksson and A. Hengel. Efficient computation of robust weighted low-rank matrix approximations using the l_1 norm. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(9):1681–1690, 2012.

[8] X. Guo, X. Wang, L. Yang, X. Cao, and Y. Ma. Robust foreground detection using smoothness and arbitrariness constraints. In *ECCV*, 2014.

[9] Z. Harchaoui, M. Douze, M. Paulin, M. Dudik, and J. Malick. Large-scale image classification with trace-norm regularization. In *CVPR*, 2012.

[10] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(9):2117–2130, 2013.

[11] I. T. Jolliffe. *Principal Component Analysis*. John Wiley and Sons, 1986.

[12] Q. Ke and T. Kanade. Robust l_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *CVPR*, 2005.

[13] E. Kim, M. Lee, C.-H. Choi, N. Kwak, and S. Oh. Efficient l_1-norm-based low-rank matrix approximations for large-scale problems using alternating rectified gradient method. *IEEE Trans. on Neural Networks and Learning Systems*, 26(2):237–251, 2015.

[14] N. Kwak. Principal component analysis based on L_1-norm maximization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 3(9):1672–1680, 2008.

[15] H. Li, N. Chen, and L. Li. Error analysis for matrix elastic-net regularization algorithms. *IEEE Trans. on Neural Networks and Learning Systems*, 23(5):737–748, 2012.

[16] C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19:2756–2779, 2007.

[17] Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. *Mathematical Programming*, 2010.

[18] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

[19] D. Meng and F. D. la Torre. Robust matrix factorization with unknown noise. In *ICCV*, 2013.

[20] K. Mitra, S. Sheorey, and R. Chellappa. Large-scale matrix factorization with missing data under additional constraints. In *NIPS*, 2010.

[21] T. Okatani, T. Yoshida, and K. Deguchi. Efficient algorithm for low-rank matrix factorization with missing components and performance comparision of latest algorithms. In *ICCV*, 2011.

[22] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52:471–501, 2010.

[23] Y. Shen, Z. Wen, and Y. Zhang. Augmented Lagrangian alternating direction method for matrix separation based on low-rank factorization. *Optimization Methods and Software*, pages 1–26, 2012.

[24] X. Shu, F. Porikli, and N. Ahuja. Robust orthonormal subspace learning: Efficient recovery of corrupted low-rank matrices. In *CVPR*, 2014.

[25] N. Srebro. Weighted low-rank approximations. In *ICML*, 2003.

[26] T. Sun and C.-H. Zhang. Calibrated elastic regularization in matrix completion. In *NIPS*, 2012.

[27] A. Talwalkar, L. Mackey, Y. Mu, S.-F. Chang, and M. I. Jordan. Distributed low-rank subspace segmentation. In *ICCV*, 2013.

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

[29] K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Journal of Optimization*, 6(3):615–640, 2010.

[30] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[31] L. Torresani, A. Hertzmann, and C. Bregler. Learning non-rigid 3d shape from 2d motion. In *NIPS*, 2003.

[32] N. Wang and D.-Y. Yeung. Bayesian robust matrix factorization for image and video processing. In *ICCV*, 2013.

[33] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, T. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *ACCV*, 2010.

[34] Y. Zheng, G. Liu, S. Sugimoto, S. Yan, and M. Okutomi. Practical low-rank matrix approximation under robust l_1-norm. In *CVPR*, 2012.

[35] T. Zhou and D. Tao. Godec: Randomized low-rank & sparse matrix decomposition in noisy case. In *ICML*, 2011.

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