Robust Sparse PCA via Weighted Elastic Net

Ling Wang[⋆] and Hong Cheng

University of Electronic Science and Technology of China,
Gaoxin West Zone Xiyuan Avenue 2006, 611731, Chengdu, China
{eewangling,hchenguestc}@gmail.com
http://www.uestcrobot.net/

Abstract. In principal component analysis (PCA), ℓ_2/ℓ_1-norm is widely used to measure coding residual. In this case, it assume that the residual follows Gaussian/Laplacian distribution. However, it may fail to describe the coding errors in practice when there are outliers. Toward this end, this paper propose a Robust Sparse PCA (RSPCA) approach to solve the outlier problem, by modeling the sparse coding as a sparsity-constrained weighted regression problem. By using a series of equivalent transformations, we show the proposed RSPCA is equivalent to the Weighted Elastic Net (WEN) problem and thus the Least Angle Regression Elastic Net (LARS-EN) algorithm is used to yield the optimal solution. Simulation results illustrated the effectiveness of this approach.

Keywords: Principal Component Analysis, Sparse Representation, Robust statistics, Elastic Net.

1 Introduction

Sparse learning based dimension reduction approaches have drawn many attention of many researchers over world recently since it can reduce not only dimensionality of feature spaces but also the number of explicitly used feature vectors [1, 2]. As we know, Principal Component Analysis (PCA) , whose each principal component (PC) can decomposed as a linear combination of all the original features, are widely used in data processing and dimensionality reduction[3, 4]. PCA can be formulated as a regression-type optimization problem, and then we can obtain sparse loadings by imposing Least Absolute Shrinkage and Selection Operator (LASSO) constraints on the regression coefficients [1, 3]. Moreover, convex optimization is also used to improve interpretation of the standard PCA [5]. Traditionally, when the number of feature dimensions is larger than the number of observation vectors, we have to select all of the feature vectors and thus it cannot reduce their dimensions.

It would be of interest to discover sparse principal components, i.e., sets of sparse vectors spanning a low-dimensional space that explains most of the variance present in the data. In this case, Sparse Principal Component Analysis

[⋆] This work was supported by "the Fundamental Research Funds for the Central Universities" under award number ZYGX2010J016.

C.-L. Liu, C. Zhang, and L. Wang (Eds.): CCPR 2012, CCIS 321, pp. 88–95, 2012.

(SPCA) provides a direct way to reduce the number of feature vectors for feature representation and learning [1]. Accordingly, various SPCA approaches have been developed [5–9], where the residual follows Gaussian/Laplacian distribution. However, there are some outliers in the sample data in practice. In this case, the ℓ_2-norm can't give the right solution, shown in Fig. 1(b). The outliers have a very large influence on ℓ_2-norm because the residual $\mathbf{e}$ is squared. Consequently, $y = \|\mathbf{e}\|_2^2$ is increased sharply when its value is increasing, shown in Fig. 1(a). Furthermore some strategies using ℓ_1-norm is used to handle this issue in PCA [10–13]Indeed, in some cases, the ℓ_1-norm performs better than ℓ_2-norm, shown in Fig. 1(b). Unfortunately, ℓ_1-norm can not model the outlier problem explicitly. From Fig. 1(c), we can see when the outlier lies in x-axis, the ℓ_1-norm fails to obtain the right solution. In this case, robust PCA using ρ functions could handle these outliers very well. There are other robust PCA or SPCA were proposed [14–16] to solve the outlier problem. Candes et al. also given the solutions on robust PCA [17] by using Augmented Lagrangian Method on low-rank matrix decomposition. Croux et al. used grid algorithm to compute the sparse and robust principal components [18].

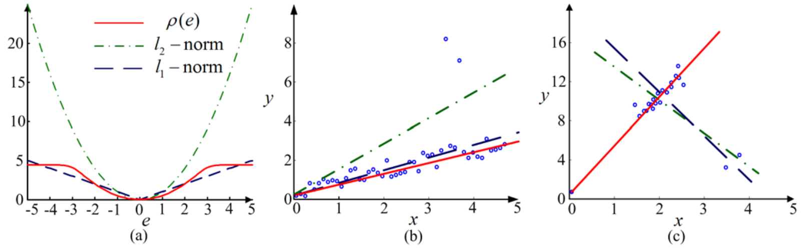

Fig. 1. An illustration for various PCA approaches: (a) Different residual functions, ℓ_1-norm, ℓ_2-norm, and the ρ-function; (b) Toy data 1; (c) Toy data 2 (best viewed in color)

In this paper, we proposed a novel Robust SPCA (RSPCA) approach, which is formulated as a weighted elastic net (WEN) approach. The proposed RSPCA approach are not only to reduce the dimensionality of data matrix but also to make it more robust under noise. Here, we consider SPCA as a regression-type optimization problem. By replacing the data item in the LARS-EN algorithm using specific robust function, the LARS-EN algorithm can be formulated as the WEN approach, which models the outlier problem explicitly thus improving the robustness of SPCA.

2 Problem Formulation

Let's first briefly review regression-type optimization framework of PCA. The PCs of data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ is $\mathbf{P} = \mathbf{X}V_k^T$, where n is the sample dimension, p is the feature dimension, $\mathbf{V}$ is the right singular vectors of $\mathbf{X}$, i.e. $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$,

$\mathbf{V}_k$ is called as loading matrix or projecting matrix. Usually $k << p$, and thus dimensionality reduction is achieved. Furthermore, the uncorrelated PCs capture the maximum variability of $\mathbf{X}$, which guarantees minimal information loss. In regression-type optimization frameworks, PCA can be formulated as an LARS-EN problem [1]. The sparse solution of j-th component $P_{\cdot j}$ can be solved by

$$\hat{\beta}_{opt} = \arg\min_{\beta_j} \|P_{\cdot j} - \mathbf{X}\beta_j\|_2^2 + \lambda_1\|\beta_j\|_2^2 + \gamma_1\|\beta_j\|_1, \tag{1}$$

where $\beta_j \in \mathbb{R}^{p\times 1}$, both λ_1 and γ_1 are non-negative Lagrange multipliers. Afterwards, we can obtain $\hat{V}_{\cdot j} = \hat{\beta}_{opt}/\|\hat{\beta}_{opt}\|_2$.

Eqn. (1) is a convex combination of the ridge penalty and ℓ_1-norm penalty. The ridge penalty is used to ensure the reconstruction of PCs, while the ℓ_1-norm penalty is used to ensure the sparsity of loadings. Larger γ_1 encourages a sparser $\hat{\beta}$. Given a fixed λ_1 and γ_1, optimizing Eqn. (1) can efficiently obtain $\hat{\beta}_{opt}$ by using the LARS-EN algorithm [19].

Now we discuss sparse PCA with noise. Assume data matrix $\mathbf{X} = \mathbf{S}+\mathbf{N}$, where $\mathbf{S}$ is original data matrix, $\mathbf{N}$ is noise matrix. The noise-free PCs of $\mathbf{S} = \overline{\mathbf{U}}\,\overline{\mathbf{D}}\,\overline{\mathbf{V}}^T$ are $\overline{\mathbf{P}} = \mathbf{S}\overline{\mathbf{V}}_k$, where $\overline{\mathbf{V}}_k = [\alpha_1, \cdots, \alpha_k]$. Then the j-th principal component $\overline{P}_{\cdot j} = \mathbf{S}\alpha_j = \mathbf{X}\alpha_j - \mathbf{N}\alpha_j$.

Similarly, we can estimate optimal PCs from observation matrix, $P_{\cdot j} = \mathbf{X}\beta_j$, where β_j is from the loading matrix $\mathbf{V}_k$ estimated from $\mathbf{X}$. Suppose the ideal PCs be estimated by $\mathbf{X}\alpha_j$, and the actual PCs be estimated by $\mathbf{X}\beta_j$, then the PCA optimization problem is formulated as

$$\arg\min_{\alpha_j,\beta_j} \|\mathbf{X}\alpha_j - \mathbf{X}\beta_j\|_2^2, \quad \text{s.t.}\ \|\alpha_j\|_2^2 = 1,\ \|\beta_j\|_2^2 = 1. \tag{2}$$

In this equation, we can add sparse constraints, $\|\alpha_j\|_1 < \xi_2$, $\|\beta_j\|_1 < \xi_1$, where ξ_1 and ξ_2 are small constants.

Theorem 1. *For any $\lambda_1 > 0, \lambda_2 > 0$, if $\hat{\beta}_j$ is given by*

$$(\hat{\alpha}_j, \hat{\beta}_j) = \arg\min_{\alpha_j,\beta_j} \|\mathbf{X}\alpha_j - \mathbf{X}\beta_j\|_2^2 + \lambda_1\|\beta_j\|_2^2 + \lambda_2\|\alpha_j\|_2^2, \tag{3}$$

then $\hat{V}_{\cdot j} = \hat{\beta}_j/\|\hat{\beta}_j\|_2$.

Define $\epsilon_j = \mathbf{X}\alpha_j - \mathbf{X}\beta_j = [e_1, \cdots, e_n]^T$, and assume that $e_i, (i = 1, \cdots, n)$ are independently identically distributed (*i.i.d.*) with probability density function (*p.d.f.*) $f(e_i)$, then the likelihood function is $\mathbf{L}(e_1, \cdots, e_n) = \prod_{i=1}^n f(e_i)$. We minimize the objective function by using maximum likelihood estimation,

$$-\ln \mathbf{L} = \sum\nolimits_{i=1}^n - \ln f(e_i) = \sum\nolimits_{i=1}^n \rho(e_i) = \mathbf{F}(\epsilon_j), \tag{4}$$

where $\rho(e_i) = -\ln f(e_i)$. Approximating $\mathbf{F}(\epsilon_j)$ by its first order Taylor expansion in the neighborhood of ϵ_0, we have

$$\tilde{\mathbf{F}}(\epsilon_j) = \mathbf{F}(\epsilon_0) + (\epsilon_j - \epsilon_0)^T\mathbf{F}'(\epsilon_0) + \mathbf{R}_1(\epsilon_j), \tag{5}$$

where $\mathbf{R}_1(\epsilon_j)$ is the high order residual term. In sparse coding, it usually assume that the fidelity term is strictly convex. We approximate the residual term as

$$\mathbf{R}_1(\epsilon_j) = \frac{1}{2}(\epsilon_j - \epsilon_0)^T \mathbf{W}_{(j)}(\epsilon_j - \epsilon_0), \tag{6}$$

where $\mathbf{W}_{(j)}$ is a diagonal matrix for that the elements in ϵ_j are independent and there is no cross term between e_i and e_l, $(i \neq l)$. Since $\mathbf{F}(\epsilon_j)$ reaches its minimal value at $\epsilon_j = \mathbf{0}$, we also require that $\tilde{\mathbf{F}}(\epsilon_j)$ has its minimal value at $\epsilon_j = \mathbf{0}$. Letting $\tilde{\mathbf{F}}(\mathbf{0}) = \mathbf{0}$, we have the diagonal elements of $\mathbf{W}_{(j)}$ as

$$W_{ii} = \omega(e_{0,i}) = \rho'(e_{0,i})/e_{0,i}. \tag{7}$$

Then $\tilde{\mathbf{F}}(\epsilon_j)$ can be written as $\tilde{\mathbf{F}}(\epsilon_j) = \frac{1}{2}\left|\mathbf{W}_{(j)}^{1/2}\epsilon_j\right|^2 + b$, where b is a scalar value determined by ϵ_0.

Since the logistic function has properties similar to the hinge loss function in SVM, we choose it as the weight function

$$\omega(e_i) = \exp(\mu\delta - \mu e_i^2)/(1 + \exp(\mu\delta - \mu e_i^2)). \tag{8}$$

For the relationship in Eqn. (7), we have

$$\rho(e_i) = -\frac{1}{2\mu}\left(\ln(1 + \exp(\mu\delta - \mu e_i^2)) - \ln(1 + \exp(\mu\delta))\right), \tag{9}$$

where μ and δ are positive scalars. μ controls the decreasing rate from 1 to 0, and δ controls the location of demarcation point. The $\rho(e)$ is shown in Fig. 1 (a), when the residual $|e|$ exceeds a threshold, it cannot increase with the residual. In other words, the outlier would be adaptively assigned with low weights to reduce their affects on the regression estimation thus resulting in more robust dimension reduction.

Then the Eqn. (3) can be approximated by

$$\begin{aligned}(\hat{\alpha}_j, \hat{\beta}_j) = \arg\min_{\alpha_j,\beta_j} \; & \|\mathbf{W}_{(j)}^{1/2}(\mathbf{X}\alpha_j - \mathbf{X}\beta_j)\|_2^2 + \lambda_1\|\beta_j\|_2^2 + \gamma_{1,j}\|\beta_j\|_1 \\ & + \lambda_2\|\alpha_j\|_2^2 + \gamma_{2,j}\|\alpha_j\|_1.\end{aligned} \tag{10}$$

This is a weighted elastic net problem, which can be solved by LARS-EN algorithm.

3 The Proposed Robust Sparse PCA

Though Eqn. (10), we formulate the RSPCA into a regression-type optimization problem, and then obtain any one of PCs. As we know, we are interested not only in the first PC, but also in the other PCs. We can get the first k PCs by optimizing the optimization equation in Theorem 2.

Theorem 2. *For any* $\lambda_1 > 0, \lambda_2 > 0$, $j = 1, 2, \cdots, k$, *let*

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \sum_{j=1}^{k} \left(\left\| \mathbf{W}_{(j)}^{1/2} (\mathbf{X}\alpha_j - \mathbf{X}\beta_j) \right\|_2^2 + \lambda_1 \|\beta_j\|_2^2 + \lambda_2 \|\alpha_j\|_2^2 \right), \qquad (11)$$

where $\mathbf{A} = \overline{\mathbf{V}}_k = [\alpha_1, \cdots, \alpha_k]$, $\mathbf{B} = \mathbf{V}_k = [\beta_1, \cdots, \beta_k]$, *and then* $\hat{\beta}_j \propto V_{\cdot j}$.

Only the first sparse loading is β_1, usually the first leading sparse loading may not be sufficient for obtain a variable support, and it is desirable to further estimate a few subsequent sparse loadings as well. Several techniques [8] have been explored for reliably deflating the covariance matrix of SPCA. Here we use the low-rank approximations [9] to eliminate the influence of the first sparse loading. Given the first sparse loading β_1, we have

$$\overline{\mathbf{X}} = \mathbf{X} - \mathbf{X}\beta_1\beta_1^T. \qquad (12)$$

Then the second sparse loading β_2 of $\mathbf{X}$ becomes the leading sparse loading of $\overline{\mathbf{X}}$, and can be estimated again by using RSPCA.

According to the Theorem 2, we propose an iterative approach to minimize the RSPCA. The detail algorithm is summarized in Table 1.

4 Experimental Results and Analysis

In this section, we evaluate the proposed RSPCA approach compared with both PCA and SPCA. We set the original data source with three components,

$$\begin{aligned} x_1(i) &= \begin{cases} 1 + \varepsilon(i), & n/2 < i \le n-1 \\ 0, & 0 \le i \le n/2 \end{cases}, \\ x_2(i) &= 0.9e^{-(i-250)^2/1250} + \varepsilon(i),\ 0 \le i \le n-1, \\ x_3(i) &= 0.7e^{-(i-75)^2/250} + 0.7e^{-(i-425)^2/250} + \varepsilon(i),\ 0 \le i \le n-1. \end{aligned} \qquad (15)$$

where n is the sample dimension of data. Three different and independent Non-Gaussian noises $\varepsilon(n)$ are added into the original data vectors respectively, uniform noise with minimum SNR$= -12dB$, Gaussian color noise with minimum SNR$= -12dB$, pink noise with minimum SNR$= -10dB$.

The original data are shown in the first column of Fig. 2. After adding noise, we re-sample the each observation data with D times, and then combine the sampled data sets as $\mathbf{X} \in \mathbb{R}^{n \times 3D}$, i.e. $p = 3D$. In this experiment, we set $n = 500$ and $D = 500$. The original data corrupted by different noises are shown in Fig. 2 (the second column to the forth column), where the rows from the first to the third is denote the x_1, x_2, x_3, respectively; the columns from second to the forth is denote the data corrupted by uniform noise, Gaussian color noise and pink noise, respectively.

In the weight function Eqn. (8), there are two parameters δ and μ, which need to be calculated in the iterative steps. δ determines the location of demarcation

Table 1. The proposed RSPCA via Weighted Elastic Net Algorithm

```
Input: observation data matrix $\mathbf{X}$
Output: $\hat{V}_i$ loadings of PCs
Initialization : $\mathbf{X} = \mathbf{UDV}^T$, $\alpha_j = V_{\cdot j}$, $\beta_j = \mathbf{0}$, $P_{ite}^{(1)}$ = means of every rows of $\mathbf{X}$
start:
 for $j = 1, \cdots, k$
     $t = 1$
     while $\hat{\beta}_j$ is not converges
       • Compute residual $\epsilon_j^{(t)} = [e_1, \cdots, e_n]^T = \mathbf{X}^{(j)}\alpha_j - P_{ite}^{(t)}$
       • Estimate weights $\omega(e_i^{(t)})$ with Eqn. (8)
       • For given fixed $\alpha_j$, solve the elastic net problem in following equation,
           $\hat{\beta_j} = \arg\min_{\beta_j} (\alpha_j - \beta_j)^T \mathbf{X}^T \mathbf{W}_{(j)} \mathbf{X}(\alpha_j - \beta_j) + \lambda_1 \|\beta_j\|_2^2 + \gamma_{1,j} \|\beta_j\|_1.$   (13)
       • For fixed $\beta_j$, update $\alpha_j$ by the following equation
           $\hat{\alpha_j} = \arg\min_{\alpha_j} (\alpha_j - \beta_j)^T \mathbf{X}^T \mathbf{W}_{(j)} \mathbf{X}(\alpha_j - \beta_j) + \lambda_2 \|\alpha_j\|_2^2 + \gamma_{2,j} \|\alpha_j\|_1.$   (14)
       • Update: $\mathbf{X}^{(j+1)} = \mathbf{X}^{(j)} - \mathbf{X}^{(j)}\hat{\beta}_j\hat{\beta}_j^T$, $P_{ite}^{(t)} = \mathbf{X}^{(j)}\hat{\beta}_j$.
       • Let $t = t + 1$.
     end while
 end for.
Normalization : $\hat{V}_{\cdot j} = \hat{\beta}_j / \|\hat{\beta}_j\|_2, j = 1, \cdots, k$.
```

point, and μ controls the decreasing rate of weight. In order to make the model robust to outliers, we compute the δ and μ as follows. Given $\psi = [e_1^2, \cdots e_n^2]$, we can get ψ_a by sorting ψ in an ascending order. Let $k = floor(\tau n)$, $\tau \in (0, 1]$, and then we set $\delta = \psi_a(k)$, and $\mu = c/\delta$. In this experiment, we set $\tau = 0.5$, $c = 8$, $\gamma_1 = 100$ in Eqn. (13) and $\gamma_2 = 100$ in Eqn. (14).

Given observation matrix $\mathbf{X}$, we use the PCA, SPCA and the proposed RSPCA to estimate the loadings $\hat{\mathbf{V}}_j$ of $\mathbf{X}$, respectively. Then the PCs is computed as $\mathbf{X}\hat{\mathbf{V}}_j$, which are used to recover the original data. In this experiment, we use the first three sparse loadings as the select vectors. The experimental results are shown in Fig. 3, where blue lines correspond to $\hat{x}_1$, green lines correspond to $\hat{x}_2$ and red lines correspond to $\hat{x}_3$. Moreover, the rows denote the PCA, SPCA and the proposed RSPCA, respectively. The columns denote the data corrupted by uniform noise, Gaussian color noise and pink noise, respectively.

From Fig. 3, we can see that the PCA approach can separate the data from uniform and Gaussian color noise, but the loadings are not sparse. Under pink noise disturbance, there are poor performance in data estimation. By using

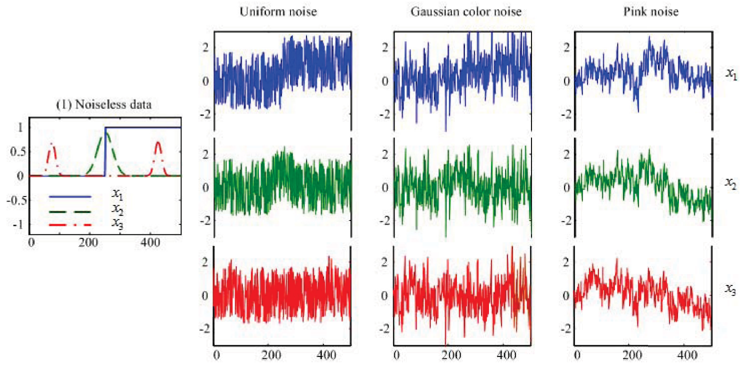

Fig. 2. Data with various Noise: The first column is the original data; the rows from the first to the third is denote the x_1, x_2, x_3, respectively; the columns from second to the forth is denote the data corrupted by uniform noise, Gaussian color noise and pink noise, respectively

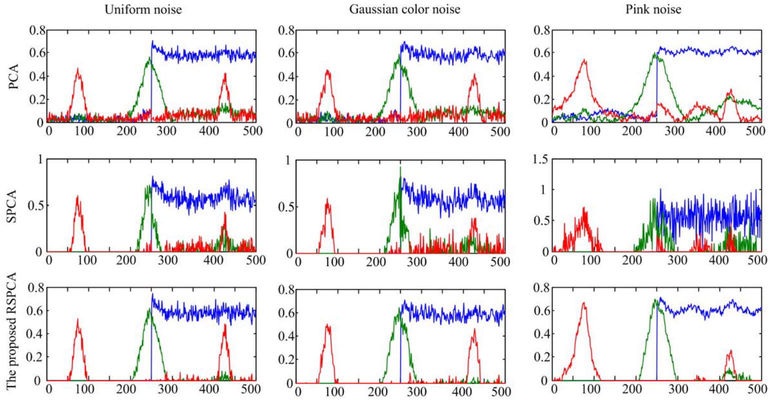

Fig. 3. The estimated PCs under Non-Gaussian color noise: the rows denote the PCA, SPCA and the proposed RSPCA, respectively; the columns denote the data corrupted by uniform noise, Gaussian color noise and pink noise, respectively

SPCA approach, the loading sparsity is better than PCA, but at the same time, the data estimated performance is poor than PCA. Finally, the proposed RSPCA using the WEN algorithm has better estimation.

5 Conclusion and Future Work

In this paper, we have proposed a novel RSPCA approach to estimate the data principal components to handle outliers. By a series of equivalent transformation, we formulate the proposed RSPCA into WEN problem. With an iterative

LARS-EN procedure, better estimation has been obtained compare with PCA and SPCA approaches. In the future work, we will use the proposed approach on object categorization.

References

1. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. Journal of Computational and Graphical Statistics 15, 265–286 (2006)
2. Zhou, T., Tao, D., Wu, X.: Manifold elastic net: a unified framework for sparse dimension reduction. In: Data Mining and Knowledge Discovery, vol. 22, pp. 340–371 (2011)
3. Jolliffe, I.T.: Principal component analysis. Wiley Online Library (2002)
4. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and Intelligent Laboratory Systems 2, 37–52 (1987)
5. d'Aspremont, A., El Ghaoui, L., Jordan, M.I., Lanckriet, G.R.G.: A direct formulation for sparse PCA using semidefinite programming. Computer Science Division (2004)
6. Moghaddam, B., Weiss, Y., Avidan, S.: Spectral bounds for sparse PCA: Exact and greedy algorithms. In: Advances in Neural Information Processing Systems, vol. 18 (2006)
7. Shen, H., Huang, J.Z.: Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis 99, 1015–1034 (2008)
8. Mackey, L.: Deflation methods for sparse PCA. In: Advances in Neural Information Processing Systems, vol. 21, pp. 1017–1024 (2009)
9. Frieze, A., Kannan, R., Vempala, S.: Fast Monte-Carlo algorithms for finding low-rank approximations. Journal of ACM 51, 1025–1041 (2004)
10. Meng, D., Zhao, Q., Xu, Z.: Robust sparse principal component analysis. Preprint (2010)
11. Ding, C., Zhou, D., He, X., Zha, H.: R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization. In: ICML (2006)
12. Ke, Q., Kanade, T.: Robust l1 norm factorization in the presence of outliers and missing data by alternative convex programming. In: IEEE CVPR (2005)
13. Kwak, N.: Principal component analysis based on L1-norm maximization. IEEE TPAMI 30, 1672–1680 (2008)
14. De La Torre, F., Black, M.J.: Robust principal component analysis for computer vision. In: IEEE ICCV (2001)
15. De La Torre, F., Black, M.J.: A framework for robust subspace learning. International Journal of Computer Vision 54, 117–142 (2003)
16. Aanæs, H., Fisker, R., Astrom, K., Carstensen, J.M.: Robust factorization. IEEE Transactions on PAMI 24, 1215–1225 (2002)
17. Candes, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? Arxiv preprint ArXiv: 0912.3599 (2009)
18. Croux, C., Filzmoser, P., Fritz, H.: Robust sparse principal component analysis. Catholic University of Leuven Department of Decision Science and Information Management Working Paper (2011)
19. Zou, H., Hastie, T.: Regression shrinkage and selection via the elastic net, with applications to microarrays. Journal of Royal Statist Society B, 1–26 (2003)