BAYESIAN LEARNING FOR ROBUST PRINCIPAL COMPONENT ANALYSIS

Martin Sundin, Saikat Chatterjee and Magnus Jansson

ACCESS Linnaeus Center, School of Electrical Engineering
KTH Royal Institute of Technology, Stockholm, Sweden
masundi@kth.se, sach@kth.se, magnus.jansson@ee.kth.se

ABSTRACT

We develop a Bayesian learning method for robust principal component analysis where the main task is to estimate a low-rank matrix from noisy and outlier contaminated measurements. To promote low-rank, we use a structured Gaussian prior that induces correlations among column vectors as well as row vectors of the matrix under estimation. In our method, the noise and outliers are modeled by a combined noise model. The method is evaluated and compared to other methods using synthetic data as well as data from the MovieLens 100K dataset. Comparisons show that the method empirically provides a significant performance improvement over existing methods.

Index Terms— Robust principal component analysis, matrix completion, Bayesian learning.

1. INTRODUCTION

Robust Principal Component Analysis (RPCA) is the problem of estimating a low-rank matrix $\mathbf{X}$ from measurements

$$\mathbf{Y} = \mathbf{X} + \mathbf{S} + \mathbf{N} \in \mathbb{R}^{p \times q}, \tag{1}$$

where $\mathbf{Y}$ is the observed matrix, $\mathbf{N}$ is additive dense noise (typically isotropic Gaussian) and $\mathbf{S}$ is a sparse matrix modeling outliers. The RPCA model (1) has been used in e.g. [1–11] to model different phenomena and has applications in e.g. image processing, collaborative filtering, face recognition [1] and machine learning [2]. The motivation for RPCA originates from the fact that many signals lie in a low-dimensional subspace and are subjected to additive dense noise. Often, some measurements are also contaminated by outliers. To estimate the signal of interest under such conditions, outlier robust estimation of the principal components of $\mathbf{X}$ is required. The system model (1) can further be used in a matrix completion setup, where only some components of $\mathbf{Y}$ are observed, by modeling the unobserved entries as outliers.

In the literature, three classes of estimation methods are typically used: greedy, convex optimization based and Bayesian. The greedy method in [3] uses alternating optimization to estimate $\mathbf{X}$ and $\mathbf{S}$ via a least-squares principle. The greedy method is highly effective but requires rank and sparsity level to be known a-priori. It may therefore not be feasible in many applications. The convex optimization based method in [1] is called principal component pursuit (PCP) and uses nuclear-norm and ℓ_1-norm penalty functions to estimate $\mathbf{X}$ and $\mathbf{S}$. One limitation of PCP is that it needs to know the dense noise power a-priori. In absence of a-priori knowledge, Bayesian methods are a suitable choice since they can learn all necessary parameters from data. Formulating a Bayesian method for (1) requires low-rank and sparsity promoting priors for $\mathbf{X}$ and $\mathbf{S}$, respectively. The method of [4] uses a variational Bayes (VB) approach where the low-rank prior is induced using block sparsity in a matrix factorization model. The use of sparsity is a secondary approach to promote low-rank and hence we have experienced that the VB method suffers in performance loss. The empirical Bayes (EB) method of [5] promotes low-rank by modeling the column vectors of $\mathbf{X}$ as correlated Gaussian vectors. Further, in [5] $\mathbf{S}$ is given a sparsity promoting prior by the usual approach where the elements are Gaussian variables with gamma distributed precisions.

In this paper we develop a new Bayesian method for RPCA. To promote low-rank, we use a model that induces correlations among the column and row vectors of $\mathbf{X}$. The low-rank prior was first introduced by us in [12], where the Bayesian learning method was named relevance singular vector machine (RSVM) in [12]. The RSVM was not investigated for robustness against sparse outliers. Further, for typical sparse representations (for example compressed sensing), we experienced in [13] that a combined model of the sparse and dense noise works better than treating the noise terms independently. In this paper, we investigate the Bayesian modeling of the RPCA problem (1) using a combined model for $\mathbf{S} + \mathbf{N}$ and the low-rank promoting prior from [12]. The new method is called robust RSVM (rRSVM) and the parameters are estimated using the expectation-maximization (EM) framework. Through numerical simulations, we investigate the performance for synthetic data as well as real data from the MovieLens 100K dataset [14]. The performance of rRSVM is found to be better than that of the competing

This work was partially supported by the Swedish Research Council under contract 621-2011-5847.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

algorithms PCP, EB and VB for both synthetic and real data.

2. ROBUST RSVM

To formulate a Bayesian learning method for the RPCA model (1), we need appropriate priors for $\mathbf{X}$, $\mathbf{S}$ and $\mathbf{N}$. We first discuss the priors in Section 2.1 and then design the learning algorithm rRSVM in Section 2.2 using the EM framework.

2.1. Appropriate priors

2.1.1. Low-rank promoting prior

We hypothesise that the structure of a low-rank matrix is determined by the subspaces in which the column and row vectors lie. To model this structure, according to our work [12], we use the following low-rank promoting prior

$$\mathbf{X} = \boldsymbol{\alpha}_L^{-1/2}\mathbf{U}\boldsymbol{\alpha}_R^{-1/2},$$

where the precision matrices $\boldsymbol{\alpha}_L \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\alpha}_R \in \mathbb{R}^{q\times q}$ are positive definite and the components of $\mathbf{U} \in \mathbb{R}^{p\times q}$ are $\mathcal{N}(0,1)$ distributed (iid Gaussian). This is equivalent to setting

$$\mathrm{vec}(\mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\alpha}_R^{-1} \otimes \boldsymbol{\alpha}_L^{-1}) \tag{2}$$

as the prior distribution, where $\mathrm{vec}(\cdot)$ denotes the standard vectorization operation of a matrix and $\otimes$ denotes the Kronecker product.

Let $\mathbf{x}_i$ denote the i'th column vector of $\mathbf{X}$. From (2), we find that two column vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ of $\mathbf{X}$ are Gaussian, zero mean and correlated as

$$E[\mathbf{x}_i\mathbf{x}_j^\top] = \boldsymbol{\alpha}_L^{-1}[\boldsymbol{\alpha}_R^{-1}]_{ij}.$$

This means that the column vectors lie in the range space of $\boldsymbol{\alpha}_L^{-1}$. Similarly, the row vectors of $\mathbf{X}$ are found to be in the range space of $\boldsymbol{\alpha}_R^{-1}$. In practice, the inverse precision matrices are highly skewed and hence only approximately low-rank. For the model (2) to promote low-rank, we assign Wishart priors to the precision matrices as follows

$$\begin{aligned} p(\boldsymbol{\alpha}_L) &\propto |\boldsymbol{\alpha}_L|^{(\nu-q)/2} e^{-\frac{\epsilon}{2}\mathrm{tr}(\boldsymbol{\alpha}_L)}, \\ p(\boldsymbol{\alpha}_R) &\propto |\boldsymbol{\alpha}_R|^{(\nu-p)/2} e^{-\frac{\epsilon}{2}\mathrm{tr}(\boldsymbol{\alpha}_R)}, \end{aligned} \tag{3}$$

where $|\cdot|$ denotes the matrix determinant and $\mathrm{tr}(\cdot)$ the trace. The Wishart distribution is a conjugate prior for a one-sided precision and is closely related to the sparsity promoting Gamma prior used by the relevance vector machine (RVM) [15] and in sparse Bayesian learning [16, 17]. Motivated by the conceptual similarity of RVM, we used the name relevance singular vector machine (RSVM) in [12] where the prior (2) was used for low-rank matrix estimation from linear measurements. The method in this paper is an outlier robust version of the RSVM for the RPCA model (1), therefore we refer to the new method as robust RSVM (rRSVM).

2.1.2. Prior for combined noise

A standard approach is to model the outlier noise $\mathbf{S}$ and dense noise $\mathbf{N}$ independently. In the Bayesian setting, a sparsity promoting prior is used for the outlier noise [5, 18]. In this paper, we model the dense and sparse outlier noise by a combined noise model as in [13], i.e.

$$S_{ij} + N_{ij} \sim \mathcal{N}(0, \beta_{ij}^{-1}), \tag{4}$$

where $\beta_{ij} > 0$ is the total noise precision of the combined noise component $S_{ij} + N_{ij}$. The motivation of using a combined model instead of an independent treatment stems from the fact that $\mathbf{S}$ and $\mathbf{N}$ need not be separated individually for estimation of $\mathbf{X}$. This approach also reduces the number of model parameters and often improves estimation performance [13]. The combined noise is only approximately sparse and is well modeled using a sparsity promoting prior. We here use the Gamma prior

$$p(\beta_{ij}) = \mathrm{Gamma}(\beta_{ij}|a+1, b) = \frac{b^{a+1}}{\Gamma(a+1)}\beta_{ij}^a e^{-b\beta_{ij}}, \tag{5}$$

for the noise precisions β_{ij}, where $\Gamma(\cdot)$ denotes the Gamma function [19], making the model sparsity promoting.

2.2. Bayesian learning algorithm for rRSVM

A common method for estimating the model parameters is the maximum a-posteriori method

$$\hat{\mathbf{X}}, \theta = \arg\max_{\mathbf{X},\theta} p(\mathbf{X}, \theta|\mathbf{Y}), \tag{6}$$

where $\theta = \{\boldsymbol{\alpha}_L, \boldsymbol{\alpha}_R, \boldsymbol{\beta}\}$. The maximization of (6) is often hard in practice and needs to be performed approximately through e.g. evidence approximation or expectation maximization (EM) [19]. To initialize EM, we make an initial choice of θ. Next, in the expectation step, EM computes the posterior distribution $p(\mathbf{X}|\mathbf{Y}, \theta')$ of $\mathbf{X}$ given the measurements $\mathbf{Y}$ and the latent variables θ' from the previous iteration. In the second step (the maximization step), the latent variables θ are updated by maximizing the EM help function

$$Q(\theta, \theta') = E\left[\log p(\mathbf{Y}, \mathbf{X}|\theta)|\mathbf{Y}, \theta'\right] + \log p(\theta),$$

with respect to θ . The expectation and maximization step is repeated until convergence. An advantage of EM over e.g. evidence approximation is that it has established monotone convergence properties [19], i.e. in each iteration the cost in (6) does not increase.

For Bayesian RPCA (1) with the priors (2) and (4), the posterior distribution $p(\mathbf{X}|\mathbf{Y}, \theta')$ Gaussian with mean

$$\begin{aligned} \mathrm{vec}(\hat{\mathbf{X}}) &= \boldsymbol{\Sigma}\mathbf{B}'\mathrm{vec}(\mathbf{Y}), \\ \boldsymbol{\Sigma} &= \left((\boldsymbol{\alpha}_R' \otimes \boldsymbol{\alpha}_L') + \mathbf{B}'\right)^{-1}, \end{aligned}$$

where $\mathbf{B}' = \mathrm{diag}(\mathrm{vec}(\boldsymbol{\beta}'))$ and $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathrm{vec}(\mathbf{X})$.

The EM help function for our model becomes

$$Q(\theta,\theta') = -\frac{1}{2}\sum_{i,j}\left(\beta_{ij}(Y_{ij}-\hat{X}_{ij})^2 - \log\beta_{ij}\right) - \frac{1}{2}\mathrm{tr}(\hat{\mathbf{X}}^\top\boldsymbol{\alpha}_L\hat{\mathbf{X}}\boldsymbol{\alpha}_R) - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}[(\boldsymbol{\alpha}_R\otimes\boldsymbol{\alpha}_L)+\mathbf{B}]) + \frac{q}{2}\log|\boldsymbol{\alpha}_L| + \frac{p}{2}\log|\boldsymbol{\alpha}_R| + \log p(\theta) + \text{constant},$$

where $\mathbf{B} = \mathrm{diag}(\mathrm{vec}(\boldsymbol{\beta}))$. The priors of the precisions is denoted by $p(\theta)$, i.e.

$$\log\, p(\theta) = \log\, p(\boldsymbol{\alpha}_L) + \log\, p(\boldsymbol{\alpha}_R) + \sum_{\substack{1\le i\le p\\ 1\le j\le q}} \log\, p(\beta_{ij}).$$

Maximizing the EM help function with respect to θ we find the update equations

$$\beta_{ij} = \frac{1+2a}{(Y_{ij}-\hat{X}_{ij})^2 + [\boldsymbol{\Sigma}_\beta]_{ij} + 2b}, \tag{7}$$

$$\boldsymbol{\alpha}_L = \nu\left(\hat{\mathbf{X}}\boldsymbol{\alpha}_R\hat{\mathbf{X}}^\top + \boldsymbol{\Sigma}_L + \epsilon\mathbf{I}_p\right)^{-1}, \tag{8}$$

$$\boldsymbol{\alpha}_R = \nu\left(\hat{\mathbf{X}}^\top\boldsymbol{\alpha}_L\hat{\mathbf{X}} + \boldsymbol{\Sigma}_R + \epsilon\mathbf{I}_q\right)^{-1}, \tag{9}$$

where $[\boldsymbol{\Sigma}_\beta]_{ij}$ denotes the (i,j) component of the matrix $\boldsymbol{\Sigma}_\beta$. The matrices $\boldsymbol{\Sigma}_\beta \in \mathbb{R}^{p\times q}$, $\boldsymbol{\Sigma}_L \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\Sigma}_R \in \mathbb{R}^{q\times q}$ are defined by their elements

$$[\boldsymbol{\Sigma}_\beta]_{ij} = [\boldsymbol{\Sigma}]_{i+p(j-1),i+p(j-1)}, \tag{10}$$

$$[\boldsymbol{\Sigma}_L]_{ij} = \mathrm{tr}(\boldsymbol{\Sigma}(\boldsymbol{\alpha}_R\otimes\mathbf{E}_{ij}^L)), \tag{11}$$

$$[\boldsymbol{\Sigma}_R]_{ij} = \mathrm{tr}(\boldsymbol{\Sigma}(\mathbf{E}_{ij}^R\otimes\boldsymbol{\alpha}_R)), \tag{12}$$

where $\mathbf{E}_{ij}^L \in \mathbb{R}^{p\times p}$ and $\mathbf{E}_{ij}^R \in \mathbb{R}^{q\times q}$ are matrices with a 1 in position (i,j) and zeros otherwise. Typically, the regularization parameters a, b and ϵ are set to small values, e.g. 10^{-4}. In the simulations we initialized the algorithm by setting the matrix precisions to identity matrices and all noise precisions to one. We stopped iterating when the relative difference $||\hat{\mathbf{X}} - \hat{\mathbf{X}}^{(old)}||_F^2/||\hat{\mathbf{X}}^{(old)}||_F^2$ was less than 1%.

The algorithm can be modified for the matrix completion problem by setting the noise precisions of the unobserved components to a low value (corresponding to high noise variance) in each iteration.

3. EXPERIMENTS

We used numerical simulations to evaluate the performance of the algorithms PCP [1], VB [4], EB [5] and rRSVM. First, we generated synthetic test data for (1). We estimated the low-rank matrix $\mathbf{X}$ using the different algorithms and empirically evaluated the Normalized Mean Square Error (NMSE)

$$\mathrm{NMSE} = E[||\mathbf{X}-\hat{\mathbf{X}}||_F^2]/E[||\mathbf{X}||_F^2].$$

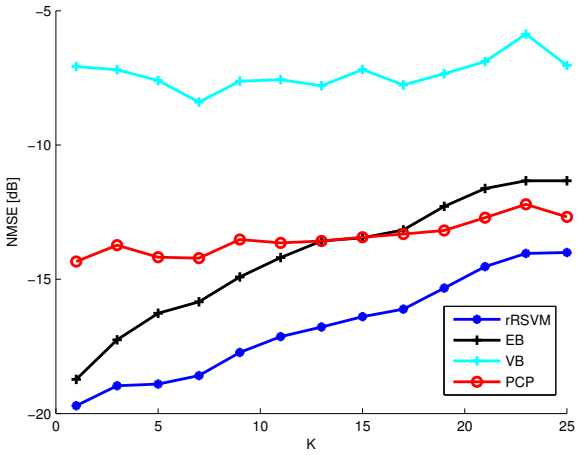

Fig. 1. NMSE vs. number of outliers, K.

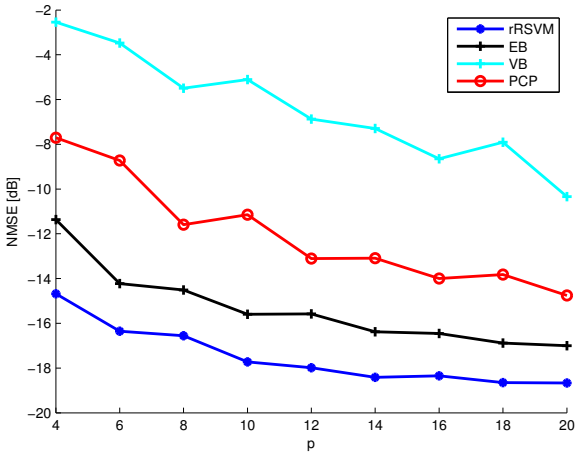

Fig. 2. NMSE vs. p, height of the matrix.

We considered the case where both the rank, sparsity and SNR is unknown. To make a broader comparisson we also compared with the PCP algorithm for which we assumed the SNR to be known a-priori. For PCP we used $\epsilon = \sigma_n\sqrt{pq+\sqrt{8pq}}$ as suggested in [20].

3.1. Synthetic data

To generate synthetic measurements (1), we generated the low rank matrix by setting $\mathbf{X} = \mathbf{AB}$, where the elements of $\mathbf{A} \in \mathbb{R}^{p\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times q}$ were drawn from a $\mathcal{N}(0,1)$ distribution. The sparse matrix $\mathbf{S}$ was generated by selecting the positions of the K non-zero coefficients uniformly at random and drawing their values from $\mathcal{N}(0,1)$. The elements of the dense noise matrix was drawn independently from $\mathcal{N}(0,\sigma_n^2)$, where σ_n^2 is chosen to fix the signal-to-noise ratio (SNR)

$$\mathrm{SNR} = \frac{E[||\mathbf{X}+\mathbf{S}||_F^2]}{E[||\mathbf{N}||_F^2]} = \frac{rpq+K}{pq\sigma_n^2}.$$

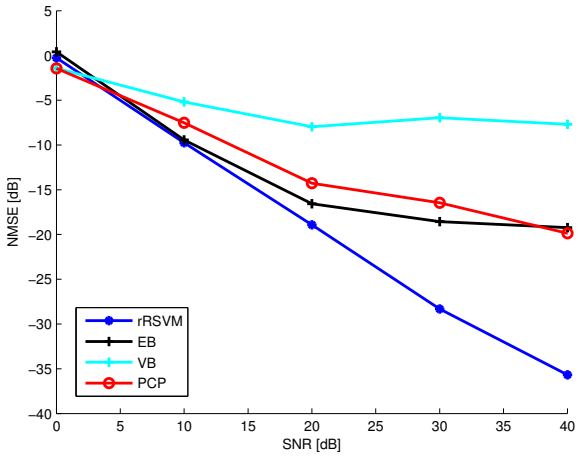

Fig. 3. NMSE vs. SNR.

We evaluated the NMSE over 100 realizations for each parameter value.

We measured how the number of outliers affect the algorithms by setting $p = 10$, $q = 20$, $r = 3$, SNR $= 20$ dB and varying K, the number of outliers. We found that EB and rRSVM gave a lower NMSE than PCP for $K \leq 14$. The NMSE of rRSVM was 2.6 dB lower than that of EB for $K \geq 5$. The NMSE of PCP was 6 dB lower than that of VB. The results are shown in Figure 1. For recovering the sparse component, VB was most efficient followed by rRSVM.

To evaluate the effect of the matrix size, we varied p, the height of the matrix, for $q = 2p$, $r = \lceil 0.15p \rceil$, $K = \lceil 0.05pq \rceil$ and SNR $= 20$ dB. We found that the NMSE of rRSVM was 1.7 to 3.3 dB lower than the NMSE of EB, the NMSE of EB was 2.4 to 5.5 dB lower than the NMSE of PCP and the NMSE of PCP was 4.4 to 6.2 dB lower than the NMSE of VB. The results are shown in Figure 2.

Finally, we measured the sensitivity to noise by setting $p = q = 10, r = 2, K = 5$ and varied the SNR. We found that rRSVM performed best for SNR ≥ 10 dB. For SNR $= 30$ dB, the NMSE of rRSVM was 9.7 dB lower than the NMSE of EB while the NMSE of PCP was 9.5 dB lower than the NMSE of VB. The results are shown in Figure 3.

3.2. MovieLens dataset

The MovieLens 100K dataset [14] consists of 100 000 ratings of 1682 movies by 943 users collected in 1997-1998. Each rating is given by an integer from 1 to 5. The movies can be modeled by certain features, e.g. genre, and each user has preferences based on these features. The preferences of a user can thus be modeled as a *linear combination* of preferences for certain (unknown) features. For a low number of relevant features, the matrix of ratings becomes low-rank.

Some users may have unique preferences for which the low-rank model is ill-suited. There are also examples of so called *schilling attacks* in which users generate ratings in order to manipulate recommendations [21]. Ratings which are not modeled well by a low-rank matrix are often few and can thus be modeled by a sparse matrix, the recommendation problem thus becomes a robust matrix completion problem.

Partition	PCP	VB	EB	rRSVM
u1	65.9	78.5	78.5	22.8
u2	79.7	58.3	58.3	22.5
u3	64.9	40.7	40.7	13.8
u4	58.5	37.2	37.2	13.4
u5	37.5	37.4	37.4	12.3

Table 1. Error when using the first 75 rows and columns of the MovieLens 100K dataset.

To test the algorithms for robust matrix completion, we used the predefined partitions $u1$, $u2$, $u3$, $u4$ and $u5$ of the MovieLens dataset into training and test data. We used only part of the dataset in order to run the algorithms in reasonable time. In simulations we performed full matrix completion on the training set and calculated the (Frobenius) error over the test set, i.e.

$$\text{Error} = \sqrt{\sum_{(i,j)\in\Omega_{\text{test}}} (\hat{X}_{ij} - X_{ij}^{(\text{test})})^2}.$$

We assumed noise-free measurements for PCP.

We found that rRSVM gave a lower error than the other algorithms. The performance of EB and VB was close to identical (differing first in the 6'th decimal place). PCP gave a lower error than EB and VB only for $u1$. The errors are shown in Table 1.

4. CONCLUSION

In this paper we developed a robust Relevance Singular Vector Machine for robust principal component analysis. The algorithm uses matrix precisions to promote low-rank and models the sparse and dense noise as a single noise source. Through Bayesian modeling, we are able to learn all parameters from data and can thus handle situations in which neither the rank, sparsity of outliers nor the noise power is known. Moreover, the Bayesian method provide error estimates of the estimated variables. The algorithm outperforms principal component pursuit, the empirical Bayes and the variational Bayes in numerical experiments with synthetic and real data.

Robust principal component analysis is a relevant problem that appears in many applications. Hence, it is important to develop more accurate methods. In many real world scenarios, such as the MovieLens dataset, neither the rank, sparsity of outliers nor noise power are known a-priori. For this reason, robust Bayesian methods are important for both signal processing and machine learning.

REFERENCES

[1] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, pp. 11, 2011.

[2] Shusen Wang, Dehua Liu, and Zhihua Zhang, "Non-convex relaxation approaches to robust matrix recovery," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press, 2013, pp. 1764–1770.

[3] Tianyi Zhou and Dacheng Tao, "Godec: Randomized low-rank & sparse matrix decomposition in noisy case," *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[4] S.D. Babacan, M. Luessi, R. Molina, and A.K. Katsaggelos, "Sparse bayesian methods for low-rank matrix estimation," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 3964–3977, Aug 2012.

[5] David Wipf, "Non-convex rank minimization via an empirical bayesian approach," *Uncertainty in Artificial Intelligence (UAI), 2012*.

[6] Xinghao Ding, Lihan He, and Lawrence Carin, "Bayesian robust principal component analysis," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3419–3430, 2011.

[7] John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, and Yi Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in neural information processing systems*, 2009, pp. 2080–2088.

[8] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[9] Ricardo Otazo, Emmanuel Candès, and Daniel K Sodickson, "Low-rank plus sparse matrix decomposition for accelerated dynamic mri with separation of background and dynamic components," *Magnetic Resonance in Medicine*, 2014.

[10] Xiaohui Shen and Ying Wu, "A unified approach to salient object detection via low rank matrix recovery," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 853–860.

[11] Morteza Mardani, Gonzalo Mateos, and Georgios B Giannakis, "Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies," *IEEE Transactions on Information Theory*, vol. 59, no. 8, pp. 5186–5205, 2013.

[12] M. Sundin, S. Chatterjee, M. Jansson, and C.R. Rojas, "Relevance singular vector machine for low-rank matrix sensing," in *2014 International Conference on Signal Processing and Communications (SPCOM)*, July 2014, pp. 1–5.

[13] Martin Sundin, Saikat Chatterjee, and Magnus Jansson, "Combined modeling of sparse and dense noise for improvement of relevance vector machine," *arXiv preprint arXiv:1501.02579*, 2015.

[14] Jonathan L Herlocker, Joseph A Konstan, Al Borchers, and John Riedl, "An algorithmic framework for performing collaborative filtering," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 230–237.

[15] Michael E Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[16] David P Wipf and Bhaskar D Rao, "Sparse bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004.

[17] David P Wipf, Bhaskar D Rao, and Srikantan Nagarajan, "Latent variable bayesian models for promoting sparsity," *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 6236–6255, 2011.

[18] Kaushik Mitra, Ashok Veeraraghavan, and Rama Chellappa, "Robust rvm regression using sparse outlier model," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 1887–1894.

[19] Christopher M Bishop, *Pattern recognition and machine learning*, vol. 4, Springer New York, 2006.

[20] Emmanuel J. Candes, Justin K. Romberg, and Terence Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[21] Paul-Alexandru Chirita, Wolfgang Nejdl, and Cristian Zamfir, "Preventing shilling attacks in online recommender systems," in *Proceedings of the 7th annual ACM international workshop on Web information and data management*. ACM, 2005, pp. 67–74.