MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

A Robust Online Subspace Estimation and Tracking Algorithm

Mansour, H.; Jiang, X.

TR2015-027 April 2015

Abstract

In this paper, we present a robust online subspace estimation and tracking algorithm (ROSETA) that is capable of identifying and tracking a time-varying low dimensional subspace from incomplete measurements and in the presence of sparse outliers. Our algorithm minimizes a robust l1 norm cost function between the observed measurements and their projection onto the estimated subspace. The projection coefficients and sparse outliers are computed using ADMM solver and the subspace estimate is updated using a proximal point iteration with adaptive parameter selection. We demonstrate using simulated experiments and a video background subtraction example that ROSETA succeeds in identifying and tracking low dimensional subspaces using fewer iterations than a state of art algorithm.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

A ROBUST ONLINE SUBSPACE ESTIMATION AND TRACKING ALGORITHM

Hassan Mansour

Mitsubishi Electric Research Laboratories
Cambridge, MA 02139
mansour@merl.com

*Xin Jiang**

Electrical and Computer Engineering
Duke University
Durham, NC 27708-0291

ABSTRACT

In this paper, we present a robust online subspace estimation and tracking algorithm (ROSETA) that is capable of identifying and tracking a time-varying low dimensional subspace from incomplete measurements and in the presence of sparse outliers. Our algorithm minimizes a robust ℓ_1 norm cost function between the observed measurements and their projection onto the estimated subspace. The projection coefficients and sparse outliers are computed using ADMM solver and the subspace estimate is updated using a proximal point iteration with adaptive parameter selection. We demonstrate using simulated experiments and a video background subtraction example that ROSETA succeeds in identifying and tracking low dimensional subspaces using fewer iterations than a state of art algorithm.

***Index Terms*—** Online subspace Identification, subspace tracking, low-rank matrix recovery, robust PCA, background subtraction

1. INTRODUCTION

The problem of identifying and tracking low-dimensional subspaces embedded in high dimensional data arises in many applications such as video background subtraction [1], anomaly detection [2], motion segmentation [3], collaborative filtering [4–6], and target localization [7]. For example, the video scene captured by a stationary or moving camera can be separated into a low rank component spanning the subspace that characterizes the background scene, and a sparse component corresponding to moving objects in the video scene.

Classical approaches to low dimensional subspace identification first organize the data into a matrix and then compute basis vectors that span the target subspace using a variety of techniques that involve low rank matrix factorization [5, 6, 8–11]. Robust extensions of these techniques factor the matrix into a low rank component corresponding to the target subspace as well as a sparse component that captures the noise [1, 10, 12, 13]. However, when the dimensionality of the data becomes too large as is the case with recommendation systems that monitor internet traffic [4], or when data arrives in streaming form and latency is an issue, as in the case of high definition video, it becomes necessary to develop online algorithms that can detect and track the target subspace as the data arrives even when the data streams are incomplete and corrupted by sparse noise. Another important benefit to online algorithms arises if the target subspace varies over time, in which case the subspace can no longer be represented by a low rank matrix when the data is grouped into a matrix.

1.1. Related Work

Recently, several algorithms have been proposed to address the online subspace estimation problem from incomplete observations [14–20]. The GROUSE algorithm [14] uses rank-one updates of the estimated subspace on the Grassmannian manifold. The PETRELS algorithm [15] minimizes the geometrically discounted sum of projection residuals on the observed entries per time index, via a recursive procedure with discounting for each row of the subspace matrix. However, neither of these algorithms are designed to be robust to data corruption or non-Gaussian distributions of noise. More recently, the GRASTA algorithm [16], or robust GROUSE, also uses updates on the Grassmannian manifold using a robust ℓ_1-norm cost to recover from outliers in the observations. Other robust online PCA techniques include recursive projection in ReProCS [17], bilinear decomposition [18], and adaptive projected subgradient STAPSM [19]. These algorithms either do not handle missing data or require relatively accurate initial estimates of the target subspace. Finally, our method is closely related to the OR-PCA algorithm [20] which uses alternating minimization to compute the subspace coefficients and sparse outliers followed by stochastic gradient updates of an ℓ_2 regularized least squares cost to track the subspace. Our technique differs from [20] in that we employ ADMM to estimate the subspace coefficients and sparse outliers followed by a proximal gradient update with adaptive step size to track the subspace.

1.2. Contributions

In this paper, we propose a robust online subspace estimation and tracking algorithm (ROSETA) that learns a low dimen-

*The authors contributed equally to this work which was conducted while Xin Jiang was a summer intern at MERL.

sional subspace from incomplete streaming measurements that may be corrupted with non-Gaussian noise. We formally define our problem and set the notation in section 2. Section 3 discusses the details of our algorithm which minimizes a robust ℓ_1 norm misfit function between the observed measurements and their projection onto the estimated subspace to compute the projection coefficients and sparse outliers. The subspace is then updated using proximal point least squares estimation with adaptive parameter selection. Our approach is inspired by the PETRELS algorithm in that it does not restrict the subspace update to the Grassmannian, and by the GRASTA algorithm in the selection of its adaptive step size. Moreover, our algorithm does not require any precomputed initial estimate of the target subspace. Finally, we demonstrate in section 4 the superior performance of our algorithm over GRASTA in identifying and tracking stationary and dynamic subspaces from incomplete measurements and noisy outliers.

2. PROBLEM FORMULATION

2.1. Problem statement

We consider the problem of identifying at every time t an r-dimensional subspace $\mathcal{U}_t$ in $\mathbb{R}^n$ with $r \ll n$ that is spanned by the columns of a rank-r matrix $U_t \in \mathbb{R}^{n\times r}$ from incomplete and noisy measurements

$$b_t = \Omega_t(U_t a_t + s_t), \tag{1}$$

where Ω_t is a selection operator that specifies the observable subset of entries at time t, $a_t \in \mathbb{R}^r$ are the coefficients specifying the linear combination of the columns of U_t, and $s_t \in \mathbb{R}^n$ is a sparse outlier vector.

When the subspace $\mathcal{U}_t$ is stationary, we drop the subscript t from U_t and the problem reduces to robust matrix completion or robust principal component analysis where the task is to separate a matrix $B \in \mathbb{R}^{n\times m}$ into a low rank component UA and a sparse component S using incomplete observations

$$B_\Omega = \Omega(UA + S).$$

Here the columns of the matrices A and S are respectively the vectors a_t and s_t stacked horizontally for all $t \in \{1 \dots m\}$, and the operator Ω specifies the observable entries for the entire matrix B.

2.2. GRASTA

GRASTA addresses the system in (1) using a robust ℓ_1 norm cost to quantify the subspace error. The algorithm proceeds by fixing the U_t and minimizing the augmented Lagrangian

$$\mathcal{L}(s_t, a_t, y_t) = \\ \|s_t\|_1 + y_t^T\left(b_t - \Omega(U_t a_t + s_t)\right) + \tfrac{\mu}{2}\|b_t - \Omega(U_t a_t + s_t)\|_2^2, \tag{2}$$

where y_t is the dual vector, and μ is a regularization constant. The subspace matrix U_t is then updated by taking a gradient step on the Grassmannian geodesic using the augmented Lagrangian (2) as the loss function (See section 3.2.2 of [16] for details). Of particular interest in GRASTA is the selection of an adaptive step size that leverages precise convergence for a stationary subspace as well as fast adaptation to a changing subspace. We develop a similar adaptive parameter selection strategy in our proposed approach to achieve the precision and adaptability goals at an even faster rate.

3. ROBUST ONLINE SUBSPACE ESTIMATION AND TRACKING

We describe in this section our proposed robust online subspace estimation and tracking algorithm (ROSETA).

3.1. Augmented Lagrangian with proximal point

ROSETA aims to minimize an augmented Lagrangian with a robust ℓ_1 norm cost in addition to a smoothing term that maintains the proximity of the update to the pervious subspace estimate over the variables (U_t, s_t, a_t, y_t). Our objective cost is given by the following expression

$$\mathcal{L}'(U_t, s_t, a_t, e_t, y_t) = \|s_t\|_1 + y_t^T\left(b_t - (U_t a_t + s_t + e_t)\right) \\ + \tfrac{\mu}{2}\|b_t - (U_t a_t + s_t + e_t)\|_2^2 + \tfrac{\mu}{2}\|U_t - U_{t-1}\|_2^2, \tag{3}$$

where e_t is supported on the complement of Ω_t, hereby denoted Ω_t^c, such that $\Omega_t(e_t) = 0$ and $\Omega_t^c(e_t) = -\Omega_t^c(U_t a_t)$.

Note that the above term in (3) is non convex in the variables U_t and a_t. Therefore, we follow the PETRELS and GRASTA approach of alternating the minimization over the variables (s_t, a_t, y_t) on the one hand, and U_t on the other. Notice that by fixing U_t, the minimizers of (3) and (2) are equal, i.e.

$$(s_t, a_t, y_t) = \arg\min_{s,a,y} \mathcal{L}(s,a,y) = \arg\min_{s,a,e,y} \mathcal{L}'(U_{t-1}, s, a, e, y). \tag{4}$$

The variable U_t is then updated by taking a gradient step to minimize the function

$$\mathcal{F}(U_t) = \frac{1}{2}\|b_t - (U_t a_t + s_t + e_t)\|_2^2 + \frac{1}{2}\|U_t - U_{t-1}\|_2^2 \tag{5}$$

using an adaptive μ.

3.2. ROSETA

In the first stage, ROSETA uses an ADMM algorithm [21] to solve (4). The variables a_t, s_t, and y_t are computed by iterating until a stopping criterion is met the following sequence of updates:

$$\begin{array}{rcl} a_t^k &=& U_{t-1}^\dagger\left(b_t - s_t^{k-1} - e_t^{k-1} + \frac{1}{\mu_{t-1}} y_t^{k-1}\right) \\ e_t^k &=& -\Omega_t^c\left(U_{t-1} a_t^k\right) \\ s_t^k &=& \mathcal{S}_{\frac{1}{\mu_{t-1}}}\left(b_t - U_{t-1} a_t^k - e_t^k - \frac{1}{\mu_{t-1}} y_t^{k-1}\right) \\ y_t^k &=& y_t^{k-1} + \mu_{t-1}\left(b_t - U_{t-1} a_t^k - s_t^k - e_t^k\right), \end{array} \tag{6}$$

where $\mathcal{S}_\tau(x) = \text{sign}(x)\cdot\max\{|x|-\tau, 0\}$ denotes the element-wise soft thresholding operator with threshold τ, k indicates the iteration number, and † is the Moore-Penrose pseudo-inverse of a matrix.

In the second stage, the variable U_t is computed by minimizing (5) using the update

$$U_t = \frac{\mu_{t-1}}{\mu_t}\left(U_{t-1} + (b_t - s_t - e_t)a_t^T\right)\left(I_r + a_t a_t^T\right)^{-1}, \tag{7}$$

where I_r is an $r \times r$ identity matrix, and μ_t is updated adaptively as will be discussed in the following section.

3.3. Adaptive parameter selection

Inspired by the adaptive step size selection in GRASTA, we developed a corresponding adaptive parameter for ROSETA.

Contrary to GRASTA, our parameter is specifically the regularizer μ_t. The regularizer μ_t controls the speed of convergence of the subspace estimate. In particular, a smaller value of μ allows for faster adaptability of U_t to a changing subspace (larger descent direction), whereas a larger value of μ only permits a small variation in U_t. Consider the descent direction

$$D_t = \left(U_{t-1} + (b_t - s_t - e_t)a_t^T\right)\left(I_r + a_t a_t^T\right)^{-1} - U_{t-1} \tag{8}$$

and compute its projection onto the orthogonal complement of the previous subspace estimate to obtain the subspace update

$$G_t = (I - U_{t-1}U_{t-1}^\dagger)D_t. \tag{9}$$

The parameter μ_t can then be updated according to

$$\mu_t = \frac{C2^{-l}}{1+\eta_t}, \tag{10}$$

where $\eta_t = \eta_{t-1} + \text{sigmoid}\left(\frac{\langle G_{t-1}, G_t\rangle}{\|G_{t-1}\|_F\|G_t\|_F}\right)$, and $l \in \{-1, 0, 1, 2\}$ is set according to pre specified thresholds for η_t. Here $\text{sigmoid}(x) = f + 2f/(1+e^{10x})$, for some predefined f.

Similar to GRASTA, the intuition behind choosing such an update rule comes from the idea that if two consecutive subspace updates G_{t-1} and G_t have the same direction, i.e. $\langle G_{t-1}, G_t\rangle > 0$, then the target subspace is still far from the current subspace estimate. Consequently, the new μ_t should be smaller to allow for fast adaptability which is achieved by increasing η_t. Similarly, when $\langle G_{t-1}, G_t\rangle < 0$, the subspace update seems to bounce around the target subspace and hence a larger μ_t is needed. Note that when the product of the norms of the subspace updates ($\|G_{t-1}\|_F \cdot \|G_t\|_F$) is too small, e.g. smaller than 10^{-6}, we assume that our subspace estimate is close to the target and we force η_t to decrease by the magnitude of the sigmoid. The ROSETA algorithm is summarized in Algorithm 1.

Algorithm 1 Robust Subspace Estimation and Tracking

Input Sequence of measurements $\{b_t\}$, η_{LOW}, η_{HIGH}
Output Sequences $\{U_t\}$, $\{a_t\}$, $\{s_t\}$
Initialize U_0, μ_0, η_0, $l = 0$
for $t = 1 \dots N$ **do**
  Solve (4) **using ADMM:**
  while not converged **do**
    $a_t^k = U_{t-1}^\dagger\left(b_t - s_t^{k-1} - e_t^{k-1} + \frac{1}{\mu_{t-1}}y_t^{k-1}\right)$
    $e_t^k = -\Omega_t^c\left(U_{t-1}a_t^k\right)$
    $s_t^k = \mathcal{S}_{\frac{1}{\mu_{t-1}}}\left(b_t - U_{t-1}a_t^k - e_t^k - \frac{1}{\mu_{t-1}}y_t^{k-1}\right)$
    $y_t^k = y_t^{k-1} + \mu_{t-1}\left(b_t - U_{t-1}a_t^k - s_t^k - e_t^k\right))$
  end while
  Compute adaptive step-size:
    $D_t = \left(U_{t-1} + (b_t - s_t - e_t)a_t^T\right)\left(I_r + a_t a_t^T\right)^{-1}$
    $G_t = (I - U_{t-1}U_{t-1}^\dagger)D_t$
    $\eta_t = \eta_{t-1} + \text{sigmoid}\left(\frac{\langle G_{t-1}, G_t\rangle}{\|G_{t-1}\|_F\|G_t\|_F}\right)$
    $l = \begin{cases} \min\{l+1, 2\}, & \text{if } \eta_t \geq \eta_{\text{HIGH}} \\ \max\{l-1, -1\}, & \text{if } \eta_{\text{LOW}} \leq \eta_t \leq \eta_{\text{HIGH}} \end{cases}$
    $\mu_t = \frac{C2^{-l}}{1+\eta_t}$
  Update subspace estimate:
    $U_t = \frac{\mu_{t-1}}{\mu_t}\left(U_{t-1} + (b_t - s_t - e_t)a_t^T\right)\left(I_r + a_t a_t^T\right)^{-1}$
end for

4. NUMERICAL EXAMPLES

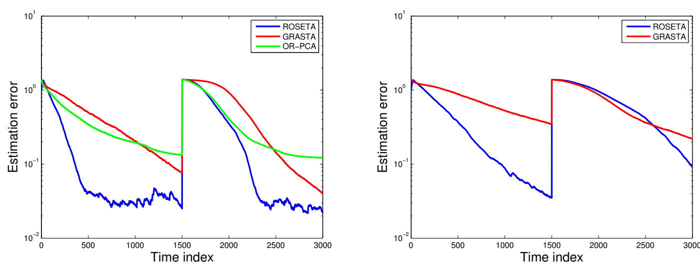
We tested the performance of ROSETA in tracking synthetically generated stationary as well as rotating subspaces in the presence of sparse outliers. We also tested its performance in extracting the stationary background of a video sequence and separating out the moving foreground objects. We also compare the performance of ROSETA to that of GRASTA [16] and ORPCA [20]. We note here that we first implemented ORPCA according to its description in [20] but found that the algorithm fails converge. Therefore, we modified the algorithm by changing $\tilde{a}_j$ to a_j in the basis update step [20, equation (9)] to allow convergence, and applying a discount factor $\gamma = 0.5$ to past observations of the matrices A_t and B_t to speed up convergence. Also note that ORPCA requires fully sampled measurements so we do not present its performance in the subsampled case.The distance between the estimated subspace and the target subspace is measured using the relative error between the projection matrices of the estimated U_t and the target U_t^* given by

$$\text{Error}_t = \frac{\|U_t U_t^\dagger - U_t^* U_t^{*\dagger}\|_F}{\|U_t^* U_t^{*\dagger}\|_F} \tag{11}$$

In all our experiments, we use the following parameters for ROSETA: $C = 8$, $\eta_0 = 99$, $\mu_0 = \frac{C}{1+\eta_0}$, $\eta_{\text{LOW}} = 50$, $\eta_{\text{HIGH}} = 100$, $f = 100$. We also add the condition that if $(\|G_{t-1}\|_F \cdot \|G_t\|_F) < 10^{-6}$, then l is updated such that $l = \begin{cases} \max\{l-1, -1\}, & \text{if } \eta_t \leq \eta_{\text{LOW}} \\ \min\{l+1, 2\}, & \text{if } \eta_{\text{LOW}} \leq \eta_t \leq \eta_{\text{HIGH}} \end{cases}$.

4.1. Synthetic examples

In the first experiment, we simulate streaming measurements from one stationary subspace followed by a sudden jump to a second subspace. We generate 3000 random streaming measurements $\{b_t\}_{t=1:3000}$ from two stationary subspaces each of dimension 20 in $\mathbb{R}^{500}$ spanned by the columns of two random matrices $U^{*(1)}, U^{*(2)} \in \mathbb{R}^{500\times 20}$ each having orthogonal columns. The first 1500 measurements belong to the subspace spanned by $U^{*(1)}$ and the second 1500 belong to the subspace spanned by $U^{*(2)}$, such that, $\{b_t\}_{1:1500} = U^{*(1)}a_t + s_t$ and $\{b_t\}_{1501:3000} = U^{*(2)}a_t + s_t$, where a_t are Gaussian random vectors in $\mathbb{R}^{20}$, and $s_t \in \mathbb{R}^{500}$ are sparse outlier vectors with nonzero Gaussian random coefficients in 20% of their entries. Fig. 1(a) compares the subspace estimation error averaged over 10 runs of ROSETA, GRASTA, and ORPCA. Fig. 1(b) illustrates the estimation error for both algorithms when every column in b_t is subsampled by 50%. Notice how in both cases, the estimation error of ROSETA (blue line) decreases faster than that of GRASTA (red line).



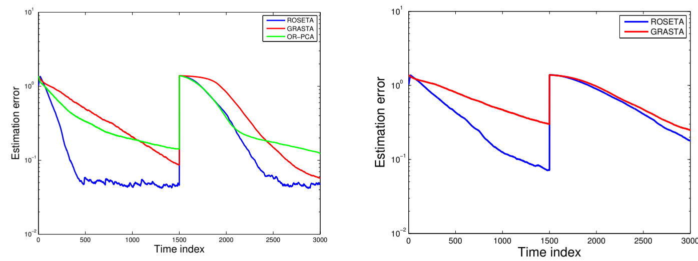
(a) (b)

Fig. 1: Subspace estimation error averaged over 10 runs of ROSETA (blue), GRASTA (red), and ORPCA (green) in identifying a stationary subspace in $\mathbb{R}^{500\times 20}$ corrupted by 20% sparse outliers with a sudden jump at $t = 1500$ from (a) fully sampled measurements, and (b) 50% subsampled measurements.

In the second experiment, we allow the target subspaces to rotate over time by multiplying $U^{*(1)}$ and $U^{*(2)}$ by a skew symmetric matrix R and update U_t^* according to $U_t^* = (I + \delta R)U_{t-1}^*$, such that, $U_1^* = U^{*(1)}$ and $U_{1501}^* = U^{*(2)}$. Here, we choose $\delta = 10^{-2}$. The subspace estimation errors of ROSETA, GRASTA, and ORPCA are shown in Figs. 2(a-b) for the fully sampled and 50% subsampled measurements.

4.2. Video background subtraction

Next, we consider the online video background subtraction problem. We are given a video sequence captured by a stationary camera. The video scene is generally stationary except for foreground moving objects. If we vectorize the video frames and stack them into a matrix, the resulting matrix can be decomposed into a low rank component corresponding to the background and a sparse component corresponding to the foreground moving objects. We compare the performance of ROSETA and GRASTA in extracting the background of the Shopping Mall video sequence[1]. Every video frame is composed of 320×256 pixels. We choose a rank 5 for the background subspace and initialize U_0 for both algorithms to an 81920×5 Gaussian random matrix. Fig. 3 demonstrates ROSETA's performance in extracting the video background compared to GASTA. It can be seen that ROSETA succeeds in capturing the video background much earlier in the video sequence (by frame 41) compared to GRASTA (frame 152).



(a) (b)

Fig. 2: Subspace estimation error averaged over 10 runs of ROSETA (blue), GRASTA (red), and ORPCA (green) in identifying a rotating subspace in $\mathbb{R}^{500\times 20}$ corrupted by 20% sparse outliers with a sudden jump at $t = 1500$ from (a) fully sampled, and (b) 50% subsampled measurements.

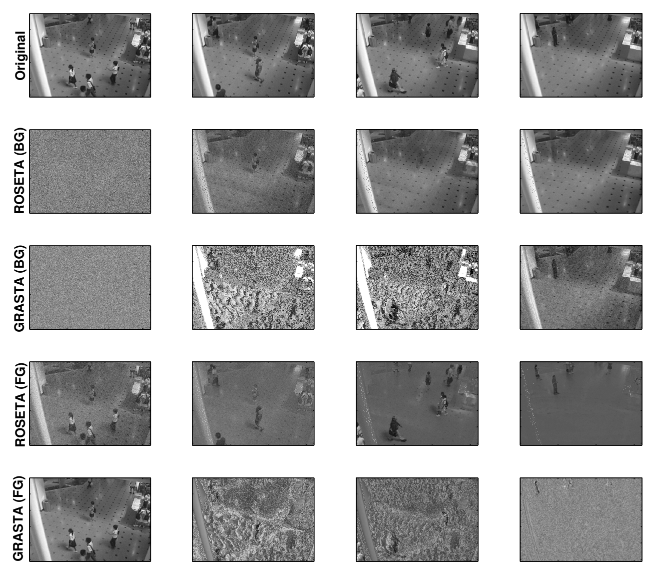


Fig. 3: Background subtraction of frames 1, 41, 92, and 152 from the Shopping Mall sequence. Row one shows the original four frames. Rows two and three show the background extracted by ROSETA and GRASTA, respectively. Rows four and five show the foreground extracted by ROSETA and GRASTA, respectively.

[1] Available from: http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html

5. REFERENCES

[1] J. Wright, A. Ganesh, S. Rao, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in Neural Information Processing Systems 22*, 2009.

[2] T. Ahmed, M. Coates, and A. Lakhina, "Multivariate online anomaly detection using kernel recursive least squares.," in *INFOCOM*. 2007, pp. 625–633, IEEE.

[3] E. Learned-Miller M. Narayana, A. Hanson, "Coherent motion segmentation in moving camera videos using optical flow orientations," in *ICCV*, 2013.

[4] S. Funk, "Netflix update: Try this at home," http://sifter.org/ simon/journal/20061211.html, December 2006.

[5] D. Gross, "Recovering Low-Rank Matrices From Few Coefficients in Any Basis," *IEEE Transactions on Information Theory*, vol. 57, pp. 1548–1566, 2011.

[6] Benjamin Recht and Christopher Ré, "Parallel stochastic gradient algorithms for large-scale matrix completion," *Mathematical Programming Computation*, 2013.

[7] S. Shahbazpanahi, S. Valaee, and M.H. Bastani, "Distributed source localization using esprit algorithm," *IEEE Transactions on Signal Processing*, vol. 49, no. 10, pp. 2169–2178, Oct 2001.

[8] M. Fazel, *Matrix rank minimization with applications*, Ph.D. thesis, Stanford University, 2002.

[9] B. Recht, M Fazel, and P.A. Parrilo, "Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization.," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[10] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, June 2011.

[11] A.Y. Aravkin, R. Kumar, H. Mansour, B. Recht, and F.J. Herrmann, "A robust svd free approach to matrix completion, with applications to interpolation of large scale data," *preprint. http://arxiv.org/abs/1302.4886*, 2013.

[12] H. Mansour and A. Vetro, "Video background subtraction using semi-supervised robust matrix completion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. May 2014, pp. 6528–6532, IEEE.

[13] Aleksandr Y. Aravkin, Stephen Becker, Volkan Cevher, and Peder Olsen, "A variational approach to stable principal component pursuit," *To appear in UAI Proceedings 2014*, 2014.

[14] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2010, pp. 704–711.

[15] Y. Chi, Y. C. Eldar, and A. R. Calderbank, "Petrels: Parallel subspace estimation and tracking by recursive least squares from partial observations.," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5947–5959, 2013.

[16] J. He, L. Balzano, and J. C. S. Lui, "Online robust subspace tracking from partial information," *preprint, http://arxiv.org/abs/1109.3827*, 2011.

[17] Chenlu Qiu and Namrata Vaswani, "Reprocs: A missing link between recursive robust pca and recursive sparse recovery in large but correlated noise," *CoRR*, vol. abs/1106.3286, 2011.

[18] Gonzalo Mateos and Georgios B. Giannakis, "Robust pca as bilinear decomposition with outlier-sparsity regularization," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5176–5190, 2012.

[19] S. Chouvardas, Y. Kopsinis, and S. Theodoridis, "An adaptive projected subgradient based algorithm for robust subspace tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 5497–5501.

[20] Jiashi Feng, Huan Xu, and Shuicheng Yan, "Online robust pca via stochastic optimization," in *Advances in Neural Information Processing Systems 26*, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, Eds., pp. 404–412. 2013.

[21] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.