Given the noise is Gaussian, Z_Δ is of order $O(\sqrt{N \ln N})$ [17], $\mathrm{Re}\left\{\frac{Z_\Delta}{b_\Delta}\right\}$ is $O\left(N^{(-\frac{1}{2})}\sqrt{\ln N}\right)$, whereas $\left|\frac{Z_\Delta}{b_\Delta}\right|^2$ is of order $O(N^{(-1)} \ln N)$. Hence, ignoring the low order term in (35), we have

$$\left|1 + \frac{\alpha - \Delta}{b_\Delta} Z_\Delta\right|^2 \approx 1 + 2(\alpha - \Delta)\mathrm{Re}\left\{\frac{Z_\Delta}{b_\Delta}\right\}. \tag{37}$$

Substituting (37) into (34) and expanding yields

$$|Y_\Delta|^2 \approx \frac{|b_\Delta|^2}{(\alpha-\Delta)^2} + 2\frac{|b_\Delta|^2}{\alpha - \Delta}\mathrm{Re}\left\{\frac{Z_\Delta}{b_\Delta}\right\}. \tag{38}$$

ACKNOWLEDGMENT

The authors sincerely appreciate the reviewers for their contributive comments in making the paper more appealing.

REFERENCES

[1] D. C. Rife and R. R. Boorstyn, "Single tone parameter estimation from discrete-time observations," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 591–598, 1974.
[2] S. Nandi, S. K. Iyer, and D. Kundu, "Estimation of frequencies in presence of heavy tail errors," *Statist. Probab. Lett.*, no. 8, pp. 265–282, 2002.
[3] A. M. Walker, "On the estimation of a harmonic component in a time series with stationary independent residuals," *Biometrika*, vol. 58, no. 1, pp. 21–36, 1971.
[4] A. Mitra and D. Kundu, "Consistent method for estimating sinusoidal frequencies: A non-iterative approach," *J. Statist. Comput. Simulat.*, vol. 58, pp. 171–194, 1997.
[5] S. Kay, "A fast and accurate single frequency estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1987–1990, 1989.
[6] U. Mengali and M. Morelli, "Data-aided frequency estimation for burst digital transmission," *IEEE Trans. Commun.*, vol. 45, no. 1, pp. 23–25, 1997.
[7] A. B. Awoseyila, C. Kasparis, and B. G. Evans, "Improved single frequency estimation with wide acquisition range," *Electron. Lett.*, vol. 44, no. 3, pp. 245–247, 2008.
[8] B. Volcker and P. Handel, "Frequency estimation from proper sets of correlations," *IEEE Trans. Signal Process.*, vol. 50, no. 4, pp. 791–802, Apr. 2002.
[9] T. Brown and W. Wang, "An iterative algorithm for single-frequency estimation," *IEEE Trans. Signal Process.*, vol. 50, no. 11, pp. 2671–2682, 2002.
[10] B. Chen and Y. R. Gel, "Autogressive frequency detection using regularized least squares," *J, Multivatiate Anal.*, pp. 1712–1727, 2010.
[11] V. F. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophys. J. R. Astron. Soc.*, vol. 33, pp. 347–366, 1973.
[12] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," in *Proc. RADC Spectrum Estimation Workshop*, Griffith AFB, NY, 1979, pp. 243–258, .
[13] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.
[14] B. G. Quinn, "Estimating frequency by interpolation using Fourier coefficients," *IEEE Trans. Signal Process.*, vol. 42, no. 5, pp. 1264–1268, May 1994.
[15] B. G. Quinn, "Estimation of frequency, amplitude, and phase from the DFT of a time series," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 814–817, Mar. 1997.
[16] B. G. Quinn and E. J. Hannan, *The Estimation and Tracking of Frequency*. New York: Cambridge Univ. Press, 2001.
[17] E. Aboutanios and B. Mulgrew, "Iterative frequency estimation by interpolation on Fourier coefficients," *IEEE Trans. Signal Process.*, vol. 53, no. 4, pp. 1237–1242, Apr. 2005.
[18] B. G. Quinn, "Recent advances in rapid frequency estimation," *Digit. Signal Process.*, vol. 19, no. 6, pp. 942–948, Dec. 2009.
[19] E. Aboutanios, "Generalised DFT-based estimators of the frequency of a complex exponential in noise," in *Proc. 3rd Int. Congr. Image Signal Process.*, Oct. 16–18, 2010, vol. 6, pp. 2998–3002.
[20] Y. V. Zakharov and T. C. Tozer, "Frequency estimator with dichotomous search of penodogram peak," *Electron. Lett.*, vol. 35, no. 19, pp. 1608–1609, 1999.
[21] E. Aboutanios, "A modified dichotomous search frequency estimator," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 186–188, 2004.
[22] C. Yang and G. Wei, "Frequency estimator by LS approximation of narrowband spectrum," *IEICE Trans. Fundam. Electron., Commun., Comput. Sci.*, vol. E93-A, no. 8, pp. 1553–1555, 2008.

Constrained Cramér–Rao Bound on Robust Principal Component Analysis

Gongguo Tang and Arye Nehorai

***Abstract*—We investigate the behavior of the mean-square error (MSE) of low-rank and sparse matrix decomposition, in particular the special case of the robust principal component analysis (RPCA), and its generalization matrix completion and correction (MCC). We derive a constrained Cramér–Rao bound (CRB) for any locally unbiased estimator of the low-rank matrix and of the sparse matrix. We analyze the typical behavior of the constrained CRB for MCC where a subset of entries of the underlying matrix are randomly observed, some of which are grossly corrupted. We obtain approximated constrained CRBs by using a concentration of measure argument. We design an alternating minimization procedure to compute the maximum-likelihood estimator (MLE) for the low-rank matrix and the sparse matrix, assuming knowledge of the rank and the sparsity level. For relatively small rank and sparsity level, we demonstrate numerically that the performance of the MLE approaches the constrained CRB when the signal-to-noise-ratio is high. We discuss the implications of these bounds and compare them with the empirical performance of the accelerated proximal gradient algorithm as well as other existing bounds in the literature.**

***Index Terms*—Accelerated proximal gradient algorithm, constrained Cramér–Rao bound, matrix completion and correction, maximum likelihood estimation, mean-square error, robust principal component analysis.**

I. INTRODUCTION

The classical principal component analysis (PCA) [1] is arguably one of the most important tools in high dimensional data analysis. However, the classical PCA is not robust to gross corruptions on even only a few entries of the underlying low-rank matrix L containing the principal components. The robust principal component analysis (RPCA) [2]–[4] models the gross corruptions as a sparse matrix S superpositioned on the low-rank matrix L. The authors of [2]–[4] presented various incoherence conditions under which the low-rank matrix L and the sparse matrix S can be accurately decomposed by solving a convex program, the principal component pursuit:

$$\min_{L,S} \|L\|_* + \lambda\|S\|_1 \text{ subject to } Y = L + S \tag{1}$$

Manuscript received November 05, 2010; revised April 18, 2011; accepted June 21, 2011. Date of publication July 14, 2011; date of current version September 14, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Benoit Champagne. This work was supported by ONR Grant N000140810849, and NSF Grants CCF-1014908 and CCF-0963742.

The authors are with the Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO 63130-1127 USA (e-mail: gt2@ese.wustl.edu; nehorai@ese.wustl.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2011.2161984

where Y is the observation matrix, λ is a tuning parameter, and $\|\cdot\|_*$ and $\|\cdot\|_1$ denote the nuclear norm and the sum of the absolute values of the matrix entries, respectively. Many algorithms have been designed to efficiently solve the resulting convex program, e.g., the singular value thresholding [5], the alternating direction method [6], the accelerated proximal gradient (APG) method [7], the augmented Lagrange multiplier method [8], and the Bayesian robust principal component analysis [9]. Some of these algorithms also apply when the observation matrix is corrupted by small, entry-wise noise [4], [9].

In this paper, we consider a general model where we observe (L, S) through the linear measurement mechanism:

$$\boldsymbol{y} = \mathcal{A}(L+S) + \boldsymbol{w}. \tag{2}$$

Here, the noise vector $\boldsymbol{w}$ follows Gaussian distribution $\mathcal{N}(0, \Sigma)$ and $\mathcal{A} : \mathbb{R}^{n\times p} \rightarrow \mathbb{R}^m$ is a linear operator. In particular, when $\mathcal{A}$ is an entry-selection operator, we have the following observation model:

$$Y_{kl} = L_{kl} + S_{kl} + W_{kl}, (k,l) \in \Omega \tag{3}$$

which we name as the *matrix completion and correction (MCC)* model. Basically, for the low-rank matrix L, we observe a subset of its entries indexed by Ω, and among the observed entries, some of them have gross errors, while all of them are corrupted by Gaussian noise. Without loss of generality, we always assume the support $\mathcal{S}$ for the sparse matrix S satisfies $\mathcal{S} \subset \Omega$. A practical example for this observation model is that we transmit a few entries of a low-rank data matrix (e.g., an image compressed by truncating its SVD) through a communication channel that introduces both small random errors and gross transmission errors, and we do not know which entries are grossly corrupted. The following RPCA model with Gaussian noise is a special case of the MCC model (3) with $\Omega = [n] \times [p]$:

$$Y = L + S + W. \tag{4}$$

Here, $[n] \times [p] = \{(k,l) : 1 \le k \le n, 1 \le l \le p\}$. When $S = 0$ in (3), we get yet another special case, the matrix completion problem:

$$Y_{kl} = L_{kl} + W_{kl}, (k,l) \in \Omega. \tag{5}$$

We investigate the behavior of the mean-square error (MSE) in estimating L and S under the unbiased condition. More specifically, we derive a constrained Cramér–Rao bound (CRB) on the MSE that applies to any *locally* unbiased estimator.

The paper is organized as follows. In Section II, we introduce model assumptions for (2) and the constrained Cramér-Rao bound. In Section III, we derive the constrained CRB on MSE of any locally unbiased estimator. Section IV is devoted to the probabilistic analysis of the constrained CRB. We present the alternating minimization implementation of the MLE for the MCC model in Section V. In Section VI, the constrained CRB is compared with the empirical performance of the APG and the MLE. The paper is concluded in Section VII.

II. Model Assumptions and the Constrained Cramér–Rao Bound

We introduce model assumptions and preliminaries on the constrained Cramér–Rao bound in this section. Suppose we have a low-rank and sparse matrix pair $(L, S) \in \mathcal{X}_{r,s}$, where the parameter space

$$\mathcal{X}_{r,s} \stackrel{\text{def}}{=} \left\{(L,S) \in \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times p} : \operatorname{rank}(L) \le r, \|S\|_0 \le s\right\}. \tag{6}$$

Here, $\|S\|_0$ counts the number of nonzero elements in the matrix S. We use $\mathcal{S} = \operatorname{supp}(S) = \{(i,j) \in [n] \times [p] : S_{ij} \neq 0\}$ to denote the support of the sparse matrix S. For any matrix $X \in \mathbb{R}^{n\times p}$, we use $\mathbf{x} = \operatorname{vec}(X)$ to denote the vector obtained by stacking the columns of X into a single column vector. Without introducing any ambiguity we identify $(\mathbf{l}, \mathbf{s})$ with (L, S) and write $(\mathbf{l}, \mathbf{s}) \in \mathcal{X}_{r,s}$, where $\mathbf{l} = \operatorname{vec}(L)$ and $\mathbf{s} = \operatorname{vec}(S)$.

It is convenient to rewrite (2) as the following matrix-vector form:

$$\boldsymbol{y} = \boldsymbol{A}(\mathbf{l}+\mathbf{s}) + \boldsymbol{w} \tag{7}$$

where $\boldsymbol{A}$ is the matrix corresponding to the linear operator $\mathcal{A}$. Therefore, the measurement vector $\boldsymbol{y}$ follows $\mathcal{N}(\boldsymbol{A}(\mathbf{l}+\mathbf{s}), \Sigma)$. Our goal is to derive lower bounds on the MSE for any unbiased estimator $(\hat{\mathbf{l}}(\boldsymbol{y}), \hat{\mathbf{s}}(\boldsymbol{y}))$ that infers the deterministic parameter $(\mathbf{l}, \mathbf{s}) \in \mathcal{X}_{r,s}$ from $\boldsymbol{y}$.

In this paper, we are concerned with locally unbiased estimators. The local unbiased condition imposes the unbiased constraint on parameters in the neighborhood of a single point. More precisely, we require

$$\mathbb{E}\left\{\hat{\mathbf{l}}(\boldsymbol{y})\right\} = \mathbf{l},\ \ \mathbb{E}\left\{\hat{\mathbf{s}}(\boldsymbol{y})\right\} = \mathbf{s},\ \ \forall (\mathbf{l},\mathbf{s}) \in B_\varepsilon(\mathbf{l}_0, \mathbf{s}_0) \cap \mathcal{X}_{r,s} \tag{8}$$

where $B_\varepsilon(\mathbf{l}_0, \mathbf{s}_0) = \{(\mathbf{l},\mathbf{s}) \in \mathbb{R}^{np} \times \mathbb{R}^{np} : \|\mathbf{l}-\mathbf{l}_0\|_2^2 + \|\mathbf{s}-\mathbf{s}_0\|_2^2 \le \varepsilon^2\}$. Here, the expectation $\mathbb{E}\{\cdot\}$ is taken with respect to the noise.

Besides linear estimators, it is usually very difficult to verify, either theoretically or computationally, the unbiasedness of an estimator. However, universal lower bounds such as the CRB, which require the unbiasedness of the estimators, are still popular system performance benchmarks, mainly due to three reasons. First, the maximum likelihood estimator (MLE) is asymptotically unbiased under mild conditions. Second, biased estimators are suboptimal when the signal-to-noise-ratio (SNR) is high. So the CRB serves as a lower bound also for biased estimators in the high SNR region. Last but not least, the CRB allows us to express system performance in an analytical form, which is important for optimal system design.

We present the constrained CRB [10] on the error covariance matrix. We denote $\mathbf{x} = [\mathbf{l}^T \mathbf{s}^T]^T$ and identify it with $(\mathbf{l}, \mathbf{s})$. The error covariance matrix and the scalar MSE are defined as

$$\Sigma_{\mathbf{x}} \stackrel{\text{def}}{=} \mathbb{E}\left\{(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^T\right\} \tag{9}$$

and

$$\mathrm{MSE}_{\mathbf{x}} \stackrel{\text{def}}{=} \mathbb{E}\|\hat{\mathbf{x}}-\mathbf{x}\|_2^2 = \operatorname{tr}(\Sigma_{\mathbf{x}}) \tag{10}$$

respectively, for any unbiased estimator $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\boldsymbol{y})$. Then, the error covariance matrix for any locally unbiased estimator $\hat{\mathbf{x}}$ satisfies [10, Lemma 2]

$$\Sigma_{\mathbf{x}} \ge P[P^T J_{\mathbf{x}} P]^\dagger P^T, \tag{11}$$

as long as the $2np \times 2np$ Fisher information matrix

$$J_{\mathbf{x}} = -\mathbb{E}\{\nabla^2 \ln f_{\mathbf{x}}\} \tag{12}$$

has bounded elements, and $\mathrm{rank}(P^T J_{\mathbf{x}} P) = \mathrm{rank}(P)$. Here, $f_{\mathbf{x}}$ is the probability density function for given parameter $\mathbf{x}$ and P is any $2np \times N$ matrix whose column space equals $\mathrm{span}\{\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_N\}$ with $\{\boldsymbol{\nu}_i\}_{i=1}^N$ N linearly independent vectors such that $\mathbf{x} + \Delta_i \boldsymbol{\nu}_i \in \mathcal{X}_{r,s}$ for all sufficiently small $\Delta_i > 0, i = 1, \ldots, N$.

III. The Constrained Cramér–Rao Bound for Matrix Completion and Correction

In this section, we apply (11) to derive the constrained CRB on the MSE for model (7). For any $\mathbf{x} \in \mathcal{X}_{r,s}$, let N linearly independent directions $\{\boldsymbol{\nu}_i\}_{i=1}^N$ satisfy $\{\mathbf{x}_i = \mathbf{x} + \Delta_i \boldsymbol{\nu}_i\}_{i=1}^N \subset \mathcal{X}_{r,s}$ for sufficiently small $\Delta_i > 0$. To obtain a tighter lower bound according to (11), the following lemma [11] implies that we should select $\{\boldsymbol{\nu}_i\}_{i=1}^N$ to span a subspace as large as possible:

Lemma 1 [11]: Suppose $P_1 \in \mathbb{R}^{2np \times N_1}$ and $P_2 \in \mathbb{R}^{2np \times N_2}$ are two full rank matrices with $\mathrm{span}\{P_1\} \subseteq \mathrm{span}\{P_2\}$. If $P_2^T J_{\mathbf{x}} P_2$ is positive definite, then $P_1 [P_1^T J_{\mathbf{x}} P_1]^{\dagger} P_1^T \leq P_2 [P_2^T J_{\mathbf{x}} P_2]^{\dagger} P_2^T$.

We first find the directions for L. Suppose that $L = U_0 \Lambda_0 V_0^T$ is the singular value decomposition of L with $U_0 = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r] \in \mathbb{R}^{n \times r}$, $\Lambda_0 = \mathrm{diag}([\lambda_1, \cdots, \lambda_r]) \in \mathbb{R}^{r \times r}$, and $V_0 = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r] \in \mathbb{R}^{p \times r}$. If we define

$$\boldsymbol{\nu}_i^{\mathbf{l}} = \mathrm{vec}(U_0 R_i), \quad i = 1, \ldots, N \tag{13}$$

where $\{R_i \in \mathbb{R}^{r \times p}\}_{i=1}^N$ are N linearly independent matrices of size $r \times p$, then the rank of the matrix corresponding to $\mathbf{l} + \Delta_i \boldsymbol{\nu}_i^{\mathbf{l}}$ is still less than r. To see this, note that $\{U_0 R_i\}_{i=1}^N$ (hence $\{\mathrm{vec}(U_0 R_i)\}_{i=1}^N$) are linearly independent, since we have

$$\sum_{i=1}^N \alpha_i U_0 R_i = 0 \Rightarrow \sum_{i=1}^N \alpha_i R_i = 0. \tag{14}$$

Therefore, we can find, at most, $r \times p$ linearly independent directions in this manner. Similarly, if we take $\boldsymbol{\nu}_i^{\mathbf{l}} = \mathrm{vec}(L_i V_0^T)$, we find another $n \times r$ linearly independent directions. However, the union of the two sets of directions $\{\mathrm{vec}(U_0 R_i)\}_{i=1}^{pr} \cup \{\mathrm{vec}(L_j V_0)\}_{j=1}^{nr}$ are linearly dependent. As a matter of fact, we only have $(n+p)r - r^2$ linearly independent directions, which are explicitly given as the union of

$$\begin{aligned} &\left\{\mathrm{vec}\left(\boldsymbol{u}_i \boldsymbol{v}_j^T\right)\right\}_{\substack{1 \leq i \leq r \\ r+1 \leq j \leq p}} \\ &\left\{\mathrm{vec}\left(\boldsymbol{u}_i \boldsymbol{v}_j^T\right)\right\}_{\substack{1 \leq i \leq r \\ 1 \leq j \leq r}} \\ \text{and } &\left\{\mathrm{vec}\left(\boldsymbol{u}_i \boldsymbol{v}_j^T\right)\right\}_{\substack{r+1 \leq i \leq n \\ 1 \leq j \leq r}}. \end{aligned} \tag{15}$$

Here, $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are the columns of $U = [U_0 \ U_1] \in \mathbb{R}^{n \times n}$ and $V = [V_0 \ V_1] \in \mathbb{R}^{p \times p}$, respectively, and U_1 and V_1 are the unit orthogonal bases for the spaces orthogonal to $\mathrm{span}\{U_0\}$ and $\mathrm{span}\{V_0\}$, respectively. To connect with (13), we note that $\mathrm{vec}(\boldsymbol{u}_i \boldsymbol{v}_j^T) = \mathrm{vec}(U_0 R)$ with $R = \mathbf{e}_i^r (V \mathbf{e}_j^p)^T$ for $1 \leq i \leq r, r+1 \leq j \leq p$. We have used $\mathbf{e}_i^r$ to denote the ith canonical basis for $\mathbb{R}^r$. Other directions can be explained in a similar manner.

By using the Kronecker product, we express the directions in the three sets in (15) as the column vectors of $V_1 \otimes U_0$, $V_0 \otimes U_0$, and $V_0 \otimes U_1$, and define

$$Q_{\mathbf{l}} = [V_1 \otimes U_0 \ V_0 \otimes U_0 \ V_0 \otimes U_1] \in \mathbb{R}^{np \times [(n+p)r - r^2]}. \tag{16}$$

Note that the columns of $Q_{\mathbf{l}}$ are indeed linearly independent since algebraic manipulation shows $Q_{\mathbf{l}}^T Q_{\mathbf{l}} = \mathrm{I}_{(n+p)r - r^2}$, which implies $\mathrm{rank}(Q_{\mathbf{l}}) = (n+p)r - r^2$.

We now find linearly independent directions for S. Recall $\mathcal{S} = \mathrm{supp}(S) = \{(i,j) \in [n] \times [p] : S_{ij} \neq 0\}$. Clearly, the directions

$$\boldsymbol{\nu}_{i,j}^{\mathbf{s}} = \mathrm{vec}\left(\mathbf{e}_i^n \mathbf{e}_j^{pT}\right), \quad (i,j) \in \mathcal{S} \tag{17}$$

satisfy that $\mathbf{s} + \Delta_{i,j} \boldsymbol{\nu}_{i,j}^{\mathbf{s}}$ has at most s nonzero components for sufficiently small $\Delta_{i,j}$. Define $Q_{\mathbf{s}} \in \mathbb{R}^{np \times s}$ as the matrix whose columns are $\mathrm{vec}(\mathbf{e}_i^n \mathbf{e}_j^{pT})$, $(i,j) \in \mathcal{S}$, and define

$$P = \begin{bmatrix} Q_{\mathbf{l}} & \mathbf{0} \\ \mathbf{0} & Q_{\mathbf{s}} \end{bmatrix} \in \mathbb{R}^{2np \times [(n+p)r - r^2 + s]}. \tag{18}$$

Then, the columns of P, $\{\boldsymbol{\nu}_i\}$, form a set of unit orthogonal directions such that $\{\mathbf{x}_i = \mathbf{x} + \Delta_i \boldsymbol{\nu}_i\}_{i=1}^N \subseteq \mathcal{X}_{r,s}$ with $N = (n+p)r - r^2 + s$.

We have the following theorem, whose proof is given by applying the constrained CRB (11):

Theorem 1: For $(L,S) \in \mathcal{X}_{r,s}$, the MSE at (L,S) for any locally unbiased estimator $(\hat{L}, \hat{S})$ satisfies

$$\mathrm{MSE}_{(L,S)} \geq \mathrm{tr}\left(\begin{bmatrix} Q_{\mathbf{l}}^T \boldsymbol{A}^T \\ Q_{\mathbf{s}}^T \boldsymbol{A}^T \end{bmatrix} \Sigma^{-1} [\boldsymbol{A} Q_{\mathbf{l}} \quad \boldsymbol{A} Q_{\mathbf{s}}]\right)^{-1} \tag{19}$$

with $Q_{\mathbf{l}}$ and $Q_{\mathbf{s}}$ defined before, as long as $\mathrm{rank}([\boldsymbol{A} Q_{\mathbf{l}} \, \boldsymbol{A} Q_{\mathbf{s}}]) = (n+p)r - r^2 + s$.

The condition $\mathrm{rank}([\boldsymbol{A} Q_{\mathbf{l}} \, \boldsymbol{A} Q_{\mathbf{s}}]) = (n+p)r - r^2 + s$ is satisfied for selection operators if the operation $\boldsymbol{A}[Q_{\mathbf{l}} \ Q_{\mathbf{s}}]$ selects $(n+p)r - r^2 + s$ linearly independent rows of $[Q_{\mathbf{l}} \ Q_{\mathbf{s}}]$. This requires the selection operator to at least select all entries indexed by $\mathcal{S}$. If we are only interested in the low-rank part L, we can always focus on the operator's observed entries, because gross errors of the unobserved entries make no sense. With this in mind, and if, in addition, the singular vectors of X are not very spiky, then a random selection operator $\boldsymbol{A} \in \mathbb{R}^{m \times np}$ with m being sufficiently large selects $(n+p)r - r^2 + s$ linearly independent rows of $[Q_{\mathbf{l}} \ Q_{\mathbf{s}}]$ with high probability. In particular, for the robust principal component analysis, where $\mathcal{A}$ is the identity operator, the condition $\mathrm{rank}([Q_{\mathbf{l}} \ Q_{\mathbf{s}}]) = (n+p)r - r^2 + s$ is satisfied if $\mathrm{range}(Q_{\mathbf{l}}) \cap \mathrm{range}(Q_{\mathbf{s}}) = \{0\}$.

Now we consider the MCC model with white noise. For any index set Ω, we use $\mathcal{P}_\Omega \in \mathbb{R}^{np \times np}$ to denote the diagonal matrix whose diagonal entries are ones for indexes in Ω and zeros otherwise. We have the following corollary, which is proved by applying the block matrix inversion formula to (19).

Corollary 1: If $\mathcal{A}$ is the selection operator that observes the entries of L indexed by Ω, $\mathcal{S} = \mathrm{supp}(S)$, and $\Sigma = \sigma^2 \mathrm{I}_m$, then we always have

$$\begin{aligned} \mathrm{MSE}_{(L,S)} \geq & \mathrm{CRB} \\ & \stackrel{\mathrm{def}}{=} \left\{s - \left[(n+p)r - r^2\right]\right\} \sigma^2 + \mathrm{tr}\left(Q_{\mathbf{l}}^T \mathcal{P}_{\Omega/\mathcal{S}} Q_{\mathbf{l}}\right)^{-1} \sigma^2 \\ & \quad + \mathrm{tr}\left(\left(Q_{\mathbf{l}}^T \mathcal{P}_{\Omega/\mathcal{S}} Q_{\mathbf{l}}\right)^{-1} \left(Q_{\mathbf{l}}^T \mathcal{P}_{\Omega} Q_{\mathbf{l}}\right)\right) \sigma^2. \end{aligned} \tag{20}$$

For the RPCA with $\Omega = [n] \times [p]$, the bound reduces to

$$\left\{s - \left[(n+p)r - r^2\right]\right\} \sigma^2 + 2\mathrm{tr}\left(\left(Q_{\mathbf{l}}^T \mathcal{P}_{\mathcal{S}^c} Q_{\mathbf{l}}\right)^{-1}\right) \sigma^2. \tag{21}$$

IV. Probabilistic Analysis of the Constrained Cramér–Rao Bound

We analyze the probabilistic behavior of the bound (20) for the matrix completion and correction problem with white noise. Throughout this section, we denote $N = (n+p)r - r^2$.

We first fix the low-rank matrix L and the sparse matrix S. Since we would not be able to infer anything about the nonzero elements of S if they are not observed, we assume $\mathcal{S} \subset \Omega$. We further assume $\Omega/\mathcal{S}$ is a random subset of $[n] \times [p]/\mathcal{S}$.

We have the following theorem whose proof is given in the Appendix:

Theorem 2: Under the following assumption:

$$\|\mathrm{vec}(U)\|_\infty \leq \sqrt{\frac{\mu_B}{n}}, \quad \|\mathrm{vec}(V)\|_\infty \leq \sqrt{\frac{\mu_B}{p}} \tag{22}$$

if $Q_1^T \mathcal{P}_{\mathcal{S}^c} Q_1$ is nonsingular and

$$m - s \geq \frac{C(N \log^3 N)}{\varepsilon^2} \times \log\left(\frac{N}{\varepsilon^2}\right), \tag{23}$$

we have

$$\{s + 2/3g(L,S)\}\,\sigma^2 \leq \mathrm{CRB} \leq \{s + 2g(L,S)\}\,\sigma^2 \tag{24}$$

with probability greater than $1 - 10e^{-c/\varepsilon^2}$ for some constant $c > 0$. Here, $g(L,S)$ is defined as

$$\frac{np - s}{m - s}\mathrm{tr}\left(\left(Q_1^T \mathcal{P}_{\mathcal{S}^c} Q_1\right)^{-1}\left(\mathrm{I}_N + Q_1^T \mathcal{P}_{\mathcal{S}} Q_1\right)\right). \tag{25}$$

Note the CRB for the RPCA (21) satisfies (24) if we take $m = np$ and use $\mathcal{P}_{\mathcal{S}} = \mathrm{I}_{np} - \mathcal{P}_{\mathcal{S}^c}$.

Next we consider the second case where the entries of the low-rank matrix L are randomly and uniformly observed, among which a random and uniform subset of observations are grossly corrupted, i.e., the observation index set Ω is a random subset of $[n] \times [p]$ with size m, and the support $\mathcal{S}$ of S is a random and uniform subset of Ω. This makes $\Omega/\mathcal{S}$ a random and uniform subset of $[n] \times [p]$.

Theorem 3: In the above probability model and under the assumptions (22) and (23), we have

$$\begin{aligned}&\left\{s - N + \frac{1}{3}\frac{mN}{m-s} + \frac{2}{3}\frac{npN}{m-s}\right\}\sigma^2 \\ &\quad\leq \mathrm{CRB} \\ &\quad\leq \left\{s - N + 3\frac{mN}{m-s} + 2\frac{npN}{m-s}\right\}\sigma^2\end{aligned} \tag{26}$$

with probability greater than $1 - 10e^{-\frac{c}{\varepsilon^2}}$ for some constant $c > 0$.

The proof is very similar to that of Theorem 2 and hence is omitted.

By taking expectation of the constrained CRB (20) and using Jensen's inequality, we find that

$$\left\{s - N + \frac{mN}{(m-s)} + \frac{npN}{(m-s)}\right\}\sigma^2 \tag{27}$$

is a good approximation of the bound, especially when s is small (the matrix completion problem where $s = 0$), r is small, or m is large (the RPCA problem where $m = np$). This is because in these cases, the large number effect is stronger.

For the RPCA problem, the approximated bound becomes

$$\left\{s - N + \frac{2npN}{(np - s)}\right\}\sigma^2. \tag{28}$$

We compare our approximated bound (26) and (28) with an existing result for the stable principal component pursuit [4]. For simplicity, we consider model (4) with $n = p$. In [4], the noise is assumed to satisfy $\|W\|_{\mathrm{F}} \leq \delta$ for some $\delta > 0$. The following optimization problem is used to find L and S:

$$\min_{L,S} \|L\|_* + \lambda\|S\|_1 \text{ subject to } \|Y - L - S\|_{\mathrm{F}} \leq \delta. \tag{29}$$

The main result of [4, Equation (6)] states that with high probability, the solution $(\hat{L}, \hat{S})$ to (29) obeys

$$\frac{\left(\|\hat{L} - L\|_{\mathrm{F}}^2 + \|\hat{S} - S\|_{\mathrm{F}}^2\right)}{n^2} \leq C\delta^2. \tag{30}$$

Note that when the noise W follows i.i.d. Gaussian distribution with variance σ^2, the typical value of δ is of the order $n\sigma$. Therefore, in the Gaussian case, the bound (30) is comparable to

$$\frac{\left(\|\hat{L} - L\|_{\mathrm{F}}^2 + \|\hat{S} - S\|_{\mathrm{F}}^2\right)}{n^2} \leq Cn^2\sigma^2. \tag{31}$$

By substituting $r = \rho_r n$, $s = \rho_s n^2$, and $m = np$ into (26), we conclude that for *any* locally unbiased estimator of the RPCA, its error approximately satisfies

$$\frac{\left(\|\hat{L} - L\|_{\mathrm{F}}^2 + \|\hat{S} - S\|_{\mathrm{F}}^2\right)}{n^2} \geq C\sigma^2 \tag{32}$$

with high probability. Notice the gap in the order between the upper bound $n^2\sigma^2$ and the approximated lower bound σ^2. Of course, for the assumption (23) to hold, we need stricter assumptions on r and s. A careful reexamination of Theorem 3 shows that

$$r \leq \rho_r n(\log n)^{-5} \text{ and } s \leq \rho_s n^2 \tag{33}$$

are sufficient.

We next argue that under some proper assumptions, our conditions (22) and (33) guarantee exact recovery for the algorithm (1) with high probability. Note that since $U = [U_0\ U_1]$ and $V = [V_0\ V_1]$, condition (22) clearly implies the so called *weak incoherence property* [12]

$$\|\mathrm{vec}(U_0)\|_\infty \leq \sqrt{\frac{\mu}{n}} \text{ and } \|\mathrm{vec}(V_0)\|_\infty \leq \sqrt{\frac{\mu}{n}}. \tag{34}$$

As shown in [13], [14], "with the exception of a very few peculiar matrices" [12], the weak incoherence property (34) implies the *strong incoherence property* with $\mu_B = O(\log n)$:

$$\max_i \left\|U_0^T \mathbf{e}_i^n\right\|^2 \leq \frac{\mu_B r}{n}, \ \max_i \left\|V_0^T \mathbf{e}_i^p\right\|^2 \leq \frac{\mu_B r}{n}, \text{ and}$$
$$\left\|U_0 V_0^T\right\|_\infty \leq \sqrt{\frac{\mu_B r}{n^2}}. \tag{35}$$

Theorem 1.1 of [3] states that under the strong coherence property, the principal component pursuit (1) with $\lambda = 1/\sqrt{n}$ is exact with high probability (over the choice of the support of S), provided that

$$r \leq \rho_r n \mu_B^{-1} (\log n)^{-2} \text{ and } s \leq \rho_s n^2. \tag{36}$$

As a consequence of (36), under (22) we have if

$$r \leq \rho_r n (\log n)^{-3} \text{ and } s \leq \rho_s n^2, \tag{37}$$

then with high probability the recovery using principal component pursuit is exact in the noise-free case. Clearly, condition (33) is stronger than condition (37). Therefore, "with the exception of a very few peculiar matrices" [12], our conditions (22) and (33) can guarantee exact recovery in the noise-free case with high probability. The phrase "with the exception of a very few peculiar matrices" refers to excluding a set of matrices with a low probability in the probabilistic model defined in [12]–[14].

V. Maximum Likelihood Estimation for Matrix Completion and Correction

In this section, assuming knowledge of the rank r and the sparsity level s, we present a simple alternating minimization algorithm to heuristically compute the MLE of the low-rank matrix L and the sparse matrix S based on the matrix completion and correction model (3). Our MLE algorithm minimizes

$$\ell(L,S) = \sum_{(k,l)\in\Omega} (Y_{kl} - L_{kl} - S_{kl})^2 \tag{38}$$

with respect to $(L,S) \in \mathcal{X}_{r,s}$. We adopt an alternating minimization procedure. With a fixed S, we minimize $\ell(L,S)$ with respect to L, subject to the constraint that L is of a rank at most r; with a fixed L, we minimize $\ell(L,S)$ with respect to S, assuming S is s-sparse. For the latter minimization in estimating S, the MLE solution simply keeps the s components with largest absolute values (within those indexed by Ω) and sets the rest to zero, the so-called hard thresholding procedure.

For the former minimization in estimating L with a fixed S, we adopted another alternating minimization procedure developed for the matrix completion problem in [11]. For completeness, we briefly describe the algorithm in the following. Since $L \in \mathbb{R}^{n\times p}$ is of rank r, we write L as

$$L = FG \text{ where } F \in \mathbb{R}^{n\times r},\ G \in \mathbb{R}^{r\times p}. \tag{39}$$

Denote $\tilde{Y} = Y - S$. We have

$$\tilde{Y}_{kl} = L_{kl} + W_{kl} = F_k^T G_l + W_{kl}, \quad (k,l) \in \Omega \tag{40}$$

where F_k^T is the kth row of F and G_l is the lth column of G. The minimizers F and G are obtained by using another alternating minimization procedure. First, for a fixed G, setting the derivative of $\tilde{\ell}(F,G) = \sum_{(k,l)\in\Omega} (\tilde{Y}_{kl} - (FG)_{kl})^2$ with respect to F_k to zero gives

$$F_k = \left(\sum_{l:(k,l)\in\Omega} G_l G_l^T \right)^{-1} \left(\sum_{l:(k,l)\in\Omega} \tilde{Y}_{kl} G_l \right) \tag{41}$$

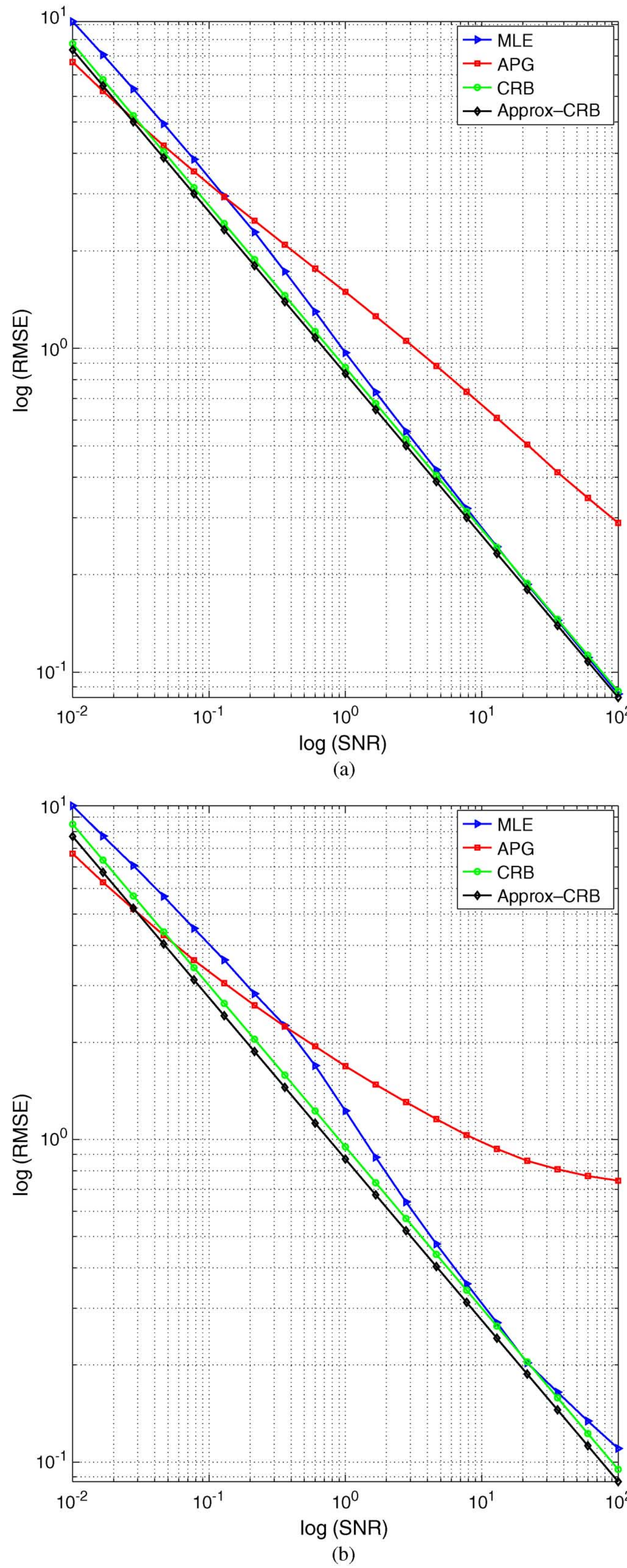

Fig. 1. Performance of the MLE and the APG algorithms for the RPCA compared with the constrained Cramér–Rao bound and its approximation. (a) RMSE versus $1/\sigma^2$ for $n = p = 200$, $r = 80$, and $s = 0.025np = 1000$; (b) RMSE versus $1/\sigma^2$ for $n = p = 200$, $r = 80$, and $s = 0.05np = 2000$.

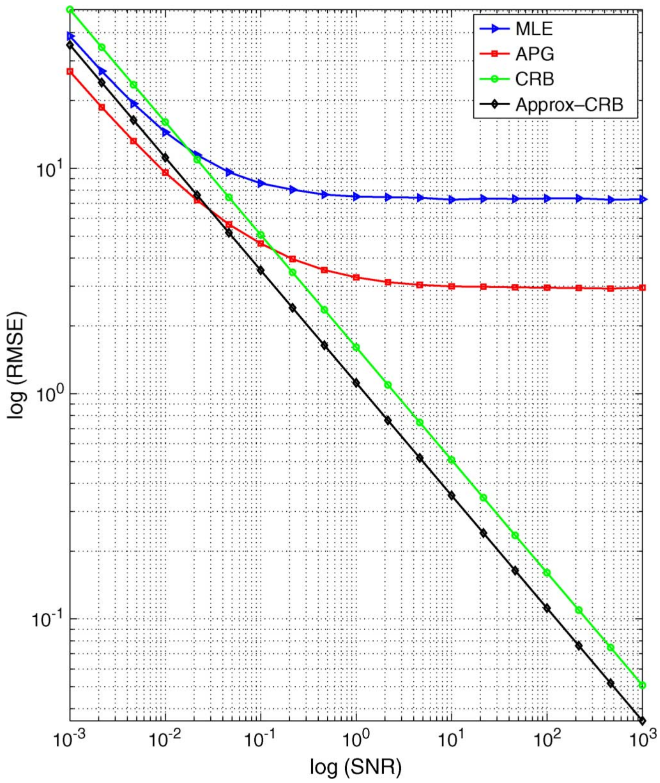

Fig. 2. Performance of the MLE and the APG algorithms compared with the constrained Cramér–Rao bound for $n = p = 100$, $r = 30$, and $s = 0.3np = 3000$.

for $k = 1, \cdots, n$. Similarly, when F is fixed, we get G_l

$$G_l = \left(\sum_{k:(k,l)\in\Omega} F_k F_k^T \right)^{-1} \left(\sum_{k:(k,l)\in\Omega} \tilde{Y}_{kl} F_k \right) \tag{42}$$

for $l = 1, \cdots, p$.

For the RPCA, as pointed out by a reviewer, the solution is given by thresholding the singular values.

Note that instead of taking inverses in (41) and (42), we solve systems of linear equations with positive definite system matrices of size $r \times r$. The overall computational complexity is $O(T_1 np \log(np) + T_1 T_2 m r^2)$, where T_1 and T_2 are, respectively, the maximal iteration numbers of the outer and inner loops.

VI. Numerical Examples

In this section, we show numerical examples to demonstrate the performance of both the APG and the MLE for both the MCC and the RPCA, and we compare it with the derived constrained CRB. We choose the APG to compare because it is the most stable one based on our numerical experiments.

We describe the experimental setup with specific values of the parameters given in the figures. The low-rank matrix L with rank r is generated as the product of two Gaussian matrices (i.e., entries follow $\mathcal{N}(0,1)$) of sizes $n \times r$ and $r \times p$. The support of the sparse matrix S is a uniform random subset of $[n] \times [p]$ of size $s = \rho_s np$, and the nonzero entries follow the uniform distribution on $[-12, 12]$. For the MCC model, we randomly observe $m = \rho_m np$ entries and the support of the sparse matrix S is a subset of these observed entries. The algorithms' MSEs are averaged over $T = 8$ runs. The parameter λ and μ for the APG are set to be $\lambda = \sqrt{\log n / n}/2$ and $\mu = \sqrt{2n}\sigma$ for $n = p$, respectively. We vary the noise level σ^2 and plot the root mean-square error $\mathrm{RMSE} \overset{\text{def}}{=} \sqrt{\text{MSE}/(np)}$ as a function of the SNR $\overset{\text{def}}{=} 1/\sigma^2$.

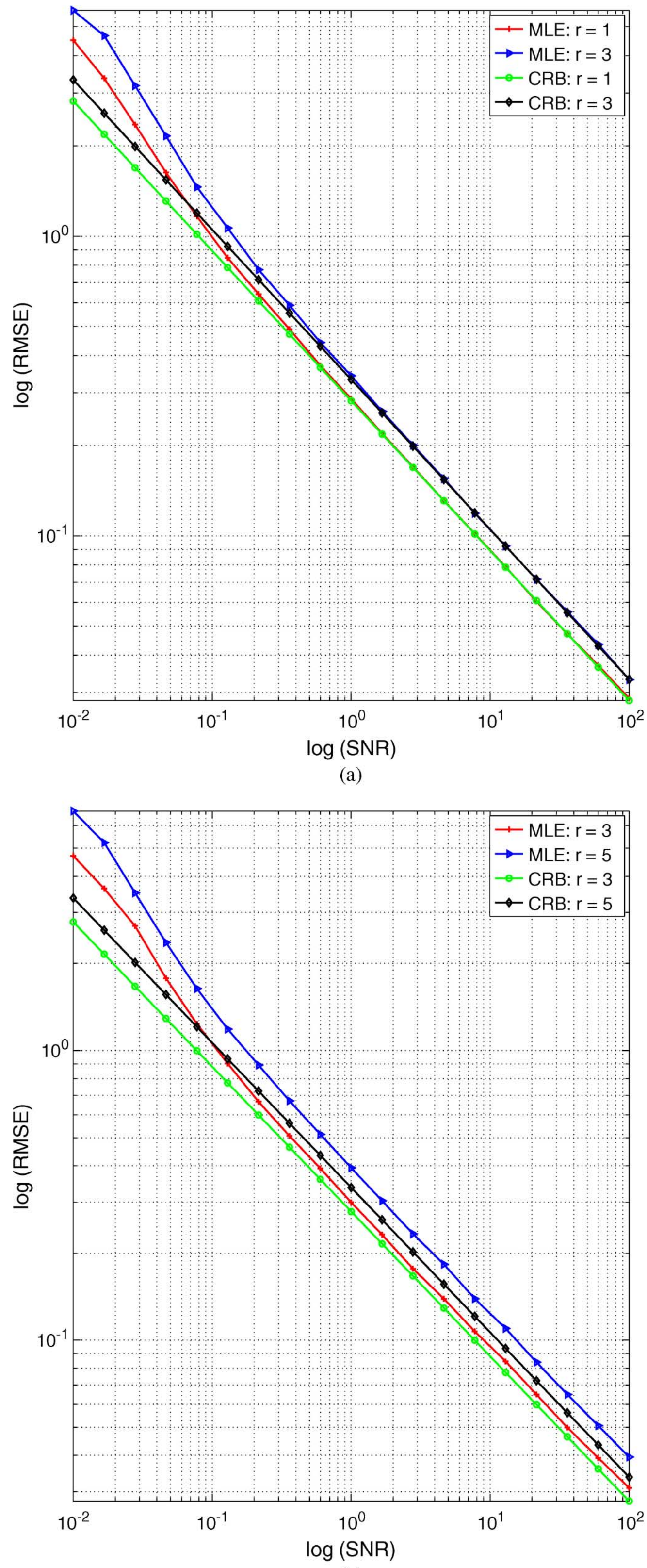

Fig. 3. Performance of the MLE for the MCC compared with the constrained Cramér–Rao bound. (a) RMSE versus $1/\sigma^2$ for $n = p = 200$, $r = 1, 3$, $s = 20$, and $\rho_m = 0.4$; (b) RMSE versus $1/\sigma^2$ for $n = p = 200$, $r = 3, 5$, $s = 1000$, and $\rho_m = 0.4$.

The CRB (21) and approximated CRB (28) are normalized in a similar manner.

In Fig. 1(a) and (b), we see that the APG performs better than the MLE for high levels of noise. This is probably achieved by introducing bias toward the zero matrix for L and S. However, for relatively high SNR, the performance of the MLE is better than that of the APG. This confirms that biased estimators are suboptimal when the signal is strong. The constrained CRB (21) predicts the behavior of the MLE well. The approximated CRB is a very good approximation of the CRB. When the sparsity level s and the rank r are small, the performance of the MLE converges to that given by the CRB while the SNR is getting large. The performance of both algorithms tend to deviate from the CRB for large s or r in the high SNR region. This phenomenon is better illustrated in Fig. 2, where we have intentionally used a large sparsity level with $\rho_s = 0.3$. We also see in Fig. 2 that the MLE performs uniformly worse than the APG. Note that unlike in other experiments where L and S are fixed and the MSE is averaged over only the noise realizations, the MSE in Fig. 2 is averaged over $T = 30$ different realizations of both the signals L, S and the noise W. We comment that averaging over only the noise results in a plot similar to that of Fig. 2.

In Fig. 3(a), we see that the MLE algorithm for the MCC model converges to the performance bound predicted by the constrained CRB when the noise is getting small. The convergence is better for smaller s and r. Actually, Fig. 3(b) demonstrates that for large s and r, there is a noticeable gap within the given noise levels.

VII. Conclusion

We analyzed the behavior of the MSE of locally unbiased estimators of the low-rank and sparse matrix decomposition. Compared with the performance analysis for specific algorithms, the lower bound applies to any locally unbiased estimator. We focused on the matrix completion and correction problem and its special case: the robust principal component analysis problem. We derived a constrained CRB and showed that the predicted performance bound is approached by the MLE. By using a concentration of measure argument, we identified the typical values of the constrained CRB when the observations and/or gross corruptions are uniform on the entries.

Appendix
Proof of Theorem 2

Proof: Recall the CRB (20) is σ^2 multiplied by

$$s - N + \operatorname{tr}\left(Q_1^T \mathcal{P}_{\Omega/S} Q_1\right)^{-1} + \operatorname{tr}\left(\left(Q_1^T \mathcal{P}_{\Omega/S} Q_1\right)^{-1}\left(Q_1^T \mathcal{P}_{\Omega} Q_1\right)\right).$$

Applying the concentration of measure result of [11, Theorem 2] to the process defined by

$$\left(Q_1^T \mathcal{P}_{S^c} Q_1\right)^{-\frac{1}{2}} \left(Q_1^T \mathcal{P}_{\Omega\setminus S} Q_1\right) \left(Q_1^T \mathcal{P}_{S^c} Q_1\right)^{-\frac{1}{2}}, \quad (43)$$

we have that with a probability of at least $1 - 10e^{-c/\varepsilon^2}$

$$\frac{1}{2}\frac{m-s}{np-s} Q_1^T \mathcal{P}_{S^c} Q_1 \le Q_1^T \mathcal{P}_{\Omega/S} Q_1 \le \frac{3}{2}\frac{m-s}{np-s} Q_1^T \mathcal{P}_{S^c} Q_1. \quad (44)$$

Taking inverse yields that $\left(Q_1^T \mathcal{P}_{\Omega/S} Q_1\right)^{-1}$ lies between

$$\frac{2}{3}\frac{np-s}{m-s}\left(Q_1^T \mathcal{P}_{S^c} Q_1\right)^{-1} \text{ and } 2\frac{np-s}{m-s}\left(Q_1^T \mathcal{P}_{S^c} Q_1\right)^{-1}. \quad (45)$$

We then use $\mathcal{P}_\Omega = \mathcal{P}_{\Omega/S} + \mathcal{P}_S$ to obtain

$$\operatorname{tr}\left(\left(Q_1^T \mathcal{P}_{\Omega/S} Q_1\right)^{-1}\left(Q_1^T \mathcal{P}_{\Omega} Q_1\right)\right) = N + \operatorname{tr}\left(\left(Q_1^T \mathcal{P}_{\Omega/S} Q_1\right)^{-1}\left(Q_1^T \mathcal{P}_{S} Q_1\right)\right). \quad (46)$$

The conclusion of Theorem 2 follows from (45) and (46). ∎

Acknowledgment

The authors would like to thank the anonymous reviewers for their insightful comments and suggestions, some of which have been incorporated into the paper.

References

[1] I. T. Jolliffe*, Principal Component Analysis, Springer Series in Statistics*. New York: Springer, 2002.
[2] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, “Sparse and low-rank matrix decompositions,” in *Proc. 47th Annu. Allerton Conf. Commun., Control, Comput. (Allerton 2009)*, Allerton, IL, Sep. 2009, pp. 962–967.
[3] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?,” *J. ACM*, vol. 58, no. 3, pp. 1–37, Jun. 2011.
[4] Z. Zhou, X. Li, J. Wright, E. J. Candès, and Y. Ma, “Stable principal component pursuit,” in *Proc. Int. Symp. Inf. Theory (ISIT 2010)*, Austin, TX, Jun. 2010, pp. 1518–1522.
[5] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2008.
[6] X. Yuan and J. Yang, “Sparse and low-rank matrix decomposition via alternating direction methods,” *Pacific J. Optim.*, 2009, to be published.
[7] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, “Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix,” Univ. of Illinois, Urbana-Champaign, IL, UIUC Tech. Rep. UILU-ENG-09-2214, Aug. 2009.
[8] Z. Lin, M. Chen, L. Wu, and Y. Ma, “The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices,” Univ. of Illinois, Urbana-Champaign, IL, UIUC Tech. Rep. UILU-ENG-09-2215, Nov. 2009.
[9] X. Ding, L. He, and L. Carin, “Bayesian robust principal component analysis,” *IEEE Trans. Image Process.* [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5771110, preprint, DOI: 10.1109/TIP.2011.2156801, to be published
[10] J. D. Gorman and A. O. Hero, “Lower bounds for parametric estimation with constraints,” *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1285–1301, Nov. 1990.
[11] G. Tang and A. Nehorai, “Lower bounds on mean-squared error of low-rank matrix reconstruction,” *IEEE Trans. Signal Process.*, vol. 59, no. 10, Oct. 2011.
[12] E. J. Candes and Y. Plan, “Matrix completion with noise,” *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.
[13] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.
[14] E. J. Candès and T. Tao, “The power of convex relaxation: Near-optimal matrix completion,” *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, May 2010.