ESTIMATING HIGH-DIMENSIONAL COVARIANCE MATRICES WITH MISSES FOR KRONECKER PRODUCT EXPANSION MODELS

Mahdi Zamanighomi[†]*, Zhengdao Wang*[‡]*, and Georgios B. Giannakis*[♮]

[†]Department of Statistics, Stanford University, Stanford, CA, USA
[‡]Department of Elec. and Comp. Engr., Iowa State University, Ames, IA, USA
[♮]Dept. of Elec. and Comp. Engr., University of Minnesota, Minneapolis, MN, USA
Email: mzamani@stanford.edu, zhengdao@iastate.edu, georgios@umn.edu

ABSTRACT

We study the problem of high-dimensional covariance matrix estimation from partial observations. We consider covariance matrices modeled as Kronecker products of matrix factors, and rely on observations with missing values. In the absence of missing data, observation vectors are assumed to be i.i.d multivariate Gaussian. In particular, we propose a new procedure computationally affordable in high dimension to extend an existing permuted rank-penalized least-squares method to the case of missing data. Our approach is applicable to a large variety of missing data mechanisms, whether the process generating missing values is random or not, and does not require imputation techniques. We introduce a novel unbiased estimator and characterize its convergence rate to the true covariance matrix measured by the spectral norm of a permutation operator. We establish a tight outer bound on the square error of our estimate, and elucidate consequences of missing values on the estimation performance. Different schemes are compared by numerical simulations in order to test our proposed estimator.

Index Terms— covariance matrices, Kronecker product, missing data, high-dimensional

1. INTRODUCTION

The problem of covariance estimation with partial observations is fundamental, and occurs in variety of applications, such as gene expression profile analysis [12, 16], machine learning [7], climate studies [14], and graphical models [9]. In practice, measurements may not be fully available in their entirety, which results in an observation data vector with missing entries. We can view the observation vector as having less entries provided that missing entries take place at fixed positions of the vector. However, missing entries may occur at positions that randomly change with time, requiring more complex estimation methods.

Recently, [15] has proposed a convex optimization approach to estimate covariance matrices with Kronecker Product (KP) structure and has derived a tight high-dimensional square error (SE) convergence rate. This method, termed the permuted rank-penalized least squares (PRLS), shows promising results in the spatial-temporal linear least-squares prediction of multivariate wind speed datasets. The PRLS approach however is not applicable in a large variety of problems entailing different patterns of misses.

In this paper, we generalize PRLS to the case of missing data. In particular, we seek to estimate high dimensional covariance matrices with KP structure through partial observations. We develop a novel method for the treatment of missing data, which requires neither imputing missing observations [11] nor discarding any available observations to recover the sought covariance matrix. Notably, this novel approach utilizes the empirical covariance matrix (ECM), even though the latter is not available as a result of missing observations. Furthermore, we show that our estimator achieves the same SE convergence rate as [15], wherein all observations are fully captured. In addition, we establish that the estimator convergence rate holds with a different probability depending on the missing patterns. Interestingly, our analysis reveals circumstances under which high convergence probability is guaranteed.

Notation: Column vectors and matrices are indicated by bold lower-case and upper-case letters, respectively. Symbol $\mathbf{x}(i)$ indicates the ith entry of vector $\mathbf{x}$ and $\mathbf{X}(i,j)$ denotes the (i,j)th element of matrix $\mathbf{X}$. We use $\mathbf{X}^T$ to denote the transpose of matrix $\mathbf{X}$, $\mathrm{vec}(\mathbf{X})$ the vectorized form of matrix $\mathbf{X}$ (stacking the columns of $\mathbf{X}$ into one column), $\|\mathbf{X}\|_F$ the Frobenius norm of matrix $\mathbf{X}$, $\|\mathbf{X}\|_*$ the nuclear norm of matrix $\mathbf{X}$, $\|\mathbf{X}\|_\infty$ the largest singular value of matrix $\mathbf{X}$, and $\|\mathbf{X}\|_0$ the smallest singular value of matrix $\mathbf{X}$. The operator $\circ$ indicates the Hadamard product, and $\otimes$ stands for the Kronecker product.

For a $d_1d_2 \times d_1d_2$ matrix $\mathbf{X}$, $\{\mathbf{X}[i,j]\}_{i,j=1}^{d_1}$ represents its $d_2 \times d_2$ block submatrices, where submatrices are in the form of $\mathbf{X}[i,j] = \mathbf{X}(1+(i-1)d_2 : id_2, 1+(j-1)d_2 : jd_2)$. We define the permutation map $\mathcal{P} : \mathbb{R}^{d_1d_2 \times d_1d_2} \to \mathbb{R}^{d_1^2 \times d_2^2}$, in which the $(i-1)d_1 + j$ row of $\mathcal{P}(\mathbf{X})$ is equal to $\mathrm{vec}(\mathbf{X}[i,j])^T$. We use $\mathrm{vec}^{-1}(\cdot)$ and $\mathcal{P}^{-1}(\cdot)$ to denote the inverse operator for $\mathrm{vec}(\cdot)$ and $\mathcal{P}(\cdot)$, respectively. The matrix $\mathbf{X}^{(\alpha)}$ is formed with (i,j) entry $\mathbf{X}(i,j)^\alpha$ and similarly, $\mathbf{x}^{(\alpha)}$ has its i-th entry $\mathbf{x}(i)^\alpha$. We let $\mathcal{S}_d := \{\mathbf{X} \in \mathbb{R}^{d_1 \times d} : \mathbf{X} = \mathbf{X}^T\}$ denote the set of real symmetric matrices, $\mathcal{S}_d^+$ the set of

real symmetric positive semidefinite matrices, $\mathcal{S}_d^{++}$ the set of real symmetric positive definite matrices, and $\mathcal{N}_d \coloneqq \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}^T\mathbf{x} = 1\}$ the unit Euclidean sphere.

2. SYSTEM MODEL

Let $\{\mathbf{x}_t\}_{t=1}^n$, $\mathbf{x}_t \in \mathbb{R}^d$, be multivariate Gaussian vectors with zero mean and unknown covariance matrix $\mathbf{\Sigma}_0$, uncorrelated across time. We observe n i.i.d random vectors $\{\mathbf{z}_t\}_{t=1}^n$ as

$$\mathbf{z}_t = \mathbf{\Gamma}_t\mathbf{x}_t, \quad 1 \le t \le n \tag{1}$$

where $\mathbf{\Gamma}_t$ is defined as the $d \times d$ diagonal matrix with $\mathbf{\Gamma}_t(i,i) = 0$, $1 \le i \le d$, if $\mathbf{x}_t(i)$ is missing and 1 otherwise. We emphasize that our analysis is not limited to any particular missing mechanism such as missing completely at random (MCR), missing at random (MR), or not missing at random (NMR) [1]. Particularly, we consider model (1) for all possible arrangements of $\mathbf{\Gamma}_t$ since several random and non-random processes could simultaneously give rise to missing values and even further, we may be unable to model the missing data mechanism [4].

Our goal is to estimate $\mathbf{\Sigma}_0$ given partial observations $\{\mathbf{z}_t\}_{t=1}^n$. We assume that (i) the positions of missing data, $\mathbf{\Gamma}_t$ for all $1 \le t \le n$, are known; and (ii) the covariance matrix can be written as a sum of KPs of lower dimensional pairs of matrices

$$\mathbf{\Sigma}_0 = \sum_{i=1}^{r} \mathbf{A}_i \otimes \mathbf{B}_i \tag{2}$$

where $\{\mathbf{A}_i\}_{i=1}^r$ are $d_1 \times d_1$ linearly independent matrices, $\{\mathbf{B}_i\}_{i=1}^r$ are $d_2 \times d_2$ linearly independent matrices, and $d = d_1 d_2$. We additionally assume that the factor dimensions d_1 and d_2 are given. The integer r denotes the total number of KPs in the summation, and is less than $\min(d_1^2, d_2^2)$ [15]. The mentioned model (2) can be interpreted as a low rank principal component decomposition, where its components are KPs, but neither orthogonal nor normalized. Such KP models show up naturally when the correlation is structured in individual dimensions (e.g., space and time, or transmit and receive sides of multiple antenna channels).

Given observations with no missing data, a sufficient statistic to estimate the true covariance matrix $\mathbf{\Sigma}_0$ is the ECM

$$\hat{\mathbf{\Sigma}}_0 = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t\mathbf{x}_t^T \tag{3}$$

Albeit unbiased, $\hat{\mathbf{\Sigma}}_0$ cannot be obtained since we only have access to $\mathbf{z}_t$. We thus consider the following alternative

$$\mathbf{\Sigma}_n^{\Gamma} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{z}_t\mathbf{z}_t^T. \tag{4}$$

This estimator concentrates around its mean, $\mathbf{\Sigma}_0^{\Gamma} \coloneqq E[\mathbf{\Sigma}_n^{\Gamma}]$, which could be far away from $\mathbf{\Sigma}_0$, and leads to unacceptably large biases in parameter estimates [3, 5]. To remove the introduced bias, let us first rewrite $\mathbf{\Sigma}_0^{\Gamma}$ as [c.f. (1)]

$$\begin{aligned}\mathbf{\Sigma}_0^{\Gamma} = E[\frac{1}{n}\sum_{t=1}^{n} \mathbf{z}_t\mathbf{z}_t^T] = E[\frac{1}{n}\sum_{t=1}^{n} \mathbf{\Gamma}_t\mathbf{x}_t\mathbf{x}_t^T\mathbf{\Gamma}_t] \\ = \frac{1}{n}\sum_{t=1}^{n} \mathbf{\Gamma}_t\mathbf{\Sigma}_0\mathbf{\Gamma}_t = \mathbf{W} \circ \mathbf{\Sigma}_0\end{aligned} \tag{5}$$

where $\mathbf{W}$ is the weight matrix with entries $\mathbf{W}(i,j) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{\Gamma}_t(i,i)\mathbf{\Gamma}_t(j,j)$. Although entries of $\mathbf{W}$ belong to the interval $[0,1]$, we assume $\mathbf{W}(i,j) \in (0,1]$ which holds provided that all variables are successfully measured in at least one time point. Therefore (5) can be represented as

$$\mathbf{\Sigma}_0 = \mathbf{W}^{(-1)} \circ \mathbf{\Sigma}_0^{\Gamma} \tag{6}$$

leading to the following unbiased estimator of $\mathbf{\Sigma}_0$ when the dataset contains missing observations [cf. (4)]

$$\hat{\mathbf{\Sigma}}_n \coloneqq \mathbf{W}^{(-1)} \circ \mathbf{\Sigma}_n^{\Gamma}. \tag{7}$$

This unbiased estimator not only takes advantage of all available information to estimate $\mathbf{\Sigma}_0$, but also can be employed whether missing patterns are random or not, so long as they are not systematic. The model (7) suffers from high variance when the number of samples, n, is smaller than the number of dimensions, d. To tackle this challenge, a low rank approximation to $\hat{\mathbf{\Sigma}}_n$ is usually considered. The popular low rank approximation relies on principal component analysis (PCA), which involves eigen-decomposition of $\hat{\mathbf{\Sigma}}_n$ to retain the top r principal components. The PCA-based estimator then takes the form of

$$\hat{\mathbf{\Sigma}}_n^{PCA} = \sum_{i=1}^{r} \sigma_i^2 \mathbf{v}_i\mathbf{v}_i^T \tag{8}$$

where σ_i is the ith largest singular value associated with the right singular vector $\mathbf{v}_i$. In high-dimensional settings, however, PCA can be severely affected by excessive bias [6]. This is mainly connected to known inconsistency of the sample eigenvalues and eigenvectors as d increases. Another issue is the possible high complexity associated with eigen-decomposition. The fact that the estimator does not account for the low-dimensional structure of (2) indicates that better estimator may be possible.

In [10], an alternative method to derive the low rank covariance estimation is proposed as the solution of the following penalized minimization problem (see also [11]):

$$\hat{\mathbf{\Sigma}}_n^{\lambda} \coloneqq \arg\min_{\mathbf{\Sigma} \in \mathbb{S}_{++}^p} \|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\|_F^2 + \lambda \operatorname{Tr}(\mathbf{\Sigma}) \tag{9}$$

in which λ is a tuning parameter and $\operatorname{Tr}(\mathbf{\Sigma})$ is equivalent to the l_1-norm on the eigenvalues of $\mathbf{\Sigma}$. The estimator (9) is developed when all $\mathbf{\Gamma}_t(i,i)$ are i.i.d Bernoulli (0-1) random variables with parameter δ and independent of $\{\mathbf{x}_t\}_{t=1}^n$. For this scenario, an unbiased estimator $\hat{\mathbf{\Sigma}}_n$ is simplified to

$$(\delta^{-1} - \delta^{-2})\operatorname{diag}(\mathbf{\Sigma}_n^{\Gamma}) + \delta^{-2}\mathbf{\Sigma}_n^{\Gamma}. \tag{10}$$

Corollary 1 in [10] proves that the solution to the convex problem (9) converges to $\mathbf{\Sigma}$ with a minimax optimal rate.

In this paper, we put forth a penalized empirical risk minimization problem analogous to (9), but applicable to any $\mathbf{\Gamma}_t$, and generalize PRLS [15] to accommodate misses, cf. models (1) and (2). Specifically, we propose the following convex optimization formulation to estimate the permuted version of $\mathbf{\Sigma}_0$:

$$\text{(P1)} \qquad \hat{\mathbf{P}}_n^{\gamma} = \arg \min_{\mathbf{P} \in \mathbb{R}^{d_1^2 \times d_2^2}} \|\hat{\mathbf{P}}_n - \mathbf{P}\|_F^2 + \gamma \|\mathbf{P}\|_*$$

where $\hat{\mathbf{P}}_n \coloneqq \mathcal{P}(\hat{\mathbf{\Sigma}}_n)$ (cf. Notation), $\mathbf{P} \coloneqq \mathcal{P}(\mathbf{\Sigma})$, and γ is a rank-controlling parameter. The term $\|\hat{\mathbf{P}}_n - \mathbf{P}\|_F^2$ is equivalent to $\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\|_F^2$ (Theorem 2.1 in [8]). To shed light on the need of $\|\mathbf{P}\|_*$, let us consider (2). It is easy to show that $\mathbf{P}_0 \coloneqq \mathcal{P}(\mathbf{\Sigma}_0) = \sum_{i=1}^{r} \text{vec}(\mathbf{A}_i)\text{vec}(\mathbf{B}_i)^T$. This suggests that $\mathbf{P}$ must be of rank r at most, and therefore (P1) is a convex relaxation of

$$\min_{\mathbf{P} \in \mathbb{R}^{d_1^2 \times d_2^2}} \|\hat{\mathbf{P}}_n - \mathbf{P}\|_F^2 \quad \text{subject to} \quad \text{rank}(\mathbf{P}) \leq r. \tag{11}$$

In particular, to obtain the convex relaxation of the NP-hard problem (11), we leverage recent developments in compressive sampling [13] and substitute the ℓ_0-norm with its ℓ_1-norm surrogate, which here corresponds to the nuclear norm $\|\mathbf{P}\|_*$.

It is well known that (P1) can be solved in closed form as [2]

$$\hat{\mathbf{P}}_n^{\gamma} = \sum_{i=1}^{\min(d_1^2, d_2^2)} \max\left(0, \sigma_i(\hat{\mathbf{P}}_n) - \frac{\gamma}{2}\right) \mathbf{u}_i \mathbf{v}_i^T \tag{12}$$

where $\sigma_i(\hat{\mathbf{P}}_n)$ is the ith largest singular value of $\hat{\mathbf{P}}_n$ corresponding to the left and right singular vectors $\mathbf{u}_i$ and $\mathbf{v}_i$, respectively. The answer $\hat{\mathbf{P}}^{\gamma}$ is essentially transformed back to the original matrix space $\mathbb{R}^{d \times d}$ as $\hat{\mathbf{\Sigma}}_n^{\gamma} \coloneqq \mathcal{P}^{-1}(\hat{\mathbf{P}}_n^{\gamma})$ (cf. Notation). In the next section, we explore the symmetry and positive definiteness of our estimate $\hat{\mathbf{\Sigma}}_n^{\gamma}$.

3. SPECTRAL NORM BOUND

In this section, we establish a bound on the spectral norm of $\mathbf{D}_n \coloneqq \hat{\mathbf{P}}_n - \mathbf{P}_0$. We will take advantage of this result to derive a tight outer bound on the squared estimation error. [1]

Theorem 1. *Consider* $\epsilon \in [0, \frac{1}{2})$. *Define* $N \coloneqq \max(d_1, d_2, n)$ *and* $C_0 \coloneqq \max(C_1 C_P, C_2\sqrt{C_P})$, *where*

$$C_P \coloneqq \max_{\mathbf{u} \in \mathcal{N}_{d_1^2}, \mathbf{v} \in \mathcal{N}_{d_2^2}} \mathbf{u}^{(2)} \mathcal{P}\left(\mathbf{W}^{(-2)}\right) \mathbf{v}^{(2)}.$$

If $q \geq \max\left(\sqrt{2C_1 C_P \ln(1 + \frac{2}{\epsilon})}, 2C_2\sqrt{C_P}\ln(1 + \frac{2}{\epsilon})\right)$, *then, it holds with probability at least* $1 - 2N^{-\frac{q}{2C_0}}$ *that*

$$\|\mathbf{D}_n\|_{\infty} \leq \frac{q\|\mathbf{\Sigma}_0\|_{\infty}}{1 - 2\epsilon} \max\left(\frac{d_1^2 + d_2^2 + \log N}{n}, \sqrt{\frac{d_1^2 + d_2^2 + \log N}{n}}\right). \tag{13}$$

In the proof, the following lemma turns out to be useful. The lemma allows us to generalize the operator norm bound on the permuted ECM (3), derived in [15], to our unbiased estimator (7).

Lemma 1. *Consider observation vectors* $\{\mathbf{z}_t\}_{t=1}^n$ *as introduced in the System Model. Let* $\mathbf{u} = (u_1, u_2, \ldots, u_{d_1^2})^T \in \mathcal{N}_{d_1^2}$, $\mathbf{v} = (v_1, v_2, \ldots, u_{d_2^2})^T \in \mathcal{N}_{d_2^2}$, *and recall* $\mathbf{D}_n$ *(see Section 3). We then have for all* $\delta \geq 0$,

$$\mathbb{P}\left(\left|\mathbf{u}^T \mathbf{D}_n \mathbf{v}\right| \geq \delta\right) \leq 2e^{-n\frac{\delta^2}{C_1 C_P \|\mathbf{\Sigma}_0\|_{\infty}^2 + C_2\sqrt{C_P}\delta\|\mathbf{\Sigma}_0\|_{\infty}}}. \tag{14}$$

Clearly, Theorem 1 demands no condition on $\mathbf{\Sigma}_0$. However, for the theorem to be of any practical interest, we require the outer bound in (13) to be small, leading to

$$n \geq \beta C_P (d_1^2 + d_2^2 + \log N) \tag{15}$$

where $\beta > 0$ is a sufficiently large constant number. Condition (15) reveals the impact of missing data, C_P, on the number of measurements sufficient to guarantee an accurate approximation to the spectral norm of $\mathbf{\Sigma}_0$. We note that the required number of samples does not dramatically grow as a response to missing data. This is because $C_P \leq \max_{i,j} \mathbf{W}^{(-2)}(i,j)$ is a small number in a variety of applications.

4. SE BOUND

Here, we establish a tight outer bound on the SE $\|\hat{\mathbf{\Sigma}}_n^{\gamma} - \mathbf{\Sigma}_0\|_F^2$. This result is built using a bound on the Frobenius norm of $\|\hat{\mathbf{P}}_n - \mathbf{P}_0\|_F^2$, and the fluctuation of $\mathbf{D}_n$ measured by the spectral norm in Theorem 1.

Theorem 2. *Choose*

$$\gamma = \frac{2q\|\mathbf{\Sigma}_0\|_{\infty}}{1 - 2\epsilon} \max\left(\frac{d_1^2 + d_2^2 + \log N}{n}, \sqrt{\frac{d_1^2 + d_2^2 + \log N}{n}}\right)$$

where the introduced parameters are as in Theorem 1. It then holds that with probability at least $1 - 2N^{-\frac{q}{2C_0}}$,

$$\|\hat{\mathbf{\Sigma}}_n^{\gamma} - \mathbf{\Sigma}_0\|_F^2 \leq \inf_{\mathbf{P}: \text{rank}(\mathbf{P}) \leq r} \|\mathbf{P} - \mathbf{P}_0\|_F^2 + \frac{(1 + \sqrt{2})^2}{4}\gamma^2 \text{rank}(\mathbf{P}) \tag{16}$$

Theorem 2 provides insight on the tuning of the regularization parameter γ. Clearly, the choice of γ depends on $\|\mathbf{\Sigma}_0\|_{\infty}$, which is generally unknown. Thus, we suggest using $\|\hat{\mathbf{\Sigma}}_n\|_{\infty}$ instead, so that γ can be specified based on the available information.

[1] Due to space limitation, all proofs are omitted and can be found in the journal version of this work on Arxiv.

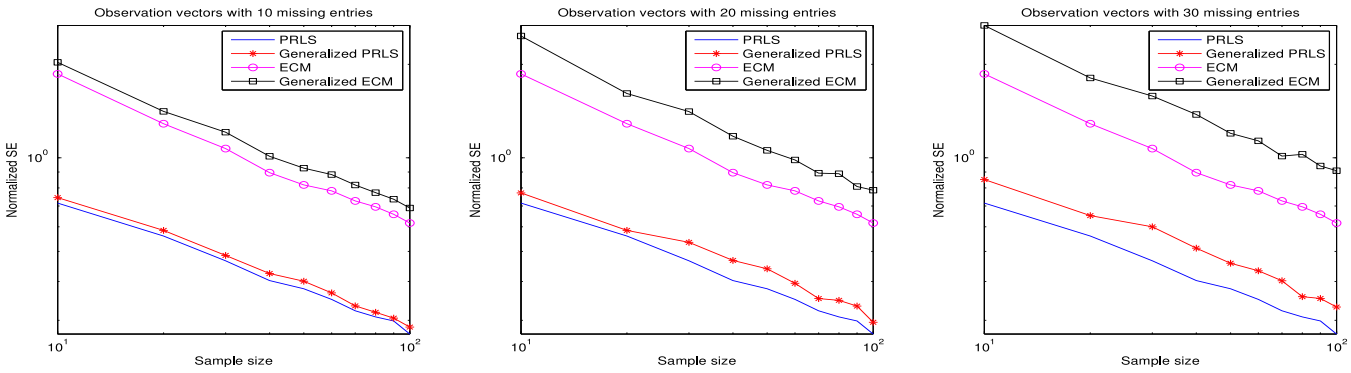

Fig. 1. SE performance normalized with respect to $\|\mathbf{\Sigma}_0\|_F^2$ versus the number of available samples n. The Generalized PRLS, $\hat{\mathbf{\Sigma}}_n$, and the Generalized ECM, $\hat{\mathbf{\Sigma}}_n$, are derived using 100-dimensional observation vectors with 10 missing entries (left), 20 missing entries (center), and 30 missing entries (right)

Given that $\mathbf{\Sigma}_0$ obeys the model (2), the estimation error $\min_{\mathbf{P}:\text{rank}(\mathbf{P})\leq r} \|\mathbf{P}-\mathbf{P}_0\|_F^2$ is zero. Therefore for large enough n, Theorem 2 offers that the SE $\|\hat{\mathbf{\Sigma}}_n^{\gamma}-\mathbf{\Sigma}_0\|_F^2$ is of order $r\frac{d_1^2+d_2^2+\log N}{n}$, with probability not less than $1-2N^{-\frac{q}{2C_0}}$. Indeed, this asymptotic SE convergence rate of the covariance estimate with partial observations coincides with one achieved in [15], where all observations are available. However, the probability $1-2N^{-\frac{q}{2C_0}}$ exhibits a change, that is the consequence of missing data. Specifically, q and C_0 have greater values, compared to the non-missing case. Furthermore, the larger q causes the regularization parameter γ to increase, placing a greater emphasis on the rank constraint in (P1).

The order of mean-square error (MSE) convergence rate for the standard sample covariance matrix is $\frac{d_1^2 d_2^2}{n}$, which is clearly less than the convergence rate of $\hat{\mathbf{\Sigma}}_n$. Therefore, we realize from Theorem 2 that the SE convergence rate of (P1) is significantly lower than the MSE convergence rate of the unbiased sample covariance matrix $\hat{\mathbf{\Sigma}}_n$, provided that rank $r \ll \min(d_1^2, d_2^2)$.

We finally deduce from Theorem 2 that the solution of (P1) takes a structure similar to (2) to satisfy the infimum (16), where each term in the expansion, $\mathbf{A}_i$ and $\mathbf{B}_i$, can be of arbitrary rank. This freedom, nevertheless, can not be offered by PCA since each term is limited to rank one.

5. NUMERICAL RESULTS

In order to provide a quantitative illustration of the results in this paper, we compare the SE performance obtained by the PRLS (solution of (4) in [15]), the Generalized PRLS (solution of (P1)), the ECM (equation (3)), and the Generalized ECM (equation (7)). We emphasize that the PRLS and ECM methods can not tolerate missing values while the Generalized PRLS and the Generalized ECM are applicable to missing data.

We construct the true covariance matrix $\mathbf{\Sigma}_0$ employing model (2) with $d_1 = d_2 = 10$ and $r = 3$. Factors A_i and B_i take the form of $\mathbf{S}\mathbf{S}^T$, $\mathbf{S}$ is a square random matrix whose columns follow a Gaussian distribution, which results in positive definite $\mathbf{\Sigma}_0$. We then generate 100-dimensional observation vectors based on the Gaussian distribution with zero mean and covariance matrix $\mathbf{\Sigma}_0$. To include missing values, we randomly force 10, 20, and 30 entries of each generated vector to be zero. For these three scenarios, the SE performance as a function of sample size is shown in Figure 1. As predicted by Theorem 2, the Generalized PRLS performs quite close to the PRLS when 10 and 20 percent of entries are missing. Furthermore for datasets containing a large number of missing values, such as right panel in Figure 1, we still achieve an acceptable performance in comparison with the PRLS. We finally observe that the Generalized PRLS notably outperforms the ECM and Generalized ECM.

6. CONCLUSIONS

We have generalized the PRLS method to datasets with missing values. The novel estimator is applicable to a large variety of missing data patterns, such as MCR, MR, and NMR, as long as all variables are observed in at least one time point. We performed an analysis of the concentration of measure phenomenon for observation vectors that are not multivariate Gaussian due to the presence of missing data. Using this result, we were able to establish a spectral norm bound along with a SE bound to illustrate the performance of our estimator. We have established that our generalized PRLS achieves the same convergence rate as the PRLS, but it holds with a different probability because of missing data. We observed from numerical results that the Generalized PRLS preforms quite close to the full-data PRLS even with a significant percentage of missing data.

Acknowledgment: The research in this paper was supported in part by NSF Grants 1523374, 1500713, and 1514056.

7. REFERENCES

[1] P. D. Allison, "Missing data: Quantitative applications in the social sciences," *British Journal of Mathematical and Statistical Psychology*, vol. 55, no. 1, pp. 193–196, 2002.

[2] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[3] M. Glasser, "Linear regression analysis with missing observations among the independent variables," *Journal of the American Statistical Association*, vol. 59, no. 307, pp. 834–844, 1964.

[4] J. W. Graham, S. M. Hofer, and D. P. MacKinnon, "Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures," *Multivariate Behavioral Research*, vol. 31, no. 2, pp. 197–218, 1996.

[5] M. P. Jones, "Indicator and stratification methods for missing explanatory variables in multiple linear regression," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 222–230, 1996.

[6] S. Lee, F. Zou, and F. A. Wright, "Convergence and prediction of principal component scores in high-dimensional settings," *Annals of statistics*, vol. 38, no. 6, pp. 3605, 2010.

[7] R. J. Little, "Robust estimation of the mean and covariance matrix from data with missing values," *Applied Statistics*, pp. 23–38, 1988.

[8] C. V. Loan and N. Pitsianis, *Approximation with Kronecker products*, Springer, 1993.

[9] P.-L. Loh, M. J. Wainwright, et al., "Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses," *The Annals of Statistics*, vol. 41, no. 6, pp. 3022–3049, 2013.

[10] K. Lounici, "High-dimensional covariance matrix estimation with missing observations," *arXiv preprint arXiv:1201.2577*, 2012.

[11] M. Mardani, G. Mateos, and G. B. Giannakis, "Subspace learning and imputation for streaming big data matrices and tensors," *IEEE Trans. Signal Processing*, vol. 63, no. 10, pp. 2663–2677, Oct. 2015.

[12] S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-i. Matsubara, and S. Ishii, "A bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no. 16, pp. 2088–2096, 2003.

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

[14] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values.," *Journal of Climate*, vol. 14, no. 5, 2001.

[15] T. Tsiligkaridis and A. O. Hero, "Covariance estimation in high dimensions via kronecker product expansions," *Signal Processing, IEEE Transactions on*, vol. 61, no. 21, pp. 5347–5360, Nov 2013.

[16] J. Xie and P. M. Bentler, "Covariance structure models for gene expression microarray data," *Structural Equation Modeling*, vol. 10, no. 4, pp. 566–582, 2003.