Low Separation Rank Covariance Estimation using Kronecker Product Expansions

Theodoros Tsiligkaridis and Alfred O. Hero III
Dept. of Electrical Engineering & Computer Science
University of Michigan, Ann Arbor, MI, USA
Email: {ttsili, hero}@umich.edu

***Abstract*—This paper presents a new method for estimating high dimensional covariance matrices. Our method, permuted rank-penalized least-squares (PRLS), is based on Kronecker product series expansions of the true covariance matrix. Assuming an i.i.d. Gaussian random sample, we establish high dimensional rates of convergence to the true covariance as both the number of samples and the number of variables go to infinity. For covariance matrices of low separation rank, our results establish that PRLS has significantly faster convergence than the standard sample covariance matrix (SCM) estimator. In addition, this framework allows one to tradeoff estimation error for approximation error, thus providing a scalable covariance estimation framework in terms of separation rank, an analog to low rank approximation of covariance matrices [1]. The MSE convergence rates generalize the high dimensional rates recently obtained for the ML Flip-flop algorithm [2], [3].**

I. Introduction

Covariance estimation is a fundamental problem in multivariate statistical analysis. It has received attention in diverse fields including economics and financial time series analysis (e.g., portfolio selection, risk management and asset pricing [4]), bioinformatics (e.g. gene microarray data [5], [6], functional MRI [7]) and machine learning (e.g., face recognition [8], recommendation systems [9]). In many modern applications, data sets are very large with both large number of samples n and large dimension d, often with $d \gg n$, leading to a number of covariance parameters that greatly exceeds the number of observations. The search for good low-dimensional representations of these data sets has recently yielded breakthroughs in multivariate statistics and signal processing. This modern theme of studying high-dimensional objects having small intrinsic dimension has sparked novel results and methodologies in signal processing. A good example being compressed sensing, where s-sparse vectors of dimension d can be recovered with $n = \Omega(s \log(d/s))$ appropriately designed measurements [10], [11], [12]. Similar results have appeared for the matrix completion problem, where a low-rank $d \times d$ matrix $\mathbf{C}$ can be recovered by nuclear norm minimization given only $n = \Omega(rd \log^2(d))$ observed entries, assuming $r = \text{rank}(\mathbf{C})$ and $\mathbf{C}$ satisfies an incoherence condition [13], [14], [15].

Kronecker product (KP) structure assumes that the covariance can be represented as the Kronecker product of two lower dimensional covariance matrices, i.e. $\mathbf{\Sigma}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0$, with $p \times p$ p.d. matrix $\mathbf{A}_0$ and $q \times q$ p.d. matrix $\mathbf{B}_0$ [16], [17]. When the data is a Gaussian random matrix having a Kronecker product covariance, the model is called the matrix normal distribution [18]. The model has applications in channel modeling for MIMO wireless communications [19], genomics [20], multi-task learning [21] and collaborative filtering [22]. The main difficulty in estimating KP-structured covariances via the maximum likelihood principle is the non-convex optimization problem that arises; thus, an alternating optimization approach is usually adopted. In the case of no missing data, an extension of the alternating optimization algorithm of Werner *et al* [17], that the authors call the flip flop (FF) algorithm, can be applied to estimate the parameters of this combined sparse and Kronecker product model, called KGlasso in [2]. Tsiligkaridis *et al* [2], [3] established the high dimensional convergence rate of FF and KGlasso, showing that only $n = \Omega((p^2 + q^2) \log(\max(p, q, n)))$ samples suffice for accurate covariance estimation (wrt. Frobenius norm) for the FF algorithm for the unstructured KP case, and only $n = \Omega((p+q) \log(\max(p, q, n)))$ is sufficient for the KGlasso algorithm for the sparse KP structured Gaussian graphical model.

In this paper, we propose a model that represents the covariance matrix as a sum of Kronecker products, where the number of terms in the summation, called the *separation rank*, may depend on the factor dimensions, and thus could potentially go to infinity. As in [17], [2] we assume n multivariate Gaussian observations, with $d = pq$ variables, whose $d \times d$ covariance $\mathbf{\Sigma}_0$ has the sum of Kronecker product representation:

$$\mathbf{\Sigma}_0 = \sum_{\gamma=1}^{r} \mathbf{A}_{0,\gamma} \otimes \mathbf{B}_{0,\gamma}, \tag{1}$$

where $\{\mathbf{A}_{0,\gamma}\}$ are $p \times p$ linearly independent matrices and $\{\mathbf{B}_{0,\gamma}\}$ are $q \times q$ linearly independent matrices [1]. We assume that the factor dimensions p, q are known. We note that the separation rank r satisfies $1 \leq r \leq r_0 = \min(p^2, q^2)$. The model is also relevant to other transposable models arising in recommendation systems like NetFlix and in gene expression analysis [9]. The model (1) with $r \geq 1$ has been proposed in spatiotemporal MEG/EEG covariance modeling [23], [24], [25] and SAR data analysis [26]. We finally note that Van Loan and Pitsianis [27] have shown that any $pq \times pq$ matrix

[1]Linear independence is understood with respect to the trace inner product defined in the space of symmetric matrices.

$\mathbf{\Sigma}_0$ can be written as an orthogonal expansion of Kronecker products of the form (1).

The principal contributions of this paper are twofold. First, we propose a novel convex optimization procedure, called the Permuted Rank-Penalized Least Squares (PRLS) method, for estimating covariance matrices with additive KP structure of the form (1). Second, we derive tight high-dimensional MSE convergence rates as n, p and q go to infinity. We establish high dimensional consistency of PRLS with a convergence rate guarantee of $O_P\left(\frac{r(p^2+q^2+\log\max(p,f,n))}{n}\right)$ as contrasted to the naive SCM rate $O_P\left(\frac{p^2q^2}{n}\right)$. To the best of our knowledge, this convex approach has not been proposed or studied in the high dimensional covariance estimation problem for estimating matrices of the form (1).

The high dimensional probabilistic analysis requires two large deviations results (see Lemma 1 and Thm. 2). We emphasize that our analysis is non-asymptotic, in the sense that probabilistic bounds are derived that holds with certain probability and this probability becomes higher as the number of sample and/or variables tend to infinity.

II. Permuted Rank-penalized Least-squares

Available are n i.i.d. multivariate Gaussian observations $\{\mathbf{z}_t\}_{t=1}^n$, where $\mathbf{z}_t \in \mathbb{R}^{pq}$, having zero-mean and covariance equal to (1). A sufficient statistic for covariance estimation is the well-known sample covariance matrix (SCM):

$$\hat{\mathbf{S}}_n = \frac{1}{n}\sum_{t=1}^{n} \mathbf{z}_t\mathbf{z}_t^T \tag{2}$$

A penalized least-squares approach was proposed in [1] for estimating a low rank covariance matrix by solving:

$$\hat{\mathbf{\Sigma}}_n^\lambda \in \arg\min_{\mathbf{S}\in S_{++}^d} \|\hat{\mathbf{S}}_n - \mathbf{S}\|_F^2 + \lambda\|\mathbf{S}\|_*$$

where $\lambda > 0$ is a regularization parameter and $\|\cdot\|_*$ denotes the spectral norm. For $\lambda = C\|\mathbf{\Sigma}_0\|_2\sqrt{\frac{r(\mathbf{\Sigma}_0)\log(2d)}{n}}$, where $C > 0$ is large enough, and $n \geq cr(\mathbf{\Sigma}_0)\log^2(\max(2d,n))$ for some constant $c > 0$ sufficiently large, Cor. 1 in [1] establishes a tight Frobenius norm error bound, which states that with probability $1 - \frac{1}{2p}$:

$$\|\hat{\mathbf{\Sigma}}_n^\lambda - \mathbf{\Sigma}_0\|_F^2 \leq \inf_{\mathbf{S}\in S_{++}^d} \|\mathbf{\Sigma}_0 - \mathbf{S}\|_F^2 + C_1\|\mathbf{\Sigma}_0\|_2^2 \mathrm{rank}(\mathbf{S})\frac{r(\mathbf{\Sigma}_0)\log(2d)}{n}$$

where $r(\mathbf{\Sigma}_0) = \frac{\mathrm{tr}(\mathbf{\Sigma}_0)}{\|\mathbf{\Sigma}_0\|_2} \leq \min\{\mathrm{rank}(\mathbf{\Sigma}_0), d\}$ is the effective rank [1].

Here we propose a similar nuclear norm penalization approach to estimate low separation-rank covariance matrices. Motivated by Van Loan and Pitsianis's work [27], we propose:

$$\hat{\mathbf{R}}_n^\lambda \in \arg\min_{\mathbf{R}\in\mathbb{R}^{p^2\times q^2}} \|\hat{\mathbf{R}}_n - \mathbf{R}\|_F^2 + \lambda\|\mathbf{R}\|_*, \tag{3}$$

where $\hat{\mathbf{R}}_n = \mathcal{R}(\hat{\mathbf{S}}_n)$ is the permuted SCM of size $p^2 \times q^2$. The permutation operator $\mathcal{R} : \mathbb{R}^{pq\times pq} \to \mathbb{R}^{p^2\times q^2}$ is defined by setting the $(i-1)p+j$ row of $\mathcal{R}(\mathbf{M})$ equal to $\mathrm{vec}(\mathbf{M}(i,j))^T$

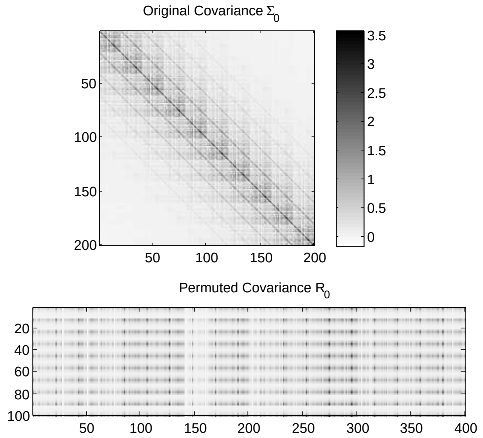

Fig. 1. Original (top) and permuted covariance (bottom) matrix. The original covariance is $\mathbf{\Sigma}_0 = \mathbf{A}_0 \times \mathbf{B}_0$, where $\mathbf{A}_0$ is a 10×10 Toeplitz matrix and $\mathbf{B}_0$ is a 20×20 unstructured p.d. matrix. Note that the permutation operator $\mathcal{R}$ maps a symmetric p.s.d. matrix $\mathbf{\Sigma}_0$ to a non-symmetric rank 1 matrix $\mathbf{R}_0 = \mathcal{R}(\mathbf{\Sigma}_0)$.

[17], [2]. An illustration of this permutation operator is shown in Fig. 1.

The minimum-norm problem considered in [27] is:

$$\min_{\mathbf{R}\in\mathbb{R}^{p^2\times q^2}:\mathrm{rank}(\mathbf{R})\leq r} \|\hat{\mathbf{R}}_n - \mathbf{R}\|_F^2 \tag{4}$$

We note that (3) is a convex relaxation of (4) and is more amenable to analysis. Furthermore, we show a tradeoff between approximation error (i.e., the error induced by model mismatch between the true covariance and the model (**??**)) and estimation error (i.e., the error due to finite sample size) by analyzing the solution of (3). We note that (3) is a strictly convex problem, so there exists a unique solution that can be found using various methods [28].

The closed form solution of (3) is given by singular value thresholding (SVT):

$$\hat{\mathbf{R}}_n^\lambda = \sum_{j=1}^{r_0}\left(\sigma_j(\hat{\mathbf{R}}_n) - \frac{\lambda}{2}\right)_+ \mathbf{u}_j\mathbf{v}_j^T$$

where $\mathbf{u}_j$ and $\mathbf{v}_j$ are the left and right singular vectors of $\hat{\mathbf{R}}_n$. Efficient methods of solving such problems have been recently studied in the literature [29], [30]. In practice, the separation rank r_0 may not be large [2]. Although empirically fast, the computational complexity of the algorithms presented in [29] and [30] is unknown, the computation of a rank r SVD is order $O(p^2q^2r)$. Faster probabilistic-based methods for truncated SVD take $O(p^2q^2\log(r))$ computational time [31]. Thus, the computational complexity of solving (3) scales well with respect to separation rank. We remark that the depermuted solution $\hat{\mathbf{\Sigma}}_n^\lambda = \mathcal{R}^{-1}(\hat{\mathbf{R}}_n^\lambda)$ is symmetric [32].

III. High Dimensional Consistency of RPLS

In this section, we show that RPLS achieves the MSE statistical convergence rate of $O_P\left(\frac{r(p^2+q^2+\log M)}{n}\right)$. This result is

[2]More details on choosing r are included later in the paper.

clearly superior to the statistical convergence rate of the naive SCM estimator:

$$\|\hat{\mathbf{S}}_n - \mathbf{\Sigma}_0\|_F^2 = O_P\left(\frac{p^2q^2}{n}\right). \tag{5}$$

The next result provides a deterministic relation between the spectral norm of $\hat{\mathbf{R}}_n - \mathbf{R}_0$ and the Frobenius norm of the the estimation error $\hat{\mathbf{R}}_n^\lambda - \mathbf{R}_0$.

Theorem 1. *Consider the convex optimization problem (3). When $\lambda \geq 2\|\hat{\mathbf{R}}_n - \mathbf{R}_0\|_2$, the following holds:*

$$\|\hat{\mathbf{R}}_n^\lambda - \mathbf{R}_0\|_F^2 \leq \inf_{\mathbf{R}} \left\{ \|\mathbf{R} - \mathbf{R}_0\|_F^2 + \frac{(1+\sqrt{2})^2}{4}\lambda^2 \mathit{rank}(\mathbf{R}) \right\}$$

Proof: The proof generalizes Thm. 1 in [1] to nonsquare matrices and is included in [32]. ■

A. High Dimensional Operator Norm Bound

In this subsection, we establish a tight bound on the spectral norm of the error matrix

$$\mathbf{\Delta}_n = \hat{\mathbf{R}}_n - \mathbf{R}_0 = \mathcal{R}(\hat{\mathbf{S}}_n - \mathbf{\Sigma}_0). \tag{6}$$

The strong law of large numbers implies that for fixed dimensions p, q, we have $\mathbf{\Delta}_n \to 0$ almost surely as $n \to \infty$. The next result characterizes the finite sample fluctuations of this convergence (in probability) measured by the spectral norm as a function of the sample size n and factor dimensions p, q. This result will be useful for establishing a tight bound on the Frobenius norm convergence rate of PRLS and can guide the selection of regularization parameter in (3).

Theorem 2. *(Operator Norm Bound on Permuted SCM) Assume $\|\mathbf{\Sigma}_0\|_2 < \infty$ for all p, q and define $M = \max(p, q, n)$. Fix $\epsilon' = \frac{1}{3}$. Assume $t \geq \max(\sqrt{4C_1 \ln(1+\frac{2}{\epsilon'})}, 4C_2 \ln(1+\frac{2}{\epsilon'}))$ and $C = \max(C_1, C_2) > 0$. Then, with probability at least $1 - 2M^{-\frac{t}{4C}}$,*

$$\|\mathbf{\Delta}_n\|_2 \leq \frac{C_0 t}{1-2\epsilon'} \max\left\{ \frac{p^2+q^2+\log M}{n}, \sqrt{\frac{p^2+q^2+\log M}{n}} \right\} \tag{7}$$

for some absolute constant $C_0 > 0$ [3].

Proof: See Appendix B. ■

Fig. 2 empirically validates the tightness of the bound (7) under the trivial separation rank 1 covariance $\mathbf{\Sigma}_0 = \mathbf{I}_p \otimes \mathbf{I}_q$.

B. High Dimensional MSE Convergence Rate for RPLS

Using bounds in Thm. 2 and Thm. 1, we next provide a tight bound on the MSE estimation error that decomposes into error due to model mismatch (first term on RHS of (8)) and error due to finite sample size.

[3]The constant in front of the rate can be tightened by optimizing it as a function of ϵ' over the interval $(0, 1/2)$, but is left as a finite constant here.

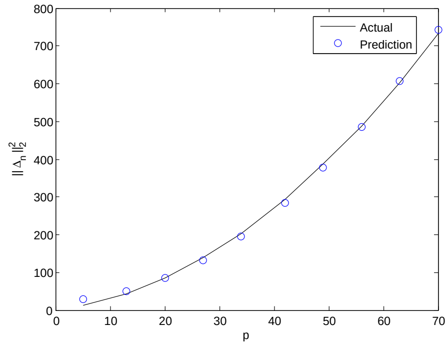

Fig. 2. Monte Carlo simulation for growth of spectral norm $\|\mathbf{\Delta}_n\|_2^2$ as a function of p for fixed $n = 10$ and $q = 5$. The predicted curve is a least-square fit of a quadratic model $y = ax^2 + b$ to the empirical curve. This demonstrates the tightness of the probabilistic bound (7).

Theorem 3. *Define $M = \max(p, q, n)$. Set $\lambda = \lambda_n = \frac{2C_0 t}{1-2\epsilon'} \max\left\{\frac{p^2+q^2+\log M}{n}, \sqrt{\frac{p^2+q^2+\log M}{n}}\right\}$ for $t > 0$ large enough (see (7)). Then, with probability at least $1 - 2M^{-\frac{t}{4C}}$:*

$$\|\hat{\mathbf{R}}_n^\lambda - \mathbf{R}_0\|_F^2 \leq \inf_{\mathbf{R}:\mathit{rank}(\mathbf{R}) \leq r} \|\mathbf{R} - \mathbf{R}_0\|_F^2 + C' r \max\left\{ \left(\frac{p^2+q^2+\log M}{n}\right)^2, \frac{p^2+q^2+\log M}{n} \right\} \tag{8}$$

for some absolute constant $C' > 0$.

Proof: See Appendix C. ■

When there is no model mismatch the approximation error $\inf_{\{\mathbf{R}:\text{rank}(\mathbf{R}) \leq r\}} \|\mathbf{R} - \mathbf{R}_0\|_F^2$ is zero and, as a result, in the large-p, q, n asymptotic regime where $p^2 + q^2 + \log M = o(n)$, it follows that $\|\hat{\mathbf{R}}_n^\lambda - \mathbf{R}_0\|_F = O_P(\sqrt{\frac{r(p^2+q^2+\log M)}{n}})$. This asymptotic MSE convergence rate of the estimated covariance to the true covariance reflects the number of degrees of freedom of the model, which is essentially of the order of $r(p^2+q^2)$ total covariance parameters. This result extends the recent results obtained in [2], [3] for the single Kronecker product model (i.e. $r = 1$).

Moreover, we note that $r \leq r_0 = \min(p^2, q^2)$. For the case of $p \sim q$, and $r \sim r_0$, we have a fully saturated Kronecker product model and the number of model parameters are of the order $p^4 \sim d^2$. In this case, the SCM convergence rate (5) coincides with the rate obtained in Thm. 3.

For covariance models of low separation rank-i.e., $r \ll r_0$, Thm. 3 establishes that the high dimensional MSE convergence rate of PRLS can be much lower than that of the naive SCM convergence rate. Thus PRLS is an attractive alternative to rank-based series expansions like PCA. We note that each term in the expansion $\mathbf{A}_{0,\gamma} \otimes \mathbf{B}_{0,\gamma}$ can be full-rank, while each term in the standard PCA expansion is rank 1.

Finally, we observe that Thm. 3 captures the trade-off between estimation error and approximation error. In other words, choosing a smaller r than the true separation rank would incur a larger approximation error

$\inf_{\{\mathbf{R}:\text{rank}(\mathbf{R})\leq r\}} \|\mathbf{R}-\mathbf{R}_0\|_F^2 > 0$, but smaller estimation error $O_P(\frac{r(p^2+q^2+\log M)}{n})$ and vice-versa.

IV. Simulation Results

We consider dense positive definite matrices $\mathbf{\Sigma}_0$ of dimension $d = 625$. Taking $p = q = 25$, we note that the number of free parameters that describe each Kronecker product is of the order $p^2 + q^2 \sim p^2$, which is essentially of the same order as the number of parameters to describe each eigenvector of $\mathbf{\Sigma}_0$, i.e., $pq \sim p^2$. The covariance matrix shown in Fig. 3 was constructed by first generating a Gaussian random matrix $\mathbf{C}$, then symmetrized to form $\mathbf{D} = \mathbf{C} + \mathbf{C}^T$, then a sparse matrix $\mathbf{M}$ was applied as $\mathbf{MDM}^T$ and finally its spectrum was perturbed from below to ensure positive definiteness. Fig. 4 compares the empirical performance of the KP estimator and the truncated eigendecomposition of the SCM for a designed separation rank 2 and eigendecomposition rank 2, respectively. We observe that the Kronecker product estimator performs much better than both the truncated eigendecomposition and naive SCM estimator. This is most likely due to the fact that the repetitive block structure of Kronecker products better summarizes the SCM. We observe from Fig. 3 that for

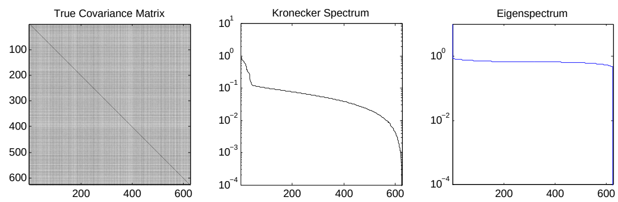

Fig. 3. True dense covariance matrix and Spectra. Left panel: True positive definite covariance matrix $\mathbf{\Sigma}_0$. Middle panel: Kronecker spectrum (eigenspectrum of $\mathbf{\Sigma}_0$ in permuted domain). Right panel: Eigenspectrum (Eigenvalues of $\mathbf{\Sigma}_0$).

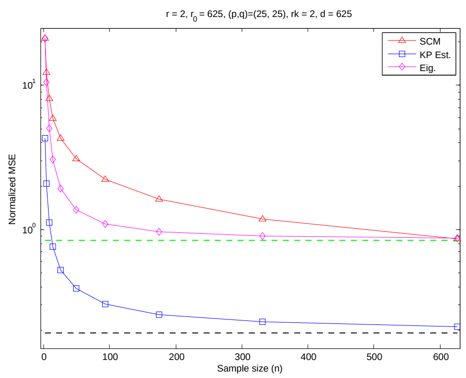

Fig. 4. Normalized MSE performance for covariance matrix as a function of sample size n. The KP estimator outperforms the truncated eigendecomposition and the standard SCM. Here, $p = q = 25$ and $N_{MC} = 80$. For $n = 49$, the KP estimator achieves a 5.433 dB MSE reduction over the truncated eigendecomposition and 8.99 dB MSE reduction over the standard SCM estimator. The error floor for the $r = 2$ eigenspectrum is 0.839 and for the $r = 2$ Kronecker spectrum is 0.19.

this arbitrarily structured covariance, the Kronecker spectrum decays more rapidly than the eigenspectrum, implying a more parsimonious (lower number of components) representation.

V. Conclusion

We have introduced a new framework for covariance estimation; separation rank decompositions using a series of Kronecker factors. We established high dimensional consistency for a penalized least squares estimator with guaranteed rates of convergence. The analysis shows that for low separation rank covariance models, our proposed method outperforms the standard SCM estimator. Future work will be to bound the approximation error term as a function of the factor dimensions p and q for different classes of covariance matrices.

Acknowledgment

The research reported in this paper was supported in part by ARO grant W911NF-11-1-0391.

Appendix A
Lemma 1

Lemma 1. *(Concentration of Measure for Coupled Gaussian Chaos) Let $\mathbf{X}$ and $\mathbf{Y}$ be arbitrary unit-Frobenius norm matrices and let $\mathbf{x} \in \mathbb{R}^{p^2}$ and $\mathbf{y} \in \mathbb{R}^{q^2}$ be reshaped versions of $\mathbf{X}$ and $\mathbf{Y}$. In the SCM (2) assume that $\{\mathbf{z}_t\}$ are i.i.d. multivariate normal $\mathbf{z}_t \sim N(0, \mathbf{\Sigma}_0)$. Recall $\mathbf{\Delta}_n$ in (6). For all $\tau \geq 0$:*

$$\mathbb{P}(|\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y}| \geq \tau) \leq 2\exp\left(\frac{-n\tau^2/2}{C_1\|\mathbf{\Sigma}_0\|_2^2 + C_2\|\mathbf{\Sigma}_0\|_2\tau}\right) \quad (9)$$

where $C_1 = \frac{4e}{\sqrt{6\pi}}$ and $C_2 = e\sqrt{2}$ are absolute constants.

Proof: This proof is based on large deviation theory for Gaussian matrices. Define $\mathbf{M} = \mathbf{X} \otimes \mathbf{Y}$. Using the definition of the reshaping operator $\mathcal{R}(\cdot)$ we can write [32] $\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y} = \frac{1}{n}\sum_{t=1}^n \psi_t$, where $\psi_t = \mathbf{z}_t^T\mathbf{M}\mathbf{z}_t - \mathbb{E}[\mathbf{z}_t^T\mathbf{M}\mathbf{z}_t]$. The statistic ψ_t has the form of Gaussian chaos of order 2. To simplify the concentration of measure derivation, we note that the stochastic equivalent of $\mathbf{z}_t^T\mathbf{M}\mathbf{z}_t$ is $\boldsymbol{\beta}_t^T\tilde{\mathbf{M}}\boldsymbol{\beta}_t$, where $\tilde{\mathbf{M}} = \mathbf{\Sigma}_0^{1/2}\mathbf{M}\mathbf{\Sigma}_0^{1/2}$ and $\boldsymbol{\beta}_t \sim N(\mathbf{0}, \mathbf{I}_{pq})$ is a random vector with i.i.d. standard normal components. By this decoupling argument, it follows [32] $\mathbb{E}|\psi_t|^2 = \|\tilde{\mathbf{M}}\|_F^2 + \|\text{diag}(\tilde{\mathbf{M}})\|_F^2 \leq 2\|\mathbf{\Sigma}_0\|_2^2$. It can also be shown (see Appendix A in [33]) that for all $m \geq 2$, $\mathbb{E}|\psi_t|^m \leq m!W^{m-2}v_t/2$, where where $W = e\sqrt{\mathbb{E}|\psi_t|^2} \leq e\sqrt{2}\|\mathbf{\Sigma}_0\|_2$ and $v_t = \frac{2e}{\sqrt{6\pi}}\mathbb{E}|\psi_t|^2 \leq \frac{4e}{\sqrt{6\pi}}\|\mathbf{\Sigma}_0\|_2^2$. An application of Bernstein's inequality (see Thm. 1.1 in [33]) then concludes the proof.

■

Appendix B
Proof of Theorem 2

Proof: Let $\mathcal{N}(\mathcal{S}^{d'-1}, \epsilon')$ denote an ϵ'-net on the sphere $\mathcal{S}^{d'-1}$ [34]. It can be shown [32] for any fixed $\epsilon' \in (0, 1/2)$:

$$\|\mathbf{\Delta}_n\|_2 \leq (1-2\epsilon')^{-1} \max_{\mathbf{x}\in\mathcal{N}(\mathcal{S}^{p^2-1},\epsilon'),\mathbf{y}\in\mathcal{N}(\mathcal{S}^{q^2-1},\epsilon')} |\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y}|$$

From Lemma 5.2 in [34], we have $\text{card}(\mathcal{N}(\mathcal{S}^{d'-1}, \epsilon')) \leq \left(1 + \frac{2}{\epsilon'}\right)^{d'}$. Using this cardinality bound, the union bound and Lemma 1:

$$\mathbb{P}(\|\mathbf{\Delta}_n\|_2 \geq \epsilon) \leq \mathbb{P}\Big(\bigcup_{\substack{\mathbf{x}\in\mathcal{N}(\mathcal{S}^{p^2-1},\epsilon') \\ \mathbf{y}\in\mathcal{N}(\mathcal{S}^{q^2-1},\epsilon')}} |\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y}| \geq \epsilon(1-2\epsilon')\Big)$$

$$\leq 2\left(1+\frac{2}{\epsilon'}\right)^{p^2+q^2} \exp\left(\frac{-n\epsilon^2(1-2\epsilon')^2/2}{C_1\|\mathbf{\Sigma}_0\|_2^2 + C_2\|\mathbf{\Sigma}_0\|_2\epsilon(1-2\epsilon')}\right)$$

We finish the proof by considering two separate sampling regimes: Gaussian tails and exponential tails. First, consider the Gaussian tail regime which occurs when $n > (\frac{tC_2}{C_1})^2(p^2 + q^2 + \log M)$ and choose $\epsilon = \frac{t\|\mathbf{\Sigma}_0\|_2}{1-2\epsilon'}\sqrt{\frac{p^2+q^2+\log M}{n}}$. For this regime, the bound can be relaxed to:

$$\mathbb{P}\left(\|\mathbf{\Delta}_n\|_2 \geq \frac{t\|\mathbf{\Sigma}_0\|_2}{1-2\epsilon'}\sqrt{\frac{p^2+q^2+\log M}{n}}\right) \leq 2M^{-\frac{t^2}{4C_1}}$$

where we used the assumption $t \geq \sqrt{4C_1 \ln(1+2/\epsilon')}$. This concludes the bound for the first regime. The exponential tail regime follows by similar arguments [32]. The proof is complete by combining both regimes and taking $C_0 > 0$ large enough [4] and noting that $t > 1$.

■

APPENDIX C
PROOF OF THEOREM 3

Proof: Define the event

$$E_r = \left\{\|\hat{\mathbf{R}}_n^\lambda - \mathbf{R}_0\|_F^2 > \inf_{\mathbf{R}:\text{rank}(\mathbf{R})\leq r} \|\mathbf{R}-\mathbf{R}_0\|_F^2 + \frac{(1+\sqrt{2})^2}{4}\lambda_n^2 r\right\}$$

where λ_n is chosen as stated. Thm. 1 implies that on the event $\lambda \geq 2\|\mathbf{\Delta}_n\|_2$, with probability 1, we have for any $1 \leq r \leq r_0$:

$$\|\hat{\mathbf{R}}_n^\lambda - \mathbf{R}_0\|_F^2 \leq \inf_{\mathbf{R}:\text{rank}(\mathbf{R})\leq r} \|\mathbf{R}-\mathbf{R}_0\|_F^2 + \frac{(1+\sqrt{2})^2}{4}\lambda^2 r$$

Using this and Thm. 2, we obtain [32]:

$$\begin{aligned}\mathbb{P}(E_r) &= \mathbb{P}(E_r \cap \{\lambda_n \geq 2\|\mathbf{\Delta}_n\|_2\}) + \mathbb{P}(E_r \cap \{\lambda_n < 2\|\mathbf{\Delta}_n\|_2\}) \\ &\leq \mathbb{P}(\lambda_n < 2\|\mathbf{\Delta}_n\|_2) \leq 2M^{-t/4C}\end{aligned}$$

This concludes the proof.

■

REFERENCES

[1] K. Lounici, "High-dimensional covariance matrix estimation with missing observations," *arXiv:1201.2577v5*, May 2012.
[2] T. Tsiligkaridis, A. Hero, and S. Zhou, "On Convergence of Kronecker Graphical Lasso Algorithms," *to appear in IEEE Transactions on Signal Processing*, 2013.
[3] ——, "Convergence Properties of Kronecker Graphical Lasso Algorithms," *arXiv:1204.0585*, July 2012.
[4] J. Bai and S. Shi, "Estimating high dimensional covariance matrices and its applications," *Annals of Economics and Finance*, vol. 12, no. 2, pp. 199–215, 2011.
[5] J. Xie and P. M. Bentler, "Covariance structure models for gene expression microarray data," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 10, no. 4, pp. 556–582, 2003.
[6] A. Hero and B. Rajaratnam, "Hub discovery in partial correlation graphs," *IEEE Transactions on Information Theory*, vol. 58, no. 9, pp. 6064–6078, September 2012.
[7] G. Derado, F. D. Bowman, and C. D. Kilts, "Modeling the spatial and temporal dependence in fmri data," *Biometrics*, vol. 66, no. 3, pp. 949–957, September 2010.

[4]Note that this constant depends on the constants $t, C_1, C_2 > 0$.

[8] Y. Zhang and J. Schneider, "Learning multiple tasks with a sparse matrix-normal penalty," *Advances in Neural Information Processing Systems*, vol. 23, pp. 2550–2558, 2010.
[9] G. I. Allen and R. Tibshirani, "Transposable regularized covariance models with an application to missing data imputation," *The Annals of Applied Statistics*, vol. 4, no. 2, pp. 764–790, 2010.
[10] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, pp. 118–124, July 2007.
[11] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509.
[12] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306.
[13] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.
[14] E. Candès and T. Tao, "The power of convex relaxation: near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.
[15] A. Rohde and A. B. Tsybakov, "Estimation of high-dimensional low-rank matrices," *Annals of Statistics*, vol. 39, no. 2, pp. 887–930, 2011.
[16] P. Dutilleul, "The mle algorithm for the matrix normal distribution," *J. Statist. Comput. Simul.*, vol. 64, pp. 105–123, 1999.
[17] K. Werner, M. Jansson, and P. Stoica, "On estimation of covariance matrices with Kronecker product structure," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, February 2008.
[18] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions*. Chapman Hill, 1999.
[19] K. Werner and M. Jansson, "Estimation of kronecker structured channel covariances using training data," in *Proceedings of EUSIPCO*, 2007.
[20] J. Yin and H. Li, "Model selection and estimation in the matrix normal graphical model," *Journal of Multivariate Analysis*, vol. 107, pp. 119–140, 2012.
[21] E. Bonilla, K. M. Chai, and C. Williams, "Multi-task gaussian process prediction," *Advances in Neural Information Processing Systems*, pp. 153–160, 2008.
[22] K. Yu, J. Lafferty, S. Zhu, and Y. Gong, "Large-scale collaborative prediction using a nonparametric random effects model," *ICML*, pp. 1185–1192, 2009.
[23] J. C. de Munck, H. M. Huizenga, L. J. Waldorp, and R. M. Heethaar, "Estimating stationary dipoles from meg/eeg data contaminated with spatially and temporally correlated background noise," *IEEE Transactions on Signal Processing*, vol. 50, no. 7, July 2002.
[24] J. C. de Munck, F. Bijma, P. Gaura, C. A. Sieluzycki, M. I. Branco, and R. M. Heethaar, "A maximum-likelihood estimator for trial-to-trial variations in noisy meg/eeg data sets," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 12, 2004.
[25] F. Bijma, J. de Munck, and R. Heethaar, "The spatiotemporal meg covariance matrix modeled as a sum of kronecker products," *NeuroImage*, vol. 27, pp. 402–415, 2005.
[26] A. Rucci, S. Tebaldini, and F. Rocca, "Skp-shrinkage estimator for sar multi-baselines applications," in *Proceedings of IEEE Radar Conference*, 2010.
[27] C. V. Loan and N. Pitsianis, "Approximation with kronecker products," in *Linear Algebra for Large Scale and Real Time Applications*. Kluwer Publications, 1993, pp. 293–314.
[28] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[29] J.-F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal of Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.
[30] J.-F. Cai and S. Osher, "Fast singular value thresholding without singular value decomposition," UCLA, Tech. Rep., 2010.
[31] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, May 2011.
[32] T. Tsiligkaridis and A. Hero, "Covariance Estimation in High Dimensions via Kronecker Product Expansions," *ArXiv*, 2013.
[33] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *IEEE Transactions on Information Theory*, May 2008.
[34] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv:1011.3027v7*, November 2011.