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, 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].

1. 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 $C$ can be recovered by nuclear norm minimization given only $n = \Omega(rd \log^2(d))$ observed entries, assuming $r = \text{rank}(C)$ and $C$ satisfies an incoherence condition [13], [14], [15].

Kronecker product (KP) structure assumes that the covariance matrix as a sum of Kronecker products, where the number of terms in the summation, called the separation rank $r$, satisfies

$$\Sigma_0 = \sum_{\gamma=1}^{r} A_{0,\gamma} \otimes B_{0,\gamma},$$

(1)

where $\{A_{0,\gamma}\}$ are $p \times p$ linearly independent matrices and $\{B_{0,\gamma}\}$ are $q \times q$ linearly independent matrices. 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 Pitasisanis [27] have shown that any $pq \times pq$ matrix

\[\text{Definition.} \quad \text{Linear independence is understood with respect to the trace inner product defined in the space of symmetric matrices.}\]
\( \Sigma_0 \) 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 \( OP \left( r(p^2 + q^2 + \log \max(p,f,n)) \right) \) as contrasted to the naive SCM rate \( OP \left( \frac{pq^2}{n} \right) \). 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 large deviations results (see Lemma 1 and Thm. 2). We available are

\[
\sum_{i,j} \lambda_{ij} \left( \sigma_j(\hat{R}_n) - \frac{\lambda}{2} \right)^2 + \left( \sigma_j(\hat{R}_n) - \frac{\lambda}{2} \right) \leq 0, j = 1, \ldots, q
\]

where \( \lambda_1, \lambda_2, \ldots, \lambda_q \) are the left and right singular vectors of \( \hat{R}_n \). Efficient methods of solving such problems have been recently studied in the literature [29], [30]. In practice, the computational complexity of the algorithms presented in [29] and [30] is unknown, the computation of a rank \( r \) SVD is order \( O(p^2 q^2 r) \). Faster probabilistic-based methods for truncated SVD take \( O(p^2 q^2 \log(r)) \) computational time [31]. Thus, the computational complexity of solving (3) scales well with respect to separation rank. We remark that the deprived solution \( \hat{R}_n \) is symmetric [32].

III. HIGH DIMENSIONAL CONSISTENCY OF RPLS

In this section, we show that RPLS achieves the MSE statistical convergence rate of \( OP \left( \frac{(p^2 + q^2 + \log M)^{1/2}}{n} \right) \). This result is
clearly superior to the statistical convergence rate of the naïve SCM estimator:
\[
\|\hat{S}_n - \Sigma_0\|_F^2 = O_P \left(\frac{p^2 q^2}{n}\right).
\]

The next result provides a deterministic relation between the spectral norm of \(\hat{R}_n - R_0\) and the Frobenius norm of the the estimation error \(\hat{R}_n^\lambda - R_0\).

**Theorem 1.** Consider the convex optimization problem (3). When \(\lambda \geq 2\|\hat{R}_n - R_0\|_F\), the following holds:
\[
\|\hat{R}_n^\lambda - R_0\|_F^2 \leq \inf_{R} \left\{ \|R - R_0\|_F^2 + \frac{(1 + \sqrt{2})^2}{4} \lambda^2 \text{rank}(R) \right\}
\]

**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
\[
\Delta_n = \hat{R}_n - R_0 = \hat{R}(\hat{S}_n - \Sigma_0).
\]

The strong law of large numbers implies that for fixed dimensions \(p, q\), we have \(\Delta_n \to 0\) 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 \(\|\Sigma_0\|_2 < \infty\) for all \(p, q\) and define \(M = \max(p, q, n)\).

Fix \(\epsilon' = \frac{1}{3}\). Assume \(t \geq \max(\sqrt{4C_1 \ln(1 + \frac{t}{2})}, 4C_2 \ln(1 + \frac{t}{2}))\) and \(C = \max(C_1, C_2) > 0\). Then, with probability at least \(1 - 2M^{-\pi t}\),
\[
\|\Delta_n\|_2 \leq C_0 t^{-1} \max \left\{ \frac{p^2 + q^2 + \log M}{n}, \sqrt{\frac{p^2 + q^2 + \log M}{n}} \right\}
\]
for some absolute constant \(C_0 > 0\).

**Proof:** See Appendix B.

Fig. 2 empirically validates the tightness of the bound (7) under the trivial separation rank 1 covariance \(\Sigma_0 = I_p \otimes I_q\).

**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.

\(\Delta_n \to 0\) as \(n \to \infty\) and the Frobenius norm of the error matrix
\[
\Delta_n = \hat{R}_n - R_0 = \hat{R}(\hat{S}_n - \Sigma_0).
\]

The strong law of large numbers implies that for fixed dimensions \(p, q, n\), we have \(\Delta_n \to 0\) 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 \(\|\Sigma_0\|_2 < \infty\) for all \(p, q\) and define \(M = \max(p, q, n)\).

Fix \(\epsilon' = \frac{1}{3}\). Assume \(t \geq \max(\sqrt{4C_1 \ln(1 + \frac{t}{2})}, 4C_2 \ln(1 + \frac{t}{2}))\) and \(C = \max(C_1, C_2) > 0\). Then, with probability at least \(1 - 2M^{-\pi t}\),
\[
\|\Delta_n\|_2 \leq C_0 t^{-1} \max \left\{ \frac{p^2 + q^2 + \log M}{n}, \sqrt{\frac{p^2 + q^2 + \log M}{n}} \right\}
\]
for some absolute constant \(C_0 > 0\).

**Proof:** See Appendix B.

Fig. 2 empirically validates the tightness of the bound (7) under the trivial separation rank 1 covariance \(\Sigma_0 = I_p \otimes I_q\).

**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.

\(\Delta_n \to 0\) as \(n \to \infty\) and the Frobenius norm of the error matrix
\[
\Delta_n = \hat{R}_n - R_0 = \hat{R}(\hat{S}_n - \Sigma_0).
\]

The strong law of large numbers implies that for fixed dimensions \(p, q, n\), we have \(\Delta_n \to 0\) 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 \(\|\Sigma_0\|_2 < \infty\) for all \(p, q\) and define \(M = \max(p, q, n)\).

Fix \(\epsilon' = \frac{1}{3}\). Assume \(t \geq \max(\sqrt{4C_1 \ln(1 + \frac{t}{2})}, 4C_2 \ln(1 + \frac{t}{2}))\) and \(C = \max(C_1, C_2) > 0\). Then, with probability at least \(1 - 2M^{-\pi t}\),
\[
\|\Delta_n\|_2 \leq C_0 t^{-1} \max \left\{ \frac{p^2 + q^2 + \log M}{n}, \sqrt{\frac{p^2 + q^2 + \log M}{n}} \right\}
\]
for some absolute constant \(C_0 > 0\).

**Proof:** See Appendix B.

Fig. 2 empirically validates the tightness of the bound (7) under the trivial separation rank 1 covariance \(\Sigma_0 = I_p \otimes I_q\).

**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.

\(\Delta_n \to 0\) as \(n \to \infty\) and the Frobenius norm of the error matrix
\[
\Delta_n = \hat{R}_n - R_0 = \hat{R}(\hat{S}_n - \Sigma_0).
\]

The strong law of large numbers implies that for fixed dimensions \(p, q, n\), we have \(\Delta_n \to 0\) 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 \(\|\Sigma_0\|_2 < \infty\) for all \(p, q\) and define \(M = \max(p, q, n)\).

Fix \(\epsilon' = \frac{1}{3}\). Assume \(t \geq \max(\sqrt{4C_1 \ln(1 + \frac{t}{2})}, 4C_2 \ln(1 + \frac{t}{2}))\) and \(C = \max(C_1, C_2) > 0\). Then, with probability at least \(1 - 2M^{-\pi t}\),
\[
\|\Delta_n\|_2 \leq C_0 t^{-1} \max \left\{ \frac{p^2 + q^2 + \log M}{n}, \sqrt{\frac{p^2 + q^2 + \log M}{n}} \right\}
\]
for some absolute constant \(C_0 > 0\).

**Proof:** See Appendix B.

Fig. 2 empirically validates the tightness of the bound (7) under the trivial separation rank 1 covariance \(\Sigma_0 = I_p \otimes I_q\).

**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.
Kronecker Spectrum

was constructed by first generating a Gaussian random matrix C, then symmetrized to form \( D = C + C^T \), then a sparse matrix M was applied as 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

\[ \text{inf}\{R\text{rank}(R) \leq r\} \left\| R - R_0 \right\|_F^2 > 0, \] 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 \( \Sigma_0 \) 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 \( \Sigma_0 \), i.e., \( pq \sim p^2 \). The covariance matrix shown in Fig. 3 was constructed by first generating a Gaussian random matrix C, then symmetrized to form \( D = C + C^T \), then a sparse matrix M was applied as 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

\[ \text{inf}\{R\text{rank}(R) \leq r\} \left\| R - R_0 \right\|_F^2 > 0, \] but smaller estimation error \( O_P(\frac{r^{p^2+q^2+ \log M}}{n}) \) and vice-versa.

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

LEMA 1

Lemma 1. (Concentration of Measure for Coupled Gaussian Chaos) Let \( X \) and \( Y \) be arbitrary unit-Frobenius norm matrices and let \( x \in \mathbb{R}^p \) and \( y \in \mathbb{R}^q \) be reshaped versions of \( X \) and \( Y \). In the SCM (2) assume that \( \{z_k\} \) are i.i.d multivariate normal \( z_k \sim N(0, \Sigma_0) \). Recall \( \Delta_n \) in (6). For all \( t \geq 0 \):

\[ \mathbb{P}(\|x^T \Delta_n y\| \geq t) \leq 2 \exp \left( \frac{-t^2/2}{C_1\|\Sigma_0\|_2 + C_2\|\Sigma_0\|_2^2} \right) \]

where \( C_1 = \frac{e}{\sqrt{8\pi}} \) and \( C_2 = e\sqrt{2} \) are absolute constants.

Proof: This proof is based on large deviation theory for Gaussian matrices. Define \( M = X \otimes Y \). Using the definition of the reshaping operator \( R(\cdot) \) we can write [32]

\[ x^T \Delta_n y = \frac{1}{n} \sum_{t=1}^n \psi_t, \]

where \( \psi_t = z_t^T M z_t - E[z_t^T M z_t] \). The statistic \( \psi_t \) has the form of Gaussian chaos of order 2. To simplify the concentration of measure derivation, we note that the stochastic equivalent of \( z_t^T M z_t \) is \( \beta_t^T M \beta_t \), where \( M = \Sigma_0^{1/2} M \Sigma_0^{1/2} \) and \( \beta_t \sim N(0, I_p) \) is a random vector with i.i.d. standard normal components. By this decoupling argument, it follows [32]

\[ \mathbb{E}[\|\psi_t\|^m] \leq \|M\|_F^m + \|\text{diag}(M)\|^m_2 \leq 2\|\Sigma_0\|_2^m.\]

It can also be shown (see Appendix A in [33]) that for all \( m \geq 2, \mathbb{E}[\|\psi_t\|^m] \leq m!W^{m-2}v_t^2/2, \) where \( W = e\sqrt{\mathbb{E}[\|\psi_t\|^2]} \leq e\sqrt{2}\|\Sigma_0\|_2^2 \) and \( v_t = \frac{e}{\sqrt{8\pi}} \mathbb{E}[\|\psi_t\|^2] \leq \frac{e^2}{8\pi} \|\Sigma_0\|_2^2 \). An application of Bernstein’s inequality (see Thm. 1.1 in [33]) then concludes the proof.

APPENDIX B

PROOF OF THEOREM 2

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

\[ \|\Delta_n\|_2 \leq (1 - 2\epsilon')^{-1} \max_{x \in \mathcal{N}(S^{d-1}, \epsilon')} \mathbb{E}_{y \in \mathcal{N}(S^{d-1}, \epsilon')} |x^T \Delta_n y|. \]

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

\[ \mathbb{P}(\|\Delta_n\|_2 \geq \epsilon) \leq \mathbb{P} \left( \bigcup_{x \in \mathcal{N}(S^{d-1}, \epsilon')} |x^T \Delta_n y| \geq \epsilon(1 - 2\epsilon') \right) \]


\[ \leq 2 \left(1 + \frac{2}{\epsilon^2}\right) \lambda^2 \exp \left(-t \frac{\| \Sigma_0 \|_2}{2} + C_2 \frac{\| \Sigma \|_F}{\sqrt{n}} (1 - 2\epsilon)\right) \]

We finish the proof by considering two separate sampling regimes: Gaussian tails and exponential tails. First, consider the Gaussian tail regime which occurs when \( n > \left(\frac{\epsilon^2}{C_2}\right)^2 \left(\frac{(\epsilon^2)^2}{2}\right) \). For this regime, the bound can be relaxed to:

\[ P \left( \| \Delta_n \|_2 \geq \frac{t}{1 - 2\epsilon} \sqrt{\frac{\| \Sigma \|_F}{\epsilon} + q^2 + \log M} \right) \leq 2M^{-\epsilon^2} \]

where we used the assumption \( t \geq \frac{\sqrt{4C_1 \log (1 + 2/\epsilon)}}{\epsilon} \). 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

\[ E_r = \left\{ \left\| \hat{R}_n - R_0 \right\|_F^2 \geq \inf_{R_{\text{rank}(R) \leq r}} \left\| R - R_0 \right\|_F^2 + \frac{(1 + \sqrt{2})^2}{4} \lambda_n^2 r \right\} \]

where \( \lambda_n \) is chosen as stated. Thm. 1 implies that on the event \( \lambda \geq 2 \| \Delta_n \|_2 \), with probability 1, we have for any \( 1 \leq r \leq r_0 \):

\[ \left\| \hat{R}_n - R_0 \right\|_F^2 \leq \inf_{R_{\text{rank}(R) \leq r}} \left\| R - R_0 \right\|_F^2 + \frac{(1 + \sqrt{2})^2}{4} \lambda_n^2 r \]

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

\[ P(E_r) = P(E_r \cap \{ \lambda_n \geq 2 \| \Delta_n \|_2 \}) + P(E_r \cap \{ \lambda_n < 2 \| \Delta_n \|_2 \}) \leq P(\lambda_n < 2 \| \Delta_n \|_2) \leq 2M^{-t/4C} \]

This concludes the proof.

**REFERENCES**


