Compressed Covariance Estimation with Automated Dimension Learning

Gautam Sabnis
University of Michigan, Ann Arbor, USA
Debdeep Pati and Anirban Bhattacharya
Texas A&M University, College Station, USA

Abstract

We propose a method for estimating a covariance matrix that can be represented as a sum of a low-rank matrix and a diagonal matrix. The proposed method compresses high-dimensional data, computes the sample covariance in the compressed space, and lifts it back to the ambient space via a decompression operation. A salient feature of our approach relative to existing literature on combining sparsity and low-rank structures in covariance matrix estimation is that we do not require the low-rank component to be sparse. A principled framework for estimating the compressed dimension using Stein’s Unbiased Risk Estimation theory is demonstrated. Experimental simulation results demonstrate the efficacy and scalability of our proposed approach.

AMS (2000) subject classification. Primary 62J10; Secondary 62H12.
Keywords and phrases. Compressed sensing, Dimension reduction, Low-rank, Factor model, Spiked covariance models, SURE

1 Introduction

Estimating a covariance matrix based on a sample of multivariate observations is a classical problem in statistics with applications across multitude of scientific disciplines. Time series analysis (Chen et al., 2013; Basu et al., 2015), portfolio optimization (Fan et al., 2008; Bai et al., 2011), gene networks (Butte et al., 2000; Schäfer et al., 2005), climate studies (Houtekamer and Mitchell, 2001; Hamill et al., 2001; Furrer et al., 2006), spatial data analysis (Kaufman et al., 2008), longitudinal data analysis (Smith and Kohn, 2002; Wu and Pourahmadi, 2003) among many other disciplines rely critically on the knowledge of the covariance structure. Recent efforts have focussed on high-dimensional data, where the dimension $p$ can be much larger than the sample size $n$ (Pourahmadi, 2011; Cai et al., 2016; Fan et al., 2016).
An unstructured $p \times p$ covariance matrix has $O(p^2)$ free entries. In moderate to high-dimensional situations, a general idea is to assume a restricted parameter space with much lower effective degrees of freedom. Examples include tapered covariance matrices (Furrer and Bengtsson, 2007), bandable covariance matrices (Wu and Pourahmadi, 2003; Bickel and Levina, 2008b), Toeplitz covariance matrices (Wu and Pourahmadi, 2009; McMurry and Politis, 2010; Xiao et al., 2012), sparse covariance matrices (Bickel and Levina, 2008a; Karoui, 2008a, b; Rothman et al., 2009; Cai and Liu, 2011), spiked sparse covariance matrices (Johnstone, 2001; Ma et al., 2013; Cai et al., 2015), covariances with a kronecker product structure (Werner et al., 2008), penalized likelihood estimation (Huang et al., 2006; d’Aspremont et al., 2008; Lam and Fan, 2009; Ravikumar et al., 2011) and regularization of principal components (Zou et al., 2006; Hoff, 2009; Johnstone and Lu, 2012; Cai et al., 2013). The challenge lies in the selection of appropriate structure or method for a specific data domain application.

An alternative approach to controlling the complexity in covariance matrix estimation is through low intrinsic dimensionality. Low intrinsic dimensionality posits that the dependencies between the variables are captured by a small number of latent components, also called factors, explicitly separating the common variation from variable-specific noise in the observed variables. Factor models in the high-dimensional regime have been used in a myriad of applications in economics and finance (Engle and Watson, 1981; Goldfarb and Iyengar, 2003). There are two disparate literatures which express the covariance matrix as the sum of a low-rank and sparse matrix: (a) factor models (Bhattacharya and Dunson, 2011; Fan et al., 2013; Pati et al., 2014), and (b) spiked covariance model (Johnstone, 2001). It is worthwhile to point out that the literature on both factor models and spiked covariance models assume the low-rank matrix is also sparse.

While the sparsity assumptions on the covariance matrix or the low rank component are well motivated for specific applications (Zou et al., 2006; Shen and Huang, 2008; Witten et al., 2009; Bhattacharya and Dunson, 2011), recent studies reveal striking correlations between several genes/loci or gene networks and the features of the diseases they cause (Jimenez-Sanchez et al., 2001; Ideker and Sharan, 2008). The correlation between the attributes of complex disease genes is more extensive and stronger than previously thought.

In this article, we attempt to mitigate the aforementioned gaps in the literature by (i) providing an efficient way of estimating covariance matrices which admit a decomposition of the form that is written as a sum of low-rank
plus sparse matrix albeit with dense to moderately sparse low-rank structures, (ii) providing a concrete principled way of choosing the dimension of the low-rank matrix using Stein’s Unbiased Risk Estimation (SURE) theory. We show via simulations that our approach is readily scalable to massive covariance matrices. The approach is based on a “compression-decompression” (CD) mechanism. The CD mechanism proceeds by projecting the high dimensional observations to a lower dimension to form compressed measurements and then decompress them back to the original dimension. The key idea is to use the sample covariance matrix of the compressed-decompressed data as an estimator of \( \Sigma \) instead of the sample covariance matrix \( \hat{\Sigma} \).

The covariance estimation problem considered in this paper is related to the covariance sketching problem where the goal is to recover or reconstruct second-order statistics of the data from “sketches” or compressed measurements of \( p \)-variate random vectors. Examples include sketching covariance matrices for popular covariance structures, namely, Toeplitz (Bioucas-Dias et al., 2014), sparse (Dasarathy et al., 2015), low-rank (Davenport and Romberg, 2016), or simultaneously structured (e.g. Toeplitz low-rank Chen et al., 2015; Qiao and Pal, 2015, sparse and low-rank Bahmani and Romberg 2015). The literature mostly assumes either sparsity or some low rank structure on the target covariance matrix with the exception of Bahmani and Romberg (2015) which deals with simultaneous low rank and sparse target covariance matrices.

Our work deviates from the majority of work on covariance sketching in that we tackle approximate low-rank covariance matrices which can be both dense or sparse, which to the best of our knowledge, is not accommodated by the literature on covariance sketching. Another distinction is that in some cases of covariance sketching, the guarantees for the estimation algorithm depend on “sketching” or projection matrices having certain properties. Our proposed estimator’s performance is independent of the choice of the projection matrix.

Our simulation studies demonstrate the efficacy of our approach and show the CD estimator to be highly competitive to the recent state-of-the-art covariance estimators such as POET (Fan et al., 2013) and adaptive thresholding (Cai and Liu, 2011).

## 2 Compression - Decompression Covariance Estimator

We now elucidate the compression-decompression (CD) estimator. Let \( X = [x_1, \ldots, x_n] \) denote a \( p \times n \) data matrix with the columns \( x_i \) independent and identically distributed from a \( p \)-variate distribution whose covariance
we wish to estimate. We assume the data to be column-centered, so that \( \mathbb{E}(x_i) = 0 \) for all \( i \), and thereby define \( \hat{\Sigma}_x = n^{-1} \sum_{i=1}^{n} x_i x_i^T = XX^T/n \) as the sample covariance matrix. The mechanism proceeds by projecting the data to a lower-dimensional space to form compressed measurements, computing the sample covariance in the compressed space, and lifting back to the ambient space via a decompression operation.

Specifically, given \( k < p \) and a \( k \times p \) unitary matrix \( \phi \) (with \( \phi\phi^* = I_k \); \( \phi^* \) denoting the complex conjugate of \( \phi \)), project the data \( x_i \) from \( \mathbb{R}^p \mapsto \mathbb{R}^k \) to create compressed data \( w_i = \phi x_i \); let \( W = [w_1, \ldots, w_n] \) denote the corresponding \( k \times n \) matrix. The \( k \times k \) sample covariance matrix \( \Sigma_w = n^{-1} \sum_{i=1}^{n} w_i w_i^* = WW^*/n \) is expected to be more stable compared to the \( p \times p \) matrix \( \hat{\Sigma}_x \). To obtain a \( p \times p \) covariance estimate for our original problem, we decompress \( \Sigma_w \) using the transformation \( \phi^* \) to define

\[
\hat{\Sigma}(\phi; k) = \phi^* \Sigma_w \phi = \frac{1}{n} \phi^*(WW^*)\phi = \phi^*(\phi \hat{\Sigma}_x \phi^*)\phi. \tag{2.1}
\]

Observe that the regularization in this framework arises from the compression operation, which is entirely different from assuming \( \ell_q \) type sparsity on the covariance matrix or its various decompositions. One of the motivations behind this approach comes from Theorem 8.1 of Pati et al. (2014), where it is shown that if the true \( \Sigma_0 = \Lambda_0 \Lambda_0^T + \sigma^2 I_p \) with \( \Lambda_0 \in \mathbb{R}^{p \times k} \), and one chooses \( \phi = \Lambda_0 \), then \( \hat{\Sigma} \) concentrates around \( \Sigma_0 \) in operator norm with high probability when \( k \ll p \) and \( \sigma^2 \) is bounded. It is important to mention here that a similar concentration for the sample covariance matrix in a spiked covariance model requires \( \Sigma \) to be have a low effective rank (Bunea and Xiao, 2015), necessitating \( \sigma^2 = O(1/p) \), which is fairly restrictive.

Evidently, the estimator in Eq. 2.1 depends on two unknown parameters, the compression matrix \( \phi \) and the projected dimension \( k \). There has been recent work in the regression context (Guhaniyogi and Dunson, 2013) where very high-dimensional covariates are projected to a lower-dimensional space using one particular instance of a random sensing matrix. However, this approach of fixing \( \phi \) has very poor performance in our setting. An alternative way may be to estimate \( \phi \) from the data which is computationally intensive as it requires estimating \( p \times k \) many parameters. Here, instead of trying to estimate the high-dimensional parameter \( \phi \), we average over the ensemble of unitary matrices which can be performed in closed-form using random matrix results (Marzetta et al., 2011). Specifically, let \( h(\cdot) \) denote the Haar measure on the space of \( k \times p \) unitary matrices satisfying \( h(\phi \Psi) = h(\phi) \) for all
non-stochastic $p \times p$ unitary matrices $\Psi$. Letting $E_\phi$ denote the expectation with respect to $h$, define the CD estimator

$$\hat{\Sigma}_{CD}(k) = E_\phi[\hat{\Sigma}_{CD}(\phi; k)].$$

Based on recent random matrix techniques as in Marzetta et al. (2011), the above expectation can be computed as

$$\hat{\Sigma}_{CD}(k) = \frac{k}{(p^2 - 1)p} \left[ (pk - 1)\hat{\Sigma}_x + (p - k)\text{Tr}(\hat{\Sigma}_x)I_p \right],$$

(2.2)

where $\text{Tr}(A)$ denotes the trace of a matrix $A$. Clearly, averaging over the Haar distribution introduces an appropriate shrinkage in the sample covariance matrix resulting in an estimator that is full rank even if $p \gg n$. As noted by Marzetta et al. (2011), Eq. 2.2 bears resemblance with shrinkage estimators of the type $a\hat{\Sigma}_x + (1 - a)I_p$, where $0 < a < 1$ and $I_p$ is a $p \times p$ identity matrix, originally proposed by Ledoit and Wolf (2004). However, a key difference is that the effect of shrinkage $a\hat{\Sigma}_x$ is compensated by $\text{Tr}(\hat{\Sigma}_x)I_p$ in Eq. 2.2 instead of just $I_p$. This helps preserving the largest eigenvalues of the resulting estimator.

A fundamental principle of statistical decision theory is that there exists an interior optimum in the trade-off between bias and estimation error. One way of attaining this optimal trade-off is simply to take a properly weighted average of the biased and unbiased estimators. Refer, for example, to the seminal work by Stein et al. (1956), who showed that shrinking sample means towards a constant can, under certain circumstances, improve accuracy. The crux of our method is to shrink the unbiased but very variable sample covariance matrix towards the biased but less variable identity covariance matrix and to thereby obtain a more efficient estimator. In addition, the resulting estimator, in Eq. 2.2 is invertible and well-conditioned, which is of crucial importance in settings where one needs to estimate the inverse, for example, in portfolio selection (Ledoit et al., 2003), Gaussian graphical models (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007) among others.

3 Choice of the Compressed Dimension

In this section, we propose a data-driven method for choosing the size of the compressed dimension. The main idea is to consider the Frobenius risk of the estimator for each choice of compressed dimension, $k$, and then choose that value which minimizes the Frobenius risk. Since the risk depends on the truth, we derive an unbiased estimator of the Frobenius risk curve and find the minimizer of the estimated risk curve to choose a value of $k$. We
use Stein’s Unbiased Risk Estimation (SURE) theory to find an unbiased estimator of the Frobenius risk associated with the CD estimator.

SURE was originally proposed in Stein (1981) in deriving an unbiased estimator of the risk of James–Stein estimate. Efron (1986, 2004) applied SURE to prediction problems which was called covariance penalty method. Li and Zou (2016) proved the asymptotic properties of SURE information criterion for large bandable covariance matrices and proposed a family of generalized SURE (SUREc) indexed by c for covariance matrix estimation, where c is some constant. For bandable covariance matrices, Li and Zou (2014) claimed that SURE2 and SURE\logern can be regarded as AIC and BIC analogues, respectively, for covariance matrix estimation problems. Xiao and Bunea (2014) proposed an improved version of the banding estimator obtained in Bickel and Levina (2008b) and used SURE-type approach for selecting the bandwidth for the banding estimator.

Let $R(k) = \mathbb{E}\|\hat{\Sigma}_{CD}(k) - \Sigma_0\|_F^2$ be the Frobenius risk associated with the proposed estimator for a fixed $k$. We use the following risk identity proved in Yi and Zou (2013),

$$R(k) = \mathbb{E}\|\hat{\Sigma}_{CD}(k) - \hat{\Sigma}_x\|_F^2 - \sum_{i,j}^{p} \text{var}(\hat{\sigma}_{ij}) + 2\sum_{i,j}^{p} \text{cov}(\hat{\sigma}_{ij}^{(k)}, \hat{\sigma}_{ij})$$

(3.1)

where $\hat{\Sigma} = [\hat{\sigma}_{ij}]$ is the usual sample covariance matrix, $\tilde{\Sigma} = [\tilde{\sigma}_{ij}]$ is the maximum likelihood estimator of the covariance matrix, with $\hat{\Sigma} = \frac{n}{n-1} \tilde{\Sigma}$, and $\hat{\Sigma}_{CD}(k) = [\hat{\sigma}_{ij}^{(k)}]$ is the proposed covariance estimator (2.2). The third term on the right hand side of Eq. 3.1 is referred to as the optimism (Efron, 2004). The second term on the right hand is the same for all estimators of $\Sigma_0$.

Standard results from multivariate statistics (Anderson, 1984) imply

$$\text{var}(\hat{\sigma}_{ij}) = \frac{\sigma_{ii}^2 \sigma_{jj}}{n-1}, \quad \text{var}(\tilde{\sigma}_{ii}) = \frac{2\sigma_{ii}^2}{n-1}, \quad \text{and} \quad \text{cov}(\tilde{\sigma}_{ll}, \tilde{\sigma}_{ii}) = \frac{2(n+2)}{n(n-1)} \sigma_{ll}^2 + \frac{(n+1)}{n-1} - 1 \sigma_{ii} \sigma_{ll}.$$

Lemma 3.1 provides unbiased estimators to each of these quantities which enables us to provide an unbiased estimator of $R(k)$. Recall $\hat{\Sigma}$ and $\tilde{\Sigma}$ defined in the previous paragraph.

**Lemma 3.1.** Let $\hat{\cdot}$ denote an unbiased estimator of the quantity of interest. Then

$$\hat{\text{var}}(\hat{\sigma}_{ij}) = \frac{n^2(n^2 - n - 4)}{(n-1)^2(n^2 + n^2 - 2n - 4)} \tilde{\sigma}_{ij}^2 + \frac{n^3}{(n-1)(n^3 + n^2 - 2n - 4)} \tilde{\sigma}_{ii} \tilde{\sigma}_{jj},$$

(3.2)

$$\hat{\text{var}}(\hat{\sigma}_{ii}) = \frac{n^2(2n^2 - 2n - 4)}{(n-1)^2(n^2 + n^2 - 2n - 4)} \tilde{\sigma}_{ii}^2,$$

(3.3)

$$\hat{\text{cov}}(\hat{\sigma}_{ll}, \hat{\sigma}_{ii}) = \frac{2n^2(n+2)}{(n-1)(n^3 + n^2 - 2n - 4)} \tilde{\sigma}_{ij}^2 + \frac{2(n-2)n^2}{(n-1)(n^3 + n^2 - 2n - 4)} \tilde{\sigma}_{ii} \tilde{\sigma}_{ll}.$$
Table 1: Simulation Setting 1 with $s = 0.5$

|                  | ktr 10 |       |       | ktr 50 |       |       |
|------------------|--------|-------|-------|--------|-------|-------|
|                  |        | 250   | 500   | 1000   | 250   | 500   | 1000 |
| $k^{opt}$        | 240    | 470   | 940   | 220    | 430   | 840   |
| $k^{sure}$       | 240    | 470   | 940   | 200    | 410   | 810   |
| SC               | 0.33   | 0.32  | 0.32  | 0.97   | 0.95  | 0.91  |
|                  | (0.05) | (0.05)| (0.05)| (0.09) | (0.08)| (0.07)|
| CD               | 0.28   | 0.27  | 0.27  | 0.62   | 0.56  | 0.52  |
|                  | (0.03) | (0.03)| (0.02)| (0.03) | (0.02)| (0.01)|
| AT               | 0.32   | 0.32  | 0.31  | 0.91   | 0.77  | 0.69  |
|                  | (0.05) | (0.05)| (0.05)| (0.06) | (0.04)| (0.04)|
| POET             | 0.32   | 0.32  | 0.31  | 0.96   | 0.94  | 0.89  |
|                  | (0.05) | (0.05)| (0.04)| (0.09) | (0.08)| (0.07)|
| SC               | 0.60   | 0.60  | 0.60  | 2.59   | 2.62  | 2.58  |
|                  | (0.05) | (0.05)| (0.05)| (0.09) | (0.08)| (0.07)|
| CD               | 0.56   | 0.55  | 0.56  | 2.10   | 2.12  | 2.09  |
|                  | (0.04) | (0.04)| (0.03)| (0.03) | (0.03)| (0.03)|
| AT               | 0.63   | 0.62  | 0.62  | 3.43   | 3.52  | 3.49  |
|                  | (0.06) | (0.06)| (0.06)| (0.10) | (0.08)| (0.09)|
| POET             | 0.59   | 0.59  | 0.59  | 2.56   | 2.60  | 2.56  |
|                  | (0.05) | (0.05)| (0.05)| (0.08) | (0.08)| (0.07)|

Comparison with SC, AT and POET estimators

We now derive the SURE criterion for the Frobenius risk $R(k)$ defined in Eq. 3.1.

**Theorem 3.2.** An unbiased estimator of the Frobenius risk associated with $\hat{\Sigma}_{CD}(k)$ is given by

\[
SURE(k) = (\eta - 1)^2 \| \hat{\Sigma} \circ \hat{\Sigma} \|_F + p\gamma^2 + 2\gamma(\eta - 1) \text{Tr}(\hat{\Sigma}) \\
+ 2\left( (a_n\eta + d_n\gamma) \left( \| \hat{\Sigma} \circ \hat{\Sigma} \|_F - \text{Tr}(\hat{\Sigma} \circ \hat{\Sigma}) \right) \\
+ (b_n\eta + c_n\gamma) \left( \text{Tr}(\hat{\Sigma})^2 - \text{Tr}(\hat{\Sigma} \circ \hat{\Sigma}) \right) + c_n(\eta + \gamma) \text{Tr}(\hat{\Sigma} \circ \hat{\Sigma}) \right)
\]

where $\eta = \frac{k(p-k-1)}{p(p^2-1)}$, $\gamma = \frac{p-k}{p(p^2-1)}$, and $\Sigma_1 \circ \Sigma_2$ denotes the Schur product between two matrices $\Sigma_1$ and $\Sigma_2$ of the same dimension.

Having obtained the SURE criterion, we can now minimize it with respect to $k$ to select the compressed dimension. Specifically, define

\[
\hat{k}^{\text{sure}} = \arg\min_k SURE(k).
\]
Table 2: Simulation Setting 1 with $s = 0.1$

| p  | 10  | 50  | 100 | 250 | 500 | 1000 |
|----|-----|-----|-----|-----|-----|------|
| k$^{\text{opt}}$ | 240 | 470 | 950 | 210 | 410 | 820  |
| $\hat{k}^{\text{sure}}$ | 240 | 470 | 940 | 210 | 410 | 810  |
| SC | 0.58 | 0.56 | 0.57 | 1.79 | 1.67 | 1.65 |
|    | (0.10) | (0.09) | (0.10) | (0.15) | (0.15) | (0.12) |
| CD | 0.49 | 0.47 | 0.47 | 1.13 | 1.00 | 0.95 |
|    | (0.05) | (0.04) | (0.04) | (0.05) | (0.03) | (0.03) |
| AT | 0.57 | 0.55 | 0.56 | 1.63 | 1.35 | 1.30 |
|    | (0.10) | (0.09) | (0.10) | (0.15) | (0.08) | (0.09) |
| POET | 0.57 | 0.55 | 0.56 | 1.76 | 1.65 | 1.62 |
|     | (0.09) | (0.09) | (0.10) | (0.15) | (0.15) | (0.12) |
| SC | 1.04 | 1.01 | 1.02 | 4.76 | 4.64 | 4.65 |
|    | (0.10) | (0.09) | (0.10) | (0.07) | (0.15) | (0.13) |
| CD | 0.97 | 0.94 | 0.95 | 3.86 | 3.77 | 3.77 |
|    | (0.08) | (0.06) | (0.06) | (0.07) | (0.06) | (0.05) |
| AT | 1.07 | 1.05 | 1.06 | 6.22 | 6.25 | 6.32 |
|    | (0.10) | (0.11) | (0.11) | (0.07) | (0.06) | (0.05) |
| POET | 1.03 | 1.00 | 1.01 | 4.71 | 4.60 | 4.60 |
|     | (0.09) | (0.09) | (0.09) | (0.14) | (0.14) | (0.12) |

Comparison with SC, AT and POET estimators

Our simulation results show below that $\hat{k}^{\text{sure}}$ provides an accurate estimate of the intrinsic dimensionality $k$ in the examples considered.

4 Experiments on Synthetic Data

In this section, we consider a number of simulation cases to compare our proposed approach in terms of (a) accuracy of the SURE method in estimating $\hat{k}^{\text{sure}}$, (b) accuracy of covariance matrix estimation in operator norm, and (c) accuracy of covariance matrix estimation in Frobenius norm. The Frobenius norm ($\| \cdot \|_F$) and the operator norm ($\| \cdot \|_2$) are defined in the usual way with $\| A \|_F = \sqrt{\text{trace}(A^T A)}$ and $\| A \|_2 = s_{\text{max}}(A)$ where $s_{\text{max}}(A)$ denotes the largest singular value of $A$.

We compare our method with Principal Orthogonal complement Thresholding (POET) of Fan et al. (2013) which is based on an additive decomposition of the covariance matrix in terms of a low rank matrix and a sparse residual covariance matrix. POET estimates the factors and the loadings by thresholding the principal components of the sample covariance matrix. We also compare with adaptive thresholding (AT) of Cai and Liu (2011) which thresholds the entries of the sample covariance matrix, with the resulting
throttled estimator \( \hat{\Sigma} \) being of the form \( \hat{\Sigma}_{jj'} = S_{jj'} \mathbb{1}(|S_{jj'}| > \delta \kappa_{jj'} ) \), where \( \delta \) is the tuning parameter and \( \kappa_{jj'} \) is a threshold specific to the corresponding entry of \( S \). We choose the tuning parameter \( \delta \) by 5-fold cross-validation as suggested by Cai and Liu (2011). We also report the performance of the sample covariance (SC) matrix for comparison.

The two simulation settings considered here are described below:

1. \( y_i, i = 1, \ldots, n \) are generated from \( N_p(0, \Sigma_0) \), where \( \Sigma_0 = \Lambda_0 \Lambda_0' + \sigma_0^2 I_p \) and \( \Lambda \) is a \( p \times k \) matrix with \( (1-s) \times 100\% \) non-zero entries. Here, \( s \in (0,1) \), is the sparsity parameter. We choose different values of \( s \) that lead to moderately sparse to dense covariance matrices. The nonzeros entries are independently drawn from a standard normal distribution.

2. This setting is designed to illustrate the performance of our approach under model misspecification. We let \( \Sigma = \Lambda_0 \Lambda_0' + \Omega_0 \), where \( \Lambda_0 \) is as in simulation setting (1), but \( \Omega_0 \) is nondiagonal corresponding to the covariance matrix of an autoregressive sequence with pure error variance 0.4 and autoregressive coefficient 0.1.
Table 4: Simulation Setting 2 with $s = 0.1$.

| $k_{\text{tr}}$ | 10   | 50   |
|-----------------|------|------|
| $p$             | 250  | 500  | 1000 | 250  | 500  | 1000 |
| $k_{\text{opt}}$ | 240  | 480  | 950  | 210  | 430  | 840  |
| $\hat{k}_{\text{sure}}$ | 240  | 470  | 940  | 210  | 410  | 820  |

| Method | 0.55 | 0.56 | 0.56 | 1.72 | 1.69 | 1.66 |
|--------|------|------|------|------|------|------|
| SC     | 0.09 | 0.10 | 0.09 | 0.17 | 0.15 | 0.13 |
| CD     | 0.54 | 0.55 | 0.55 | 1.60 | 1.45 | 1.28 |
| AT     | 0.54 | 0.55 | 0.55 | 1.69 | 1.67 | 1.63 |
| POET   | 0.96 | 0.98 | 0.98 | 4.60 | 4.65 | 4.65 |

| Method | 0.90 | 0.92 | 0.92 | 3.77 | 3.78 | 3.77 |
|--------|------|------|------|------|------|------|
| SC     | 0.07 | 0.07 | 0.08 | 0.07 | 0.06 | 0.06 |
| CD     | 0.99 | 1.01 | 1.00 | 6.13 | 6.27 | 6.28 |
| AT     | 0.95 | 0.97 | 0.97 | 4.56 | 4.61 | 4.60 |
| POET   | 0.09 | 0.10 | 0.10 | 0.14 | 0.14 | 0.12 |

Comparison with SC, AT and POET estimators

For each simulation setting, we choose the sample size $n = 100$, dimension $p = 250, 500, 1000$ and the true number of factors $k_{\text{tr}} = 10, 50$. For each $(n, p, k_{\text{tr}})$ triplet, we consider 100 simulation replicates. We consider two different values of the sparsity parameter $s$, $s \in \{0.1, 0.5\}$. $s = 0.1$ randomly sets 10% entries in the factor loadings matrix to 0. This corresponds to the extreme situation in which there is not a single sparse entry in the true covariance matrix. Similarly, $s = 0.5$ randomly sets 50% entries in the factor loadings matrix to 0 and corresponds to the moderately sparse regime.

To evaluate the accuracy of $\hat{k}_{\text{sure}}$, we compare with $k_{\text{opt}}$, defined as the minimizer of the true predictive risk function which assumes knowledge of the truth. For different settings, results across simulation replicates are summarized, in Tables 1, 2, 3 and 4, for two simulation settings to compare the matrix norm differences between the estimator resulting from different methods and the truth. In particular, normalized average operator norm error ($\| \cdot \|_2/p$), in the top panel, and normalized average frobenius norm error ($\| \cdot \|_F/p$), in the bottom panel, across 100 replicates are provided with standard error in paranthesis. The tables indicate that the SURE method is very accurate in estimating $k_{\text{opt}}$. From Tables 1 and 2 it becomes evident as the number of model parameters increases and the sparsity reduces, the
Figure 1: Operator(op) and Frobenius(fro) norm error comparisons of $\hat{\Sigma}_{CD}$ averaged over 100 replicates for different levels of sparsity for $n = 100$ and $p = 250$

performance of AT and POET deteriorates, in terms of both operator and frobenius norms, due to the sparsity assumption on both estimators, while the CD estimator has more robust performance. Even in Tables 3 and 4, where the truth is misspecified for both CD and AT, and in fact designed to favor POET, CD performs better than its competitors. We also evaluate the performance of CD estimator across various levels of sparsity on the low-rank structure for fixed sample size $n$ and dimension $p$. Figure 1 displays the superior performance of CD estimator in dense ($s = 0.1$) to moderately sparse ($s = 0.7$) regimes.

5 Discussion

In this article, we developed a simple but useful method for estimating covariance matrices with dense low-rank structures under the assumption of low intrinsic dimensionality. We also provide a principled framework for choosing the size of the low-rank dimension using SURE theory. We observe excellent performances of the proposed method in terms of scalability to high dimensions and capability of dealing with model misspecification. From Fig. 1, we see that the CD estimator outperforms AT and POET in dense ($s = 0.1$) to moderately sparse ($s = 0.7$) regimes.

6 Supporting Documents

The proofs of Lemma 3.1 and Theorem 3.2 are deferred to Appendices A and B respectively.
Appendix A: Proof of Lemma 3.1

We obtain unbiased estimators of Eqs. 3.2, 3.3, and 3.4. Suppose \( \{X_i\}_{i=1}^{n} \) is a random sample from \( N(\mu, \Sigma) \) where, without loss of generality, we let \( \mu = 0 \). We have,

\[
\mathbb{E}(\hat{\sigma}_{ij}^2) = \frac{n}{n-1} \sigma_{ij}^2 + \frac{\sigma_{ii} \sigma_{jj}}{n-1}
\]  

(A.1)

and,

\[
\mathbb{E}(\hat{\sigma}_{ii}^2 \hat{\sigma}_{jj}^2) = \frac{n+1}{n-1} \sigma_{ii} \sigma_{jj} + \frac{2(n+2)}{n(n-1)} \sigma_{ij}^2.
\]  

(A.2)

Equations A.1 and A.2 are obtained from Yi and Zou (2013). Solving (A.1) and (A.2) simultaneously, we obtain unbiased estimators for \( \sigma_{ij}^2 \) and \( \sigma_{ii} \sigma_{jj} \) below

\[
\mathbb{E} \left[ \frac{n(n^2-1)}{n^3 + n^2 - 2n - 4} (\hat{\sigma}_{ij}^2)^2 - \frac{n(n-1)}{n^3 + n^2 - 2n - 4} \hat{\sigma}_{ij}^2 \hat{\sigma}_{jj}^2 \right] = \sigma_{ij}^2.
\]  

(A.3)

\[
\mathbb{E} \left[ \frac{2(n-1)(n+2)}{2n + 4 - n^3 - n^2} (\hat{\sigma}_{ij}^2)^2 - \frac{n^2(n-1)}{2n + 4 - n^3 - n^2} \hat{\sigma}_{ii}^2 \hat{\sigma}_{jj}^2 \right] = \sigma_{ii} \sigma_{jj}.
\]  

(A.4)

An unbiased estimator of \( \text{Var}(\hat{\sigma}_{ij}^2) \) is given by

\[
\hat{\text{Var}}(\hat{\sigma}_{ij}^2) = \frac{\hat{\sigma}_{ij}^2 + \hat{\sigma}_{ii} \hat{\sigma}_{jj}}{n-1}.
\]  

(A.5)

Substituting (A.3) and (A.4) in (A.5) gives (3.2).

Equation 3.3 is obtained from Eq. 3.2 trivially. To obtain (3.4), note that

\[
\text{Cov}(\hat{\sigma}_{jj}^2, \hat{\sigma}_{ii}^2) = \frac{2(n+2)}{n(n-1)} \sigma_{ij}^2 + \frac{2}{n-1} \sigma_{ii} \sigma_{jj}.
\]  

(A.6)

Substituting the unbiased estimators of \( \sigma_{ij}^2 \) and \( \sigma_{ii} \sigma_{jj} \) in Eq. A.6, obtained from Eqs. A.3 and A.4, gives us Eq. 3.4. This completes the proof.

Appendix B: Proof of Theorem 3.2

We analyze each term in Eq. 3.1 one at a time. Consider \( \|\hat{\Sigma}_{CD}(k) - \hat{\Sigma}\|_F^2 \), a natural unbiased estimator of \( \mathbb{E}\|\hat{\Sigma}_{CD}(k) - \hat{\Sigma}\|_F^2 \). Then

\[
\|\hat{\Sigma}_{CD}(k) - \hat{\Sigma}\|_F^2 = \sum_{i,j}^p [(\eta - 1)\hat{\sigma}_{ij} + \gamma I(i = j)]^2
\]

\[
= \sum_{i \neq j}^p (\eta - 1)^2 \hat{\sigma}_{ij}^2 + \sum_{i = j}^p [(\eta - 1)\hat{\sigma}_{ii} + \gamma]^2
\]

\[
= (\eta - 1)^2 \|\hat{\Sigma} \circ \hat{\Sigma}\|_F + p\gamma^2 + 2\gamma(\eta - 1)\text{Tr}(\hat{\Sigma}),
\]  

(B.1)
where (B.1) is obtained by noting that
\[
\sum_{i \neq j} \tilde{\sigma}_{ij}^2 = \sum_{i=1}^{p} \sum_{j \neq i}^{p} \tilde{\sigma}_{ij}^2 = \|\tilde{\Sigma} \circ \tilde{\Sigma}\|_F - \text{Tr}(\tilde{\Sigma} \circ \tilde{\Sigma})
\]. Consider optimism, the third term on the right hand side of Eq. 3.1. Then

\[
\text{optimism } = \sum_{i \neq j}^{p} \eta \text{Var}(\tilde{\sigma}_{ij}) + \sum_{i=1}^{p} \left\{ \eta \text{Var}(\tilde{\sigma}_{ii}) + \gamma \text{cov} \left( \sum_{l=1}^{p} \tilde{\sigma}_{ll} \right) \right\} \\
= \sum_{i \neq j}^{p} \eta \text{Var}(\tilde{\sigma}_{ij}) + \sum_{i=1}^{p} \left\{ \eta + \gamma \right\} \text{Var}(\tilde{\sigma}_{ii}) + \gamma \sum_{l=1}^{p} \text{cov} (\tilde{\sigma}_{ll}, \tilde{\sigma}_{ii}) \right\}.
\]

Using Lemma 3.1, we have

\[
\text{optimism } = \eta a_n \sum_{i \neq j}^{p} \sigma_{ij}^2 + \eta b_n \sum_{i=1}^{p} \tilde{\sigma}_{ii} + c_n (\gamma + \eta) \sum_{i=1}^{p} \sigma_{ii}^2 + d_n \gamma \sum_{i \neq l}^{p} \tilde{\sigma}_{il} + e_n \sum_{i \neq l}^{p} \tilde{\sigma}_{il} \tilde{\sigma}_{ll} \\
= 2 \left\{ (a_n \eta + d_n \gamma) (\|\tilde{\Sigma} \circ \tilde{\Sigma}\|_F - \text{Tr}(\tilde{\Sigma} \circ \tilde{\Sigma})) + (b_n \eta + c_n \gamma) \right\} \left( \text{Tr}(\tilde{\Sigma})^2 - \text{Tr}(\tilde{\Sigma} \circ \tilde{\Sigma}) \right) + c_n (\eta + \gamma) \text{Tr}(\tilde{\Sigma} \circ \tilde{\Sigma}) \right\},
\]

where (B.3) is obtained by writing \( \sum_{i=1}^{p} \tilde{\sigma}_{ii} \sum_{j \neq i}^{p} \tilde{\sigma}_{jj} = \text{Tr}(\tilde{\Sigma})^2 - \text{Tr}(\tilde{\Sigma} \circ \tilde{\Sigma}) \) and

\[
\sum_{i=1}^{p} \tilde{\sigma}_{ii}^2 = \text{Tr}(\tilde{\Sigma} \circ \tilde{\Sigma}).
\]

The proof is completed by combining (B.1) and (B.3) to obtain an unbiased estimator of the Frobenius risk of \( \hat{\Sigma}_{CD}(k) \).

References

Anderson, T. (1984). *Multivariate Statistical Analysis*. Wiley, New York.

Bahmani, S. and Romberg, J. (2015). Sketching for simultaneously sparse and low-rank covariance matrices. In 2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP). IEEE, pp. 357–360.

Bai, J., Shi, S., et al. (2011). Estimating high dimensional covariance matrices and its applications. *Ann. Econ. Financ.* 12, 199–215.

Basu, S., Michailidis, G., et al. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Stat.* 43, 1535–1567.

Bhattacharya, A. and Dunson, D.B. (2011). Sparse bayesian infinite factor models. *Biometrika* 98, 291.

Bickel, P.J. and Levina, E. (2008a). Covariance regularization by thresholding. *Ann. Stat.*, 2577–2604.
BICKEL, P.J. and LEVINA, E. (2008b). Regularized estimation of large covariance matrices. Ann. Stat., 199–227.

BIYOUCA-DIAS, J.M., COHEN, D. and ELDAR, Y.C. (2014). Covalsa: Covariance estimation from compressive measurements using alternating minimization. In 2014 Proceedings of the 22nd European on Signal Processing Conference (EUSIPCO). IEEE, pp. 999–1003.

BUNEA, F. and XIAO, L. (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca. Bernoulli 21, 1200–1230.

BUTTE, A.J., TAMAYO, P., SLONIM, D., GOLUB, T.R. and KOHANE, I.S. (2000). Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. Proc. Natl. Acad. Sci. 97, 12182–12186.

CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. J. Am. Stat. Assoc. 106, 672–684.

CAI, T.T., MA, Z., WU, Y., et al. (2013). Sparse pca: Optimal rates and adaptive estimation. Ann. Stat. 41, 3074–3110.

CAI, T., MA, Z. and WU, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. Probab. Theory Related Fields 161, 781–815.

CAI, T.T., REN, Z., ZHOU, H.H., et al. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. Electron. J. Statist. 10, 1–59.

CHEN, X., XU, M., WU, W.B., et al. (2013). Covariance and precision matrix estimation for high-dimensional time series. Ann. Stat. 41, 2994–3021.

CHEN, Y., CHI, Y. and GOLDSMITH, A.J. (2015). Exact and stable covariance estimation from quadratic sampling via convex programming. IEEE Trans. Inf. Theory 61, 4034–4059.

DASARATHY, G., SHAH, P., BHASKAR, B.N. and NOWAK, R.D. (2015). Sketching sparse matrices, covariances, and graphs via tensor products. IEEE Trans. Inf. Theory 61, 1373–1388.

D’ASPREMONT, A., BANERJEE, O. and EL GHAOUI, L. (2008). First-order methods for sparse covariance selection. SIAM J. Matrix Anal. Appl. 30, 56–66.

DAVENPORT, M.A. and ROMBERG, J. (2016). An overview of low-rank matrix recovery from incomplete observations. IEEE J. Selected Topics Signal Process. 10, 608–622.

EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? J. Am. Stat. Assoc. 81, 461–470.

EFRON, B. (2004). The estimation of prediction error. J. Am. Stat. Assoc. 99, 467.

ENGLE, R. and WATSON, M. (1981). A one-factor multivariate time series model of metropolitan wage rates. J. Am. Stat. Assoc. 76, 774–781.

FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. J. R. Statist. Soc. Series B (Statist. Methodol.) 75, 603–680.

FAN, J., LIAO, Y. and LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. Econ. J. 19, C1–C32.

FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. J. Econ. 147, 186–197.

FURRER, R. and BENGTSSON, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. J. Multivar. Anal. 98, 227–255.

FURRER, R., GENTON, M.G. and NYCHKA, D. (2006). Covariance tapering for interpolation of large spatial datasets. J. Comput. Graph. Stat. 15, 502–523.
GOLDFARB, D. and IYENGAR, G. (2003). Robust portfolio selection problems. Math. Oper. Res. 28, 1–38.

GUHANIYOGI, R. and DUNSON, D.B. (2013). Bayesian compressed regression. arXiv:1303.0642.

HAMILL, T.M., WHITAKER, J.S. and SNYDER, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. Mon. Weather. Rev. 129, 2776–2790.

HOFF, P.D. (2009). A hierarchical eigenmodel for pooled covariance estimation. J. R. Statist. Soc. Series B (Statist. Methodol.) 71, 971–992.

HOUTEKAMER, P.L. and MITCHELL, H.L. (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. Mon. Weather. Rev. 129, 123–137.

HUANG, J.Z., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. Biometrika, 85–98.

IDEKER, T. and SHARAN, R. (2008). Protein networks in disease. Genome Res. 18, 644–652.

JIMENEZ-SANchez, G., CHILDs, B. and VALLE, D. (2001). Human disease genes. Nature 409, 853–855.

JOHNSTOne, I.M. (2001). On the distribution of the largest eigenvalue in principal components analysis. Ann. Statist, 295–327.

JOHNSTOne, I.M. and LU, A.Y. (2012). On consistency and sparsity for principal components analysis in high dimensions. J. Amer. Statist. Assoc.

KAROUI, N.E. (2008a). Operator norm consistent estimation of large-dimensional sparse covariance matrices. Ann. Stat., 2717–2756.

KAROUI, N.E. (2008b). Spectrum estimation for large dimensional covariance matrices using random matrix theory. Ann. Stat., 2757–2790.

KAUFMAN, C.G., SCHERVISH, M.J. and NYCHKA, D.W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. J. Am. Stat. Assoc. 103, 1545–1555.

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. Ann. Statist. 37, 4254.

LEDOIT, O., SANTA-CLARA, P. and WOLF, M. (2003). Flexible multivariate Garch modeling with an application to international stock markets. Rev. Econ. Statist. 85, 735–747.

LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. 88, 365–411.

LI, D. and ZOU, H. (2014). Asymptotic properties of sure information criteria for large covariance matrices. arXiv:1406.6514.

LI, D. and ZOU, H. (2016). Sure information criteria for large covariance matrix estimation and their asymptotic properties. IEEE Trans. Inf. Theory 62, 2153–2169.

MA, Z., et al. (2013). Sparse principal component analysis and iterative thresholding. Ann. Stat. 41, 772–801.

MARZETTA, T.L., TUCCI, G.H. and SIMON, S.H. (2011). A random matrix-theoretic approach to handling singular covariance estimates. IEEE Trans. Inform. Theory 57, 6256–6271.

MCMURRY, T.L. and POLITIS, D.N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. J. Time Ser. Anal. 31, 471–482.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. Ann. Statist, 1436–1462.

PATI, D., BHATTACHARYA, A., PILLAI, N.S. and DUNSON, D.B. (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. Ann. Stat. 42, 1102–1130.
COMPRESSED COVARIANCE ESTIMATION WITH AUTOMATED DIMENSION LEARNING 481

POURAHMADI, M. (2011). Covariance estimation: The glm and regularization perspectives. Stat. Sci., 369–387.

QIAO, H. and PAL, P. (2015). Generalized nested sampling for compressing low rank Toeplitz matrices. IEEE Signal Process. Lett. 22, 1844–1848.

RAVIKUMAR, P., WAINWRIGHT, M.J., RASKUTTI, G., YU, B., et al. (2011). High-dimensional covariance estimation by minimizing $l_1$-penalized log-determinant divergence. Electron. J. Statist. 5, 935–980.

ROTHMAN, A.J., LEVINA, E. and ZHU, J. (2009). Generalized thresholding of large covariance matrices. J. Am. Stat. Assoc. 104, 177–186.

SCHA¨FER, J., STRIMMER, K., et al. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statist. Appl. Genetics Molecul. Biol. 4, 32.

SHEN, H. and HUANG, J.Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. J. Multivar. Anal. 99, 1015–1034.

SMITH, M. and KOHN, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. J. Am. Stat. Assoc. 97, 1141–1153.

STEIN, C., et al. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Proceedings of the Third Berkeley symposium on mathematical statistics and probability, vol 1, pp. 197–206.

STEIN, C.M. (1981). Estimation of the mean of a multivariate normal distribution. Ann. Statist., 1135–1151.

WERNER, K., JANSSON, M. and STOICA, P. (2008). On estimation of covariance matrices with kronecker product structure. IEEE Trans. Signal Process. 56, 478–491.

WITTEN, D.M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics, kxp008.

WU, W.B. and POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. Biometrika 90, 831–844.

WU, W.B. and POURAHMADI, M. (2009). Banding sample autocovariance matrices of stationary processes. Stat. Sin., 1755–1768.

XIAO, H., WU, W.B., et al. (2012). Covariance matrix estimation for stationary time series. Ann. Stat. 40, 466–493.

XIAO, L. and BUNEA, F. (2014). On the theoretic and practical merits of the banding estimator for large covariance matrices. arXiv:1402.0844.

YI, F. and ZOU, H. (2013). Sure-tuned tapering estimation of large covariance matrices. Comput. Statist. Data Anal. 58, 339–351.

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the gaussian graphical model. Biometrika, 19–35.

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. J. Comput. Graph. Statist. 15, 265–286.

Gautam Sabnis
UNIVERSITY OF MICHIGAN, ANN ARBOR, USA
E-mail: gautamsabnis15@gmail.com

Debdeep Pati
ANIRBAN BHATTACHARYA
TEXAS A&M UNIVERSITY, COLLEGE STATION, USA

Paper received: 13 April 2017.