Regularization-Free Estimation in Trace Regression with Symmetric Positive Semidefinite Matrices

Martin Slawski  
Department of Statistics & Biostatistics  
Department of Computer Science  
Rutgers University  
Piscataway, NJ 08854, USA  
{martin.slawski@rutgers.edu, pingli@stat.rutgers.edu}

Matthias Hein  
Department of Computer Science  
Department of Mathematics  
Saarland University  
Saarbrücken, Germany  
hein@cs.uni-saarland.de

Abstract

Trace regression models have received considerable attention in the context of matrix completion, quantum state tomography, and compressed sensing. Estimation of the underlying matrix from regularization-based approaches promoting low-rankness, notably nuclear norm regularization, have enjoyed great popularity. In this paper, we argue that such regularization may no longer be necessary if the underlying matrix is symmetric positive semidefinite (\(\text{spd}\)) and the design satisfies certain conditions. In this situation, simple least squares estimation subject to an \(\text{spd}\) constraint may perform as well as regularization-based approaches with a proper choice of regularization parameter, which entails knowledge of the noise level and/or tuning. By contrast, constrained least squares estimation comes without any tuning parameter and may hence be preferred due to its simplicity.

1 Introduction

Trace regression models of the form

\[
y_i = \text{tr}(X_i^\top \Sigma^*) + \epsilon_i, \quad i = 1, \ldots, n,
\]

where \(\Sigma^* \in \mathbb{R}^{m_1 \times m_2}\) is the parameter of interest to be estimated given measurement matrices \(X_i \in \mathbb{R}^{m_1 \times m_2}\) and observations \(y_i\) contaminated by errors \(\epsilon_i, \quad i = 1, \ldots, n\), have attracted considerable interest in high-dimensional statistical inference, machine learning and signal processing over the past few years. Research in these areas has focused on a setting with few measurements \(n \ll m_1 \cdot m_2\) and \(\Sigma^*\) being (approximately) of low rank \(r \ll \min\{m_1, m_2\}\). Such setting is relevant to problems such as matrix completion [6, 23], compressed sensing [5, 17], quantum state tomography [11] and phase retrieval [7]. A common thread in these works is the use of the nuclear norm of a matrix as a convex surrogate for its rank [18] in regularized estimation amenable to modern optimization techniques. This approach can be seen as natural generalization of \(\ell_1\)-norm (aka lasso) regularization for the linear regression model [24] that arises as a special case of model (1) in which both \(\Sigma^*\) and \(\{X_i\}_{i=1}^n\) are diagonal. It is inarguable that in general regularization is essential if \(n < m_1 \cdot m_2\).

The situation is less clear if \(\Sigma^*\) is known to satisfy additional constraints that can be incorporated in estimation. Specifically, in the present paper we consider the case in which \(m_1 = m_2 = m\) and \(\Sigma^*\) is known to be symmetric positive semidefinite (\(\text{spd}\)), i.e. \(\Sigma^* \in S^m_{+}\) with \(S^m_{+}\) denoting the positive semidefinite cone in the space of symmetric real \(m \times m\) matrices \(S^m\). The set \(S^m_{+}\) deserves specific interest as it includes covariance matrices and Gram matrices in kernel-based learning [20]. It is rather common for these matrices to be of low rank (at least approximately), given the widespread use of principal components analysis and low-rank kernel approximations [28]. In the present paper, we focus on the usefulness of the \(\text{spd}\) constraint for estimation. We argue that if \(\Sigma^*\) is \(\text{spd}\) and the measurement matrices \(\{X_i\}_{i=1}^n\) obey certain conditions, constrained least squares estimation

\[
\min_{\Sigma \in S^m_{+}} \frac{1}{2n} \sum_{i=1}^n (y_i - \text{tr}(X_i^\top \Sigma))^2
\]

may perform similarly well in prediction and parameter estimation as approaches employing nuclear norm regularization with proper choice of the regularization parameter, including the interesting...
regime $n < \delta_m$, where $\delta_m = \dim(S^m) = m(m + 1)/2$. Note that the objective in (2) only consists of a data fitting term and is hence convenient to work with in practice since there is no free parameter. Our findings can be seen as a non-commutative extension of recent results on non-negative least squares estimation for linear regression [16, 21].

**Related work.** Model (1) with $\Sigma^* \in S^m_+$ has been studied in several recent papers. A good deal of these papers consider the setup of compressed sensing in which the $\{X_i\}_{i=1}^n$ can be chosen by the user, with the goal to minimize the number of observations required to (approximately) recover $\Sigma^*$. For example, in [27], recovery of $\Sigma^*$ being low-rank from noiseless observations ($\epsilon_i = 0$, $i = 1, \ldots, n$) by solving a feasibility problem over $S^n_+$ is considered, which is equivalent to the constrained least squares problem (1) in a noiseless setting.

In [3, 8], recovery from rank-one measurements is considered, i.e., for $\{x_i\}_{i=1}^n \subset \mathbb{R}^m$

$$y_i = x_i^T \Sigma^* x_i + \epsilon_i = \text{tr}(X_i^T \Sigma^*) + \epsilon_i, \quad \text{with } X_i = x_i x_i^T, \quad i = 1, \ldots, n.$$  \hspace{1cm} (3)

As opposed to [3, 8], where estimation based on nuclear norm regularization is proposed, the present work is devoted to regularization-free estimation. While rank-one measurements as in (3) are also in the center of interest herein, our framework is not limited to this case. In [3] an application of (3) to covariance matrix estimation given only one-dimensional projections $\{x_i^T z_i\}_{i=1}^n$ of the data points is discussed, where the $\{z_i\}_{i=1}^n$ are i.i.d. from a distribution with zero mean and covariance matrix $\Sigma^*$. In fact, this fits the model under study with observations

$$y_i = (x_i^T z_i)^2 = x_i^T z_i z_i^T x_i = x_i^T \Sigma^* x_i + \epsilon_i, \quad \epsilon_i = x_i^T \{z_i z_i^T - \Sigma^*\} x_i, \quad i = 1, \ldots, n.$$ \hspace{1cm} (4)

Specializing (3) to the case in which $\Sigma^* = \sigma^*(\sigma^*)^T$, one obtains the quadratic model

$$y_i = ||x_i||^2 + \epsilon_i,$$ \hspace{1cm} (5)

which (with complex-valued $\sigma^*$) is relevant to the problem of phase retrieval [14]. The approach of [7] treats (5) as an instance of (1) and uses nuclear norm regularization to enforce rank-one solutions. In follow-up work [4], the authors show a refined recovery result stating that imposing an $\text{spd}$ constraint — without regularization — suffices. A similar result has been proven independently by [10]. However, the results in [4] and [10] only concern model (5). After posting an extended version [22] of the present paper, a generalization of [4, 10] to general low-rank $\text{spd}$ matrices has been achieved in [13]. Since [4, 10, 13] consider bounded noise, whereas the analysis herein assumes Gaussian noise, our results are not directly comparable to those in [4, 10, 13].

**Notation.** $M^d$ denotes the space of real $d \times d$ matrices with inner product $\langle M, M' \rangle := \text{tr}(M^T M')$. The subspace of symmetric matrices $S^d$ has dimension $\delta_d := d(d + 1)/2$. $M \in S^d$ has an eigen-decomposition $M = U \Lambda U^T = \sum_{j=1}^d \lambda_j(M) u_j u_j^T$, where $\lambda_1(M) = \lambda_{\text{max}}(M) \geq \lambda_2(M) \geq \ldots \geq \lambda_d(M) = \lambda_{\text{min}}(M)$, $U = [u_1 \ldots u_d]$. For $q \in [1, \infty)$ and $M \in S^d$, $\|M\|_q := \left(\sum_{j=1}^d \lambda_j(M)^q\right)^{1/q}$ denotes the Schatten-$q$-norm ($q = 1$: nuclear norm, $q = 2$: Frobenius norm $\|M\|_F$, $q = \infty$: spectral norm $\|M\|_{\infty} := \max_{1 \leq j \leq d} |\lambda_j(M)|$).

Let $S_1(d) = \{M \in S^d : \|M\|_1 = 1\}$ and $S_+^d(d) = S_1(d) \cap S^d_+$. The symbols $\preceq, \succeq, \succ, \prec$ refer to the semidefinite ordering. For a set $A$ and $\alpha \in \mathbb{R}$, $\alpha A := \{\alpha a : a \in A\}$.

It is convenient to re-write model (1) as $y = X(\Sigma^*) + \epsilon$, where $y = (y_i)_{i=1}^n$, $\epsilon = (\epsilon_i)_{i=1}^n$ and $X : M^m \to \mathbb{R}^n$ is a linear map defined by $(X(M))_i = \text{tr}(X_i^T M)$, $i = 1, \ldots, n$, referred to as the sampling operator. Its adjoint $X^* : \mathbb{R}^n \to M^m$ is given by the map $v \mapsto \sum_{i=1}^n v_i X_i$.

**Supplement.** The appendix contains all proofs, additional experiments and figures.

## 2 Analysis

**Preliminaries.** Throughout this section, we consider a special instance of model (1) in which

$$y_i = \text{tr}(X_i \Sigma^*) + \epsilon_i, \quad \text{where } \Sigma^* \in S^m_+, \quad X_i \in S^m, \quad \text{and } \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n.$$ \hspace{1cm} (6)

The assumption that the errors $\{\epsilon_i\}_{i=1}^n$ are Gaussian is made for convenience as it simplifies the stochastic part of our analysis, which could be extended to sub-Gaussian errors.

Note that w.l.o.g., we may assume that $\{X_i\}_{i=1}^n \subset S^m$. In fact, since $\Sigma^* \in S^m_+$, for any $M \in M^m$ we have that $\text{tr}(M \Sigma^*) = \text{tr}(M^m \Sigma^*)$, where $M^m = (M + M^T)/2$.  

2
In the sequel, we study the statistical performance of the constrained least squares estimator

\[ \hat{\Sigma} \in \arg\min_{\Sigma \in S_+^m} \frac{1}{2n} \| y - \mathcal{X}(\Sigma) \|_2^2 \quad (7) \]

under model (6). More specifically, under certain conditions on \( \mathcal{X} \), we shall derive bounds on

\[ (a) \quad \frac{1}{n}\| \mathcal{X}(\Sigma^*) - \mathcal{X}(\hat{\Sigma}) \|_2^2, \quad \text{and} \quad (b) \quad \| \hat{\Sigma} - \Sigma^* \|_1, \quad (8) \]

where \((a)\) will be referred to as “prediction error” below. The most basic method for estimating \( \Sigma^* \) is ordinary least squares (ols) estimation

\[ \hat{\Sigma}^{ols} \in \arg\min_{\Sigma \in S_+^m} \frac{1}{2n} \| y - \mathcal{X}(\Sigma) \|_2^2, \quad (9) \]

which is computationally simpler than (7). While (7) requires convex programming, (9) boils down to solving a linear system of equations in \( \delta_m \rightarrow 0 \) as \( n \rightarrow \infty \). Moreover, the estimation error \( \| \hat{\Sigma}^{ols} - \Sigma^* \|_1 \) is unbounded unless \( n \geq \delta_m \). Research conducted over the past few years has thus focused on methods dealing successfully with the case \( n < \delta_m \) as long as the target \( \Sigma^* \) has additional structure, notably low-rankedness. Indeed, if \( \Sigma^* \) has rank \( r \ll m \), the intrinsic dimension of the problem becomes (roughly) \( mr \ll \delta_m \). In a large body of work, nuclear norm regularization, which serves as a convex surrogate of rank regularization, is considered as a computationally convenient alternative for which a series of adaptivity properties to underlying low-rankedness has been established, e.g. \cite{5, 15, 17, 18, 19}. Complementing (9) with nuclear norm regularization yields the estimator

\[ \hat{\Sigma}^1 \in \arg\min_{\Sigma \in S_+^m} \frac{1}{2n} \| y - \mathcal{X}(\Sigma) \|_2^2 + \lambda \| \Sigma \|_1, \quad (10) \]

where \( \lambda > 0 \) is a regularization parameter. In case an spd constraint is imposed (10) becomes

\[ \hat{\Sigma}^{1+} \in \arg\min_{\Sigma \in S_+^m} \frac{1}{2n} \| y - \mathcal{X}(\Sigma) \|_2^2 + \lambda \text{tr}(\Sigma), \quad (11) \]

Our analysis aims at elucidating potential advantages of the spd constraint in the constrained least squares problem (7) from a statistical point of view. It turns out that depending on properties of \( \mathcal{X} \), the behaviour of \( \hat{\Sigma} \) can range from a performance similar to the least squares estimator \( \hat{\Sigma}^{ols} \) on the one hand to a performance similar to the nuclear norm regularized estimator \( \hat{\Sigma}^1 \) with properly chosen/tuned \( \lambda \) on the other hand. The latter case appears to be remarkable: \( \hat{\Sigma} \) may enjoy similar adaptivity properties as nuclear norm regularized estimators even though \( \hat{\Sigma} \) is obtained from a pure data fitting problem without explicit regularization.

### 2.1 Negative result

We first discuss a negative example of \( \mathcal{X} \) for which the spd-constrained estimator \( \hat{\Sigma} \) does not improve (substantially) over the unconstrained estimator \( \hat{\Sigma}^{ols} \). At the same time, this example provides clues on conditions to be imposed on \( \mathcal{X} \) to achieve substantially better performance.

**Random Gaussian design.** Consider the Gaussian orthogonal ensemble (GOE)

\[ \text{GOE}(m) = \{ X = (x_{jk}), \{ x_{jk} \}_{j=1}^m \overset{\text{i.i.d.}}{\sim} N(0, 1), \{ x_{jk} = x_{kj} \}_{1 \leq j < k \leq m} \overset{\text{i.i.d.}}{\sim} N(0, 1/2) \}. \]

Gaussian measurements are common in compressed sensing. It is hence of interest to study measurements \( \{ X_i \}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \text{GOE}(m) \) in the context of the constrained least squares problem (7). The following statement points to a serious limitation associated with such measurements.

**Proposition 1.** Consider \( X_i \overset{\text{i.i.d.}}{\sim} \text{GOE}(m) \), \( i = 1, \ldots, n \). For any \( \varepsilon > 0 \), if \( n \leq (1 - \varepsilon)\delta_m / 2 \), with probability at least \( 1 - 32 \exp(-\varepsilon^2 \delta_m) \), there exists \( \Delta \in S_+^m \), \( \Delta \neq 0 \) such that \( \mathcal{X}(\Delta) = 0 \).

Proposition 1 implies that if the number of measurements drops below 1/2 of the ambient dimension \( \delta_m \), estimating \( \Sigma^* \) based on (7) becomes ill-posed; the estimation error \( \| \hat{\Sigma} - \Sigma^* \|_1 \) is unbounded, irrespective of the rank of \( \Sigma^* \). Geometrically, the consequence of Proposition 1 is that the convex cone \( \mathcal{C}_X = \{ z \in \mathbb{R}^m : z = \mathcal{X}(\Delta), \Delta \in S_+^m \} \) contains 0. Unless 0 is contained in the boundary of \( \mathcal{C}_X \) (we conjecture that this event has measure zero), this means that \( \mathcal{C}_X = \mathbb{R}^m \), i.e. the spd constraint becomes vacuous.
2.2 Slow Rate Bound on the Prediction Error

We present a positive result on the spd-constrained least squares estimator \( \hat{\Sigma} \) under an additional condition on the sampling operator \( \mathcal{X} \). Specifically, the prediction error will be bounded as

\[
\frac{1}{n} \| \mathcal{X}(\Sigma^*) - \mathcal{X}(\hat{\Sigma}) \|^2 = O(\lambda_0(\Sigma^*)_1 + \lambda_0^2), \quad \text{where} \quad \lambda_0 = \frac{1}{n} \| \mathcal{X}(\varepsilon) \|_\infty, \tag{12}
\]

with \( \lambda_0 \) typically being of the order \( O(\sqrt{m/n}) \) (up to log factors). The rate in (12) can be a significant improvement of what is achieved by \( \hat{\Sigma}_{\text{ols}} \) if \( \| \Sigma^* \|_1 = \text{tr}(\Sigma^*) \) is small. If \( \lambda_0 = o(\| \Sigma^* \|_1) \) that rate coincides with those of the nuclear norm regularized estimators (10), (11) with regularization parameter \( \lambda \geq \lambda_0 \), cf. Theorem 1 in [19]. For nuclear norm regularized estimators, the rate \( O(\lambda_0^2(\Sigma^*)_1) \) is achieved for any choice of \( \mathcal{X} \) and is slow in the sense that the squared prediction error only decays at the rate \( n^{-1/2} \) instead of \( n^{-1} \).

**Condition on \( \mathcal{X} \).** In order to arrive at a suitable condition to be imposed on \( \mathcal{X} \) so that (12) can be achieved, it makes sense to re-consider the negative example of Proposition 1, which states that for any

\[
\text{Condition 1. There exist constants } R_*, \tau_* > 0 \text{ s.t. } \tau^2(\mathcal{X}, R_*) \geq \tau_*^2, \quad \text{where for } R \in \mathbb{R}
\]

\[
\tau^2(\mathcal{X}, R) = \text{dist}(\mathcal{P}_X, 0) = \min_{\Delta \in \mathfrak{S}_+^n(m)} \| \mathcal{X}(\Delta) \|_2 = 0, \quad \text{where}
\]

\[
\mathcal{P}_X := \{ z \in \mathbb{R}^n : z = \mathcal{X}(\Delta), \ \Delta \in \mathfrak{S}_+^n(m) \}, \quad \text{and } \mathfrak{S}_+^n(m) := \{ \Delta \in \mathbb{S}_+^n : \text{tr}(\Delta) = 1 \}.
\]

In this situation, it is impossible to derive a non-trivial bound on the prediction error as dist(\( \mathcal{P}_X, 0 \)) = 0 may imply \( \mathcal{C}_X = \mathbb{R}^n \) so that \( \mathcal{X}(\Sigma^*) - \mathcal{X}(\hat{\Sigma}) \|^2 = \| \varepsilon \|^2 \). To rule this out, the condition dist(\( \mathcal{P}_X, 0 \)) > 0 is natural. More strongly, one may ask for the following:

There exists a constant \( \tau > 0 \) such that \( \tau_0^2(\mathcal{X}) := \min_{\Delta \in \mathfrak{S}_+^n(m)} \frac{1}{n} \| \mathcal{X}(\Delta) \|_2^2 \geq \tau^2. \tag{13} \)

An analogous condition is sufficient for a slow rate bound in the vector case, cf. [21]. However, the condition for the slow rate bound in Theorem 1 below is somewhat stronger than (13).

**Condition 1.** There exist constants \( R_*, \tau_* > 0 \) s.t. \( \tau^2(\mathcal{X}, R_*) \geq \tau_*^2 \), for all \( R \in \mathbb{R} \)

\[
\tau^2(\mathcal{X}, R) = \text{dist}(\mathcal{P}_X, 0) = \min_{\Delta \in \mathfrak{S}_+^n(m)} \| \mathcal{X}(\Delta) \|_2 = 0, \quad \text{where}
\]

\[
\mathcal{P}_X := \{ z \in \mathbb{R}^n : z = \mathcal{X}(\Delta), \ \Delta \in \mathfrak{S}_+^n(m) \}, \quad \text{and } \mathfrak{S}_+^n(m) := \{ \Delta \in \mathbb{S}_+^n : \text{tr}(\Delta) = 1 \}.
\]

The following condition is sufficient for Condition 1 and in some cases much easier to check.

**Proposition 2.** Suppose there exists \( a \in \mathbb{R}^n, \| a \|_2 \leq 1, \) and constants \( 0 < \phi_{\text{min}} \leq \phi_{\text{max}} \) s.t.

\[
\lambda_{\text{min}}(n^{-1/2}(a^* \Sigma^*) a) \geq \phi_{\text{min}}, \quad \text{and} \quad \lambda_{\text{max}}(n^{-1/2}(a^* \Sigma^*) a) \leq \phi_{\text{max}}.
\]

Then for any \( \zeta > 1, \mathcal{X} \) satisfies Condition 1 with \( R_* = \zeta(\phi_{\text{max}}/\phi_{\text{min}}) \) and \( \tau_*^2 = (\zeta - 1)^2 \phi_{\text{max}}^2 \).

The condition of Proposition 2 can be phrased as having a positive definite matrix in the image of the unit ball under \( \mathcal{X}^* \), which, after scaling by \( 1/\sqrt{n} \), has its smallest eigenvalue bounded away from zero and a bounded condition number. As a simple example, suppose that \( X_1 = \sqrt{n}I \). Invoking Proposition 2 with \( a = (1, 0, \ldots, 0)^T \) and \( \zeta = 2 \), we find that Condition 1 is satisfied with \( R_* = 2 \) and \( \tau_*^2 = 1 \). A more interesting example is random design where the \( \{ X_i \}_{i=1}^n \) are (sample) covariance matrices, where the underlying random vectors satisfy appropriate tail or moment conditions.

**Corollary 1.** Let \( \pi_m \) be a probability distribution on \( \mathbb{R}^m \) with second moment matrix \( \Gamma := \mathbb{E}_{z \sim \pi_m}[zz^T] \) satisfying \( \lambda_{\text{min}}(\Gamma) > 0. \) Consider the random matrix ensemble

\[
\mathcal{M}(\pi_m, q) = \left\{ \frac{1}{n} \sum_{k=1}^n z_k z_k^T, \ z_k \sim \pi_m \right\}. \tag{14}
\]

Suppose that \( \{ X_i \}_{i=1}^n \sim \mathcal{M}(\pi_m, q) \) and let \( \hat{\Gamma}_n := \frac{1}{n} \sum_{i=1}^n X_i \) and \( 0 < \epsilon_n < \lambda_{\text{min}}(\Gamma) \). Under the event \( \| \hat{\Gamma}_n - \Gamma \|_\infty \leq \epsilon_n \), \( \mathcal{X} \) satisfies Condition 1 with

\[
R_* = \frac{2(\lambda_{\text{min}}(\Gamma) + \epsilon_n)}{\lambda_{\text{min}}(\Gamma) - \epsilon_n} \quad \text{and} \quad \tau_*^2 = (\lambda_{\text{max}}(\Gamma) + \epsilon_n)^2.
\]
It is instructive to spell out Corollary 1 with $\pi_m$ as the standard Gaussian distribution on $\mathbb{R}^m$. The matrix $\hat{\Gamma}_n$ equals the sample covariance matrix computed from $N = n \cdot q$ samples. It is well-known [9] that for $m$, $N$ large, $\lambda_{\max}(\hat{\Gamma}_n)$ and $\lambda_{\min}(\hat{\Gamma}_n)$ concentrate sharply around $(1 + \eta_n)^2$ and $(1 - \eta_n)^2$, respectively, where $\eta_n = \sqrt{m/N}$. Hence, for any $\gamma > 0$, there exists $C_\gamma > 1$ so that if $N \geq C_\gamma m$, it holds that $R_\gamma \leq 2 + \gamma$. Similar though weaker concentration results for $\|\Gamma - \hat{\Gamma}_n\|_\infty$ exist for the broader class of distributions $\pi_m$ with finite fourth moments [26]. Specialized to $q = 1$, Corollary 1 yields a statement about $\mathcal{X}$ made up from random rank-one measurements $X_i = z_i z_i^\top$, $i = 1, \ldots, n$, cf. (3). The preceding discussion indicates that Condition 1 tends to be satisfied in this case.

**Theorem 1.** Suppose that model (6) holds with $\mathcal{X}$ satisfying Condition 1 with constants $R_\gamma$ and $\tau_\gamma^2$. We then have

$$\frac{1}{n} \|\mathcal{X}(\Sigma^\ast) - \mathcal{X}(\hat{\Sigma})\|_2^2 \leq \max \left\{ 2(1 + R_\gamma)\lambda_0 \|\Sigma^\ast\|_1, 2\lambda_0 \|\Sigma^\ast\|_1 + 8 \left( \lambda_0 R_\gamma \right)^2 \right\}$$

where, for any $\mu \geq 0$, with probability at least $1 - (2m)^{-\mu}$

$$\lambda_0 \leq \sigma \sqrt{(1 + \mu)2 \log(2m) \Delta^2 n}, \text{ where } V_n^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \|\|_\infty.$$  

**Remark:** Under the scalings $R_\gamma = O(1)$ and $\tau_\gamma^2 = \Omega(1)$, the bound of Theorem 1 is of the order $O(\lambda_0 \|\Sigma^\ast\|_1 + \lambda_0^2)$ as announced at the beginning of this section. For given $\mathcal{X}$, the quantity $\tau^2(\mathcal{X}, R)$ can be evaluated by solving a least squares problem with $\text{spd}$ constraints. Hence it is feasible to check in practice whether Condition 1 holds. For later reference, we evaluate the term $V_n^2$ for $\mathcal{M}(\pi_m, q)$ with $\pi_m$ as standard Gaussian distribution. As shown in the supplement, with high probability, $V_n^2 = O(m \log n)$ holds as long as $m = O(nq)$.

### 2.3 Bound on the Estimation Error

In the previous subsection, we did not make any assumptions about $\Sigma^\ast$ apart from $\Sigma^\ast \in S^m_{\text{sd}}$. Henceforth, we suppose that $\Sigma^\ast$ is of low rank $1 \leq r \ll m$ and study the performance of the constrained least squares estimator (7) for prediction and estimation in such setting.

**Preliminaries.** Let $\Sigma^\ast = UAU^\top$ be the eigenvalue decomposition of $\Sigma^\ast$, where

$$U = \begin{bmatrix} U_\parallel & U_\perp \end{bmatrix} \begin{bmatrix} m \times r & m \times (m-r) \end{bmatrix} \begin{bmatrix} \Lambda_r & 0_{r \times (m-r)} \\ 0_{(m-r) \times r} & 0_{(m-r) \times (m-r)} \end{bmatrix}$$

where $\Lambda_r$ is diagonal with positive diagonal entries. Consider the linear subspace

$$T^\perp = \{ M \in S^m : M = U_\perp U_\perp^\top, \quad A \in S^{m-r} \}.$$  

From $U_\parallel^\top \Sigma^\ast U_\perp = 0$, it follows that $\Sigma^\ast$ is contained in the orthogonal complement

$$T = \{ M \in S^m : M = U_\parallel B + B^\top U_\parallel^\top, \quad B \in \mathbb{R}^{r \times m} \},$$

of dimension $mr - r(r-1)/2 \ll \delta_m$ if $r \ll m$. The image of $T$ under $\mathcal{X}$ is denoted by $T$.

**Conditions on $\mathcal{X}$**. We introduce the key quantities the bound in this subsection depends on.

- **Separability constant.**

$$\tau^2(T) = \frac{1}{n} \text{dist}^2(T, \mathcal{P}_\Delta), \quad \mathcal{P}_\Delta := \{ z \in \mathbb{R}^m : z = \mathcal{X}(\Delta), \quad \Delta \in T^\perp \cap S^+_m(m) \} = \frac{1}{\delta_m} \min_{\Theta \in T, \Delta \in S^+_m(r(r-1)/2 \leq \delta_m)} \|\mathcal{X}(\Theta) - \mathcal{X}(\Delta)\|_2^2$$

- **Restricted eigenvalue.**

$$\phi^2(T) = \min_{0 \neq \Delta \in T} \frac{\|\mathcal{X}(\Delta)\|_2^2/n}{\|\Delta\|_F^2}.$$  

As indicated by the following statement concerning the noiseless case, for bounding $\|\hat{\Sigma} - \Sigma^\ast\|$, it is inevitable to have lower bounds on the above two quantities.
In this section, we empirically study properties of the estimator. We here empirically evaluate relative to regularization-based methods is explored. We also present an application to spiked co-

Proposition 3. Consider the trace regression model (1) with \( \varepsilon_i = 0, i = 1, \ldots, n \). Then

\[
\arg\min_{\Sigma \in \mathbb{S}_+^m} \frac{1}{2n} \left\| X(\Sigma^*) - X(\Sigma) \right\|_2^2 = \{ \Sigma^* \} \quad \text{for all } \Sigma^* \in \mathbb{T} \cap \mathbb{S}_+^m
\]

if and only if it holds that \( \tau^2(\mathbb{T}) > 0 \) and \( \phi^2(\mathbb{T}) > 0 \).

Correlation constant. Moreover, we use of the following the quantity. It is not clear to us if it is intrinsically required, or if its appearance in our bound is for merely technical reasons.

\[
\mu(\mathbb{T}) = \max \left\{ \frac{1}{n} \langle X(\Delta), X'(\Delta') \rangle : \|\Delta\|_1 \leq 1, \Delta, \Delta' \in \mathbb{S}_+^m \cap \mathbb{T} \} .
\]

We are now in position to provide a bound on \( \| \hat{\Sigma} - \Sigma^* \|_1 \).

Theorem 2. Suppose that model (6) holds with \( \Sigma^* \) as considered throughout this subsection and let \( \lambda_0 \) be defined as in Theorem 1. We then have

\[
\| \hat{\Sigma} - \Sigma^* \|_1 \leq \max \left\{ \frac{8\lambda_0}{\phi^2(\mathbb{T})} \left( \frac{3}{2} + \frac{\mu(\mathbb{T})}{\phi^2(\mathbb{T})} \right) + 4\lambda_0 \left( \frac{1}{\phi^2(\mathbb{T})} + \frac{1}{\tau^2(\mathbb{T})} \right), \frac{8\lambda_0}{\phi^2(\mathbb{T})} \left( 1 + \frac{\mu(\mathbb{T})}{\phi^2(\mathbb{T})} \right), \frac{8\lambda_0}{\tau^2(\mathbb{T})} \right\}.
\]

Rem.

Remark. Given Theorem 2 an improved bound on the prediction error scaling with \( \lambda_0^2 \) in place of \( \lambda_0 \) can be derived, cf. (26) in Appendix D.

The quality of the bound of Theorem 2 depends on how the quantities \( \tau^2(\mathbb{T}), \phi^2(\mathbb{T}) \) and \( \mu(\mathbb{T}) \) scale with \( n, m \) and \( r \), which is design-dependent. Accordingly, the estimation error in nuclear norm can be non-finite in the worst case and \( O(\lambda_0 r) \) in the best case, which matches existing bounds for nuclear norm regularization (cf. Theorem 2 in [19]).

- The quantity \( \tau(\mathbb{T}) \) is specific to the geometry of the constrained least squares problem (7) and hence of critical importance. For instance, it follows from Proposition 1 that for standard Gaussian measurements, \( \tau(\mathbb{T}) = 0 \) with high probability once \( n < \delta_m/2 \). The situation can be much better for random spd measurements (14) as exemplified for measurements \( X_i = z_i z_i^\top \) with \( z_i \sim \mathcal{N}(0, I) \) in Appendix H. Specifically, it turns out that \( \tau(\mathbb{T}) = \Omega(1/r) \) as long as \( n = \Omega(m \cdot r) \).

- It is not restrictive to assume \( \phi^2(\mathbb{T}) \) is positive. Indeed, without that assumption, even an oracle estimator based on knowledge of the subspace \( \mathbb{T} \) would fail. Reasonable sampling operators \( \mathcal{X} \) have rank \( \min \{ n, \delta_m \} \) so that the nullspace of \( \mathcal{X} \) only has a trivial intersection with the subspace \( \mathbb{T} \) as long as \( n \geq \dim(\mathbb{T}) = m r - r(r - 1)/2 \).

- For fixed \( \mathbb{T} \), computing \( \mu(\mathbb{T}) \) entails solving a biconvex (albeit non-convex) optimization problem in \( \Delta \in \mathbb{T} \) and \( \Delta' \in \mathbb{S}_+^m \cap \mathbb{T} \). Block coordinate descent is a practical approach to such optimization problems for which a globally optimal solution is out of reach. In this manner we explore the scaling of \( \mu(\mathbb{T}) \) numerically as done for \( \tau^2(\mathbb{T}) \). We find that \( \mu(\mathbb{T}) = O(\delta_m/n) \) so that \( \mu(\mathbb{T}) = \Omega(1) \) apart from the regime \( n/\delta_m \rightarrow 0 \), without ruling out the possibility of undersampling, i.e. \( n < \delta_m \).

3 Numerical Results

In this section, we empirically study properties of the estimator \( \hat{\Sigma} \). In particular, its performance relative to regularization-based methods is explored. We also present an application to spiked covariance estimation for the CBCL face image data set and stock prices from NASDAQ.

Comparison with regularization-based approaches. We here empirically evaluate \( \| \hat{\Sigma} - \Sigma^* \|_1 \) relative to well-known regularization-based methods.

Setup. We consider rank-one Wishart measurement matrices \( X_i = z_i z_i^\top, z_i \sim \mathcal{N}(0, I), i = 1, \ldots, n \), fix \( m = 50 \) and let \( n \in \{ 0.24, 0.26, \ldots, 0.36, 0.4, \ldots, 0.56 \} \cdot m^2 \) and \( r \in \{ 1, 2, \ldots, 10 \} \) vary. Each configuration of \( (n, r) \) is run with 50 replications. In each of these, we generate data

\[
y_i = \text{tr}(X_i \Sigma^*) + \sigma \varepsilon_i, \quad \sigma = 0.1, \quad i = 1, \ldots, n,
\]

where \( \Sigma^* \) is generated randomly as rank \( r \) Wishart matrices and the \( \{ \varepsilon_i \}_{i=1}^n \) are i.i.d. \( \mathcal{N}(0, 1) \).
An estimate is available. A constrained least squares estimator of methods with careful tuning. For larger values of $r$ the constrained least squares estimator does not differ much from that of the regularization-based

**Discussion of the results.** We conclude from Figure 1 that in most cases, the performance of the constrained least squares estimator does not differ much from that of the regularization-based methods with careful tuning. For larger values of $r$, the constrained least squares estimator seems to require slightly more measurements to achieve competitive performance.

**Real Data Examples.** We now present an application to recovery of spiked covariance matrices which are of the form $\Sigma^* = \sum_{j=1}^{r} \lambda_j u_j u_j^\top + \sigma^2 I$, where $r \ll m$ and $\lambda_j \gg \sigma^2 > 0$, $j = 1, \ldots, r$. This model appears frequently in connection with principal components analysis (PCA).

**Extension to the spiked case.** So far, we have assumed that $\Sigma^*$ is of low rank, but it is straightforward to extend the proposed approach to the case in which $\Sigma^*$ is spiked as long as $\sigma^2$ is known or an estimate is available. A constrained least squares estimator of $\Sigma^*$ takes the form $\hat{\Sigma} + \sigma^2 I$, where

$$\hat{\Sigma} \in \arg\min_{\Sigma \in \mathcal{S}_+^n} \frac{1}{2n} \|y - \mathcal{A}(\Sigma + \sigma^2 I)\|^2_2. \tag{17}$$

The case of $\sigma^2$ unknown or general (unknown) diagonal perturbation is left for future research.
the problem by using measurements \((18)\). Even under significant perturbations \((\beta = 0.08)\), reasonable reconstruction of \(\Sigma^*\) remains possible, albeit the number of required measurements increases accordingly. In the extreme case \(\beta = 1/N\), the error is still decreasing with \(n\), but millions of samples seems to be required to achieve reasonable reconstruction error.

The general picture is similar for the NASDAQ data set, but the difference between using measurements based on the full sample correlation matrix on the one hand and approximations based on random subsampling \((18)\) on the other hand are more pronounced.

Data sets. (i) The CBCL facial image data set \([1]\) consist of \(N = 2429\) images of \(19 \times 19\) pixels (i.e., \(m = 361\)). We take \(\Sigma^*\) as the sample covariance matrix of this data set. It turns out that \(\Sigma^*\) can be well approximated by \(\Sigma_r\), \(r = 50\), where \(\Sigma_r\) is the best rank \(r\) approximation to \(\Sigma^*\) obtained from computing its eigendecomposition and setting to zero all but the top \(r\) eigenvalues. (ii) We construct a second data set from the daily end prices of \(m = 252\) stocks from the technology sector in NASDAQ, starting from the beginning of the year 2000 to the end of the year 2014 (in total \(N = 3773\) days, retrieved from \(\texttt{finance.yahoo.com}\)). We take \(\Sigma^*\) as the resulting sample correlation matrix and choose \(r = 100\).

Experimental setup. As in preceding measurements, we consider \(n\) random Wishart measurements for the operator \(X_i\), where \(n = C(mr)\), where \(C\) ranges from 0.25 to 12. Since \(\|\Sigma_r - \Sigma^*\|_F/\|\Sigma^*\|_F \approx 10^{-3}\) for both data sets, we work with \(\sigma^2 = 0\) in \((17)\) for simplicity. To make recovery of \(\Sigma^*\) more difficult, we make the problem noisy by using observations

\[
y_i = \text{tr}(X_i S_i), \quad i = 1, \ldots, n,
\]

(18)

where \(S_i\) is an approximation to \(\Sigma^*\) obtained from the sample covariance respectively sample correlation matrix of \(\beta N\) data points randomly sampled with replacement from the entire data set, \(i = 1, \ldots, n\), where \(\beta\) ranges from 0.4 to 1/\(N\) (\(S_i\) is computed from a single data point). For each choice of \(n\) and \(\beta\), the reported results are averages over 20 replications.

Results. For the CBCL data set, as shown in Figure 2, \(\hat{\Sigma}\) accurately approximates \(\Sigma^*\) once the number of measurements crosses \(2mr\). Performance degrades once additional noise is introduced to the problem by using measurements \((18)\). Even under significant perturbations \((\beta = 0.08)\), reasonable reconstruction of \(\Sigma^*\) remains possible, albeit the number of required measurements increases accordingly. In the extreme case \(\beta = 1/N\), the error is still decreasing with \(n\), but millions of samples seems to be required to achieve reasonable reconstruction error.

The general picture is similar for the NASDAQ data set, but the difference between using measurements based on the full sample correlation matrix on the one hand and approximations based on random subsampling \((18)\) on the other hand are more pronounced.

4 Conclusion

We have investigated trace regression in the situation that the underlying matrix is symmetric positive semidefinite. Under restrictions on the design, constrained least squares enjoys similar statistical properties as methods employing nuclear norm regularization. This may come as a surprise, as regularization is widely regarded as necessary in small sample settings.

Acknowledgments

The work of Martin Slawski and Ping Li is partially supported by NSF-DMS-1444124, NSF-III-1360971, ONR-N00014-13-1-0764, and AFOSR-FA9550-13-1-0137.
References