On Compressive Ensemble Induced Regularisation: 
How Close is the Finite Ensemble Precision Matrix to the Infinite Ensemble?

Ata Kabán 
A.KABAN@CS.BHAM.AC.UK

School of Computer Science, University of Birmingham, Edgbaston, B15 2TT, Birmingham, UK

Editors: Steve Hanneke and Lev Reyzin

Abstract
Averaging ensembles of randomly oriented low-dimensional projections of a singular covariance represent a novel and attractive means to obtain a well-conditioned inverse, which only needs access to random projections of the data. However, theoretical analyses so far have only been done at convergence, implying good properties for ‘large-enough’ ensembles. But how large is ‘large enough’? Here we bound the expected difference in spectral norm between the finite ensemble precision matrix and the infinite ensemble, and based on this we give an estimate of the required ensemble size to guarantee the approximation error of the finite ensemble is below a given tolerance. Under mild assumptions, we find that for any given tolerance, the ensemble only needs to grow linearly in the original data dimension. A technical ingredient of our analysis is to upper bound the spectral norm of a matrix-variate $T$, which we then employ in conjunction with specific results from random matrix theory regarding the estimation of the covariance of random matrices.

Keywords: Ensemble Learning, Compressive Learning, Random Matrix Theory, Matrix-variate $T$

1. Introduction

Obtaining an invertible approximation to a singular covariance matrix is a classical problem in many high dimensional estimation settings where the available sample size is too small relative to the feature dimension of the data. Examples include classification and clustering with multivariate Gaussians, least squares regression, and more generally Gaussian graphical models (Yuan & Lin, 2007; Meinshausen & Bühlmann, 2006).

Many methods have been proposed. The common idea is to restrict the number of parameters that model the covariance. Two major branches of methods include rotation-sensitive methods such as sparsity or structured sparsity restrictions (which posit that only few features correlate with each other), and rotation-invariant methods, such as the Ledoit-Wolf estimator (Ledoit & Wolf, 2004), ridge regularisation, and more recently proposed random projection ensembles (Marzetta et al., 2011).

This paper is concerned with the latter approach, which may be more appropriate when sparsity or other specific structural assumptions are not known and not justified a priori. For instance, gene and protein association networks often present complex and dense interactions between many genes or proteins at a stage of disease development (Krämer et al., 2009; Ideker & Sharan, 2012). To account for such situations, and also motivated by ad-
Vances in compressive data acquisition, novel regularisation schemes have been identified in the form of aggregating a large number of compressed estimates. Regularisation then happens as a byproduct of reducing dimension, without having to impose sparsity or other structure. In particular, Marzetta et al (2011) proposes such schemes as a general-purpose approach to handling singular covariances, and interestingly, the same type of approximator / regulariser shows up when learning an ensemble of compressive linear Fisher discriminant classifiers (Durrant & Kabán, 2015) or an ensemble of compressive OLS regressors (Thanei et al, 2017).

This novel way to regularise the covariance was demonstrated empirically to outperform the Ledoit-Wolf estimator in the sense of Frobenius norm of the difference between the approximator constructed from a severely singular covariance estimate and the true covariance (Marzetta et al, 2011), and also it significantly outperformed ridge regularisation in terms of classification performance (Durrant & Kabán, 2015) on high dimensional gene expression data. However, all theoretical analyses have been conducted at the convergence of the ensemble – that is, assuming an infinitely large ensemble. This paper is concerned with the question of how close is the finite ensemble estimate from the infinite ensemble estimate, and the related question of how large one needs to grow the ensemble to achieve a desired closeness in the spectral norm.

More precisely, given a fixed $d \times d$ positive semi-definite rank $\rho < d$ matrix $M$, consider the following inverse covariance approximator:

$$ic_k(M) = E[R^T (RM R^T)^{-1} R]$$

where $R$ is a random $k \times d$, matrix with i.i.d. standard Gaussian entries, and $k < \rho - 1$.

As shown by Marzetta et al (2011), $ic_k(M)$ is always non-singular, even if $M$ was singular. The spectral properties of $ic_k(M)$ have been analysed from various angles (Marzetta et al, 2011; Durrant & Kabán, 2015; Thanei et al, 2017).

However, in practice, one can only estimate the matrix expectation in eq.(1) by employing a Monte Carlo sample average. Let $R_1, R_2, ..., R_m$ be independent copies of $R$. Then,

$$\hat{ic}_k(M) = \frac{1}{m} \sum_{i=1}^{m} R_i^T (R_i M R_i^T)^{-1} R_i$$

The question we study here is how large $m$ needs to be so that

$$E[\|ic_k(M) - \hat{ic}_k(M)\|] \leq \epsilon?$$

where $\| \cdot \|$ denotes the spectral norm.

Problems of this kind are known to be challenging in general, and are the subject of study in non-asymptotic random matrix theory (Rudelson, 1999; Youssef, 2013; Ahlswede & Winter, 2002). In covariance estimation it is known from the work of Rudelson (1999) that, for general distributions with support on the sphere of radius $\sqrt{d}$ the required size is $m = O(d \log d)$, but for many distributions $m = O(d)$ is sufficient. A lot of progress has been made in the past few years on identifying distributions in the latter category (Adamczak et al, 2012). Relatively recently, work by Youssef (2013) extended such results to the matrix-covariance
setting, and gave some generic conditions under which \( m \in O(d) \). This order is the best we can hope for, especially since we will have to deal with sums of heavy tailed matrices. Some known examples in which \( m \in O(d) \) suffices are sums of subgaussian or subexponential matrices (Ahlswede & Winter, 2002).

As we shall see, since \( \rho < d \), the random matrix elements of the sum of our interest each contain a heavy-tailed sub-matrix with dependent entries. Controlling these is where most of the difficulty lies. The next section provides the key technical ingredients needed, namely we develop a high probability upper bound on the spectral norm of a matrix-variate \( T \), and we list the specific tools from random matrix theory that we will need. Hereafter we use the shorthand \( \bar{\rho} \equiv d - \rho \geq 1 \).

2. Technical ingredients

2.1. Upper bound on the spectral norm of a matrix-variate \( T \)

Let \( P \) and \( Q \) be two independent random matrices with i.i.d. standard normal entries, of size \( k \times \rho \), and \( k \times r \) respectively, and assume that \( k < \rho - 1 \). Noting that \( PP^T \sim W(\rho, I_k) \) is a Wishart matrix independent of \( Q \), by Theorem 4.2.1 in (Gupta & Nagar, 1999), the matrix \( J := (PP^T)^{-1/2}Q \) has a zero mean matrix-variate T-distribution, \( T_{k \times r}(0, I_k, I_r, \nu) \) with

\[
\nu = \rho - k + 1. \tag{4}
\]

Here, and throughout this paper, we refer to the parametrisation from (Gupta & Nagar, 1999), so the \( k \times r \) matrix \( J \) has the following probability density:

\[
p(J) = \frac{\Gamma_k \left( \frac{\nu+k+r-1}{2} \right)}{\pi^{k^2} \Gamma_k \left( \frac{\nu+k-1}{2} \right)} \det(I_k + JJ^T)^{-\frac{\nu+k+r-1}{2}} \tag{5}
\]

where \( \Gamma_p(a) \equiv \pi^{\frac{p-1}{2}} \prod_{i=1}^{p} \Gamma(a + \frac{1-i}{2}) \) is the multivariate Gamma function. A property of this matrix-distribution is that \( J^T \sim T_{r \times k}(0, I_r, I_k, \nu) \), by Theorem 4.3.3 in (Gupta & Nagar, 1999).

The goal of this section is to prove a polynomially decaying upper bound on the following:

\[
\Pr \left\{ \lambda_{\text{max}} \left( Q^T(PP^T)^{-1}Q \right) \cdot \frac{\rho - k - 1}{k} \geq t \right\} \leq ? \tag{6}
\]

where \( \lambda_{\text{max}} \) denotes largest eigenvalue of its argument.

We should note that the matrix-variate T is different from a multivariate t vector reshaped into a matrix (Daz-García & Gutiérrez-Jáimez, 2012). Instead, the matrix-variate T-distribution implies that both the rows and the columns of \( J \) are statistically dependent on each other, so existing bounds on the spectral norm of random matrices are not readily available.
We have:

\[
\Pr\left\{ \lambda_{\max}(Q^T (PP^T)^{-1}Q) \cdot \frac{n - k - 1}{k} \geq t \right\} = \Pr\left\{ \lambda_{\max}(J^T J) \frac{\nu - 2}{k} \geq tk \right\} \\
= \Pr\left\{ \lambda_{\max}(J^T J)(\nu - 2) \geq tk \right\} \\
\leq \Pr\left\{ \text{Tr}(J^T J)(\nu - 2) \geq tk \right\} \\
= \Pr\left\{ \sum_{j=1}^{k} \|J_j\|^2 \cdot (\nu - 2) \geq tk \right\}
\]

(7)

where \(J_j\) denotes the \(j\)-th row of \(J\).

By Theorem 4.3.9 in (Gupta & Nagar, 1999), all marginal distributions of the rows (and columns) of \(J\) are distributed as multivariate t with the same degree of freedom \(\nu\). In particular, \(J_j\) above is distributed as a multivariate \(t\) with \(\nu\) degrees of freedom – in the parametrisation we are using, the pdf of this random vector is given by plugging \(k = 1\) into eq. (5). It is then easy to check that \(J_j \sqrt{\nu - 2}\) is isotropic – indeed, its variance matrix exists since \(k < \rho - 1\) (hence \(\nu \geq 2\)) and it evaluates to \(I_r\).

We first give the following lemma, which may be of independent interest. The proof follows in Section 5.1.

**Lemma 1 (Chernoff-type bound on square norm of \(t\) distributed random vectors)**

Let \(x \sim T_d(0, I_d, \nu)\). Then \(\forall t > d,\)

\[
\Pr\{\|x\|^2 > t\} \leq \left(\frac{d}{t}\right)^{-\frac{d}{2}} \left(\frac{d + \nu}{t + \nu}\right)^{\frac{\nu + d}{2}}
\]

(8)

**Remark 2** Lemma 1 is tight in the sense that in the limit when \(\nu \to \infty\) it recovers a Chernoff bound for the square norm of Gaussian random vectors:

\[
\lim_{\nu \to \infty} \left(\frac{d}{t}\right)^{-\frac{d}{2}} \left(\frac{d + \nu}{t + \nu}\right)^{\frac{\nu + d}{2}} = \left(\frac{d}{t}\right)^{-\frac{d}{2}} \exp\left(\frac{-t - d}{2}\right) \geq \Pr\{\|y\|^2 > t\}
\]

(9)

**Remark 3** For finite \(\nu\), observe that the r.h.s. in Lemma 1 tightens with increasing \(\nu\), which suggests better concentration for higher degrees of freedom. This agrees with the intuition that a heavier tail (smaller \(\nu\)) implies more mass spread-out in the tails, hence concentration becomes weaker. More formally, it is straightforward to see that we can bound:

\[
\left(\frac{d}{t}\right)^{-\frac{d}{2}} \left(\frac{d + \nu}{t + \nu}\right)^{\frac{\nu + d}{2}} \leq \left(\frac{t}{d}\right)^{\frac{d}{2}} \exp\left(-\frac{t - d}{2} \cdot \frac{d + \nu}{t + \nu}\right)
\]

(10)

where the fraction \(\frac{d + \nu}{t + \nu} \leq 1\) for all \(\nu\) since \(t > d\), and reaches 1 as \(\nu \to \infty\).

Since \(k\) is finite, using Lemma 1, and noting that \(J_j \sqrt{\nu - 2} \sim T_r(0, I_r, \nu)\), we can further bound the right hand side (RHS) of eq. (7) for all \(t > c \cdot r\), where \(c > 1\) is a constant, as
the following:

\[
\text{eq. (7)} \leq \sum_{j=1}^{k} \Pr \{ \| J_j \cdot \sqrt{\nu - 2} \|^2 \geq t \} \leq k \cdot \left( \frac{t}{r} \right)^{\frac{\nu + r}{2}} \cdot \left( \frac{r + \nu}{t + \nu} \right)^{\frac{\nu + r}{2}} \tag{11}
\]

2.2. Specific tools from random matrix theory

In addition to the tail bound developed in the previous section we will need the following tools:

Definition 4 (Matrix Strong Regularity (MSR) condition (Youssef, 2013)) A positive semidefinite random matrix \( U \) of dimension \( d \times d \) and \( E[U] = I_d \) satisfies MSR if there exist \( \eta, c_{MSR} > 0 \) constants s.t.

\[
\Pr \{ \| AU A \| \geq t \} \leq c_{MSR} \frac{t^{1+\eta}}{\nu + \eta}, \forall t \geq c_{MSR} \leq \text{rank}(A), \forall A \text{ orthogonal projection in } \mathbb{R}^d
\]

where \( \| \cdot \| \) denotes the operator norm.

Theorem 5 (Youssef (2013)) Let \( U \) be a \( d \times d \) positive semidefinite matrix having \( E[U] = I_d \) and satisfying the MSR for some \( \eta, c_{MSR} > 0 \), and let \( U_1, U_2, ..., U_m \) be independent copies of \( U \). Then, \( \forall \epsilon \in (0,1) \), for \( m = C_1 \cdot \frac{d}{\epsilon^{2+2/\eta}} \), we have:

\[
E[\| \frac{1}{m} \sum_{i=1}^{m} U_i - I_d \|] \leq \epsilon \tag{13}
\]

where \( C_1 \) is a constant that depends only on \( \eta \) and \( c_{MSR} \).

The exact expression of \( C_1 \) is given in (Youssef, 2013).

3. Estimating the ensemble size

We define the generic term \( U_{(M)} \) of the matrix sum in \( \hat{c}_k \) of eq. (2) as the following, where an appropriate transformation is included in order to satisfy \( E[U_{(M)}] = I_d \):

\[
U_{(M)} = E[R^T (RMRT)^{-1} R]^{-1/2} \cdot R^T (RMRT)^{-1} R \cdot E[R^T (RMRT)^{-1} R]^{-1/2} \tag{14}
\]

By writing \( M = \Lambda \Lambda^T \) for the SVD decomposition of \( M \), where \( \Lambda \) is the \( d \times d \) diagonal matrix of the eigenvalues of \( M \), and \( LL^T = L^T L = I_d \), observe that \( U_{(M)} \) has the same distribution as

\[
U_{(\Lambda)} \sim L \cdot U_{(\Lambda)} \cdot L^T \tag{15}
\]

since \( R \) has the same distribution as \( RL \).

Therefore it is enough to obtain a result of the form of eq. (13) for terms \( U_i \) of the form \( U_{(\Lambda)} \), in other words, we can identify \( M \) with \( \Lambda \). Indeed,

\[
E[\| \frac{1}{m} \sum_{i=1}^{m} U_{(M)} - I_d \|] = E[\| L (\frac{1}{m} \sum_{i=1}^{m} U_{i(\Lambda)} - I_d) L^T \|] = E[\| \frac{1}{m} \sum_{i=1}^{m} U_{i(\Lambda)} - I_d \|] \tag{16}
\]

and \( U_{\Lambda} \) satisfies \( E[U_{(\Lambda)}] = I_d \).
3.1. Roadmap

First we establish Lemmas 6 and 7, which split the problem into separately proving MSR for the two block-diagonal sub-matrices. Of these, the sub-matrix that corresponds to the null-space of $M$ turns out to have a matrix-variate T distribution, and we make use of our Lemma 1 to control its spectral norm. Dealing with the sub-matrix that corresponds to the range-space of $M$ is rather straightforward, and finally we leverage Theorem 5 to conclude that, provided $\rho - k + 1 \geq \Omega(\log(d - \rho))$, the required ensemble size is linear in $d$. In Section 4 we demonstrate numerical simulations that confirm order-wise tightness. Section 5 is devoted to proofs, and Section 6 concludes the paper.

3.2. Splitting up the problem

To get started with proving MSR for $U(\Lambda)$, we take a $d \times d$ projection matrix of rank $r \in \{1, \ldots, d\}$. This necessarily has the form $A = B^T (BB^T)^{-1} B$ where $B$ is an $r \times d$ full row-rank matrix. We need to develop an upper bound on $\Pr \{ \|AU(\Lambda)A\| \geq t \}$ for all $t > c_{MSR} \cdot r$.

Denote by $\Delta$ the $\rho \times \rho$ diagonal matrix of the non-zero eigenvalues of $M$, and it will be useful to split $R$ as $R = \begin{bmatrix} P & S \end{bmatrix}$ into the $k \times \rho$ matrix $P$ and the $k \times \bar{\rho}$ matrix $S$, where $\bar{\rho} = d - \rho$.

Using the fact that $E[R^T (RAR^T)^{-1} R]$ is diagonal (Marzetta et al, 2011; Durrant & Kabán, 2015), we can write:

$$U(\Lambda) = \begin{bmatrix} V(\Delta) & Z(\Delta) \\ Z(\Delta) & W(\Delta) \end{bmatrix}$$

(17)

where

$$V(\Delta) = E[P^T (P \Delta P^T)^{-1} P]^{-1/2} \cdot P^T (P \Delta P^T)^{-1} P \cdot E[P^T (P \Delta P^T)^{-1} P]^{-1/2}$$

(18)

$$W(\Delta) = E[S^T (P \Delta P^T)^{-1} S]^{-1/2} \cdot S^T (P \Delta P^T)^{-1} S \cdot E[S^T (P \Delta P^T)^{-1} S]^{-1/2}$$

(19)

$$Z(\Delta) = E[P^T (P \Delta P^T)^{-1} P]^{-1/2} \cdot P^T (P \Delta P^T)^{-1} S \cdot E[S^T (P \Delta P^T)^{-1} S]^{-1/2}$$

(20)

Then we can split the problem by showing the following:

**Lemma 6** Denote by $\Delta$ the $\rho \times \rho$ diagonal matrix of the non-zero eigenvalues of $M$. For $U$ defined as above, we have:

$$\|AU(\Lambda)A\| \leq \|A_1 V(\Delta) A_1\| + \|A_2 W(\Delta) A_2\|$$

(21)

where the $\rho \times \rho$ matrix $A_1$ and the $\bar{\rho} \times \bar{\rho}$ matrix $A_2$ are projections of rank $r$ in $\mathbb{R}^\rho$ and $\mathbb{R}^{\bar{\rho}}$ respectively.

The proof is deferred to Section 5.2, along with the construction of $A_1$ and $A_2$.

To further simplify the problem, denote by $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ the largest and the smallest eigenvalues respectively, and $\lambda_{\min \neq 0}(\cdot)$ will denote the smallest non-zero eigenvalue of its matrix argument. We shall assume that $\kappa(\Delta) \equiv \lambda_{\max}(M)/\lambda_{\min \neq 0}(M) = \lambda_{\max}(\Delta)/\lambda_{\min}(\Delta)$ is bounded by some constant independent of $d$. This is reasonable since we are talking about the condition number of the non-random $\rho \times \rho$ matrix where $\rho < d$. 
With this assumption we can show that it is enough to prove MSR for $M := M_0 \equiv \begin{bmatrix} I_\rho & 0 \\ 0 & 0 \end{bmatrix}$. That is, it is sufficient to consider the case $\Lambda = I_\rho$. Generalisation to other $M$ then presents no difficulty using the following Lemma 7.

**Lemma 7** Assume that $\kappa(\Lambda)$ is bounded independently of $d$. Then,

$$\|A_1 V(\Lambda) A_1\| \leq \kappa(\Lambda) \cdot \|A_1 V(I_\rho) A_1\|$$

(22)

$$\|A_2 W(\Lambda) A_2\| \leq \kappa(\Lambda) \cdot \|A_2 W(I_\rho) A_2\|$$

(23)

where $I_\rho$ is the $\rho$-dimensional identity matrix.

The proof is deferred to Section 5.3.

The advantage is that the expectation that appears in $ic_k(M_0)$ has a closed form in this case:

$$ic_k(M_0) = \begin{bmatrix} k \rho & 0 \\ 0 & k \rho \bar{\rho} \end{bmatrix}$$

(24)

This can be seen easily by computing:

$$E[V(I_\rho)] = \frac{k}{\rho} I_\rho$$

$$E[W(I_\rho)] = \frac{k}{\rho - k - 1} I_\rho$$

$$E[Z(I_\rho)] = 0$$

Lemma 7 implies that $M$ can only change the constant $c_{MSR}$. Indeed, if $V(I_\rho)$ satisfies the MSR with $\eta, c_{MSR} > 0$, then $V(\Lambda)$ satisfies the MSR with $\eta$ and $c_{MSR} \cdot (\kappa(\Lambda))^{1+\eta}$. Likewise, if $W(I_\rho)$ satisfies the MSR with $\eta', c_{MSR} > 0$, then $W(\Lambda)$ satisfies the MSR with $\eta'$ and $c_{MSR} \cdot (\kappa(\Lambda))^{1+\eta'}$.

With the choice $M := M_0$, we will omit the index $(I_\rho)$ from the notations of $U, V, W, Z$, and the form of $U$ then becomes the following:

$$U = \begin{bmatrix} P \left(P^T P^T\right)^{-1} P \frac{\rho}{k} & P \left(P^T P^T\right)^{-1} S \sqrt{\frac{\rho(\rho-k-1)}{k}} \\ S^T \left(P^T P^T\right)^{-1} P \frac{\sqrt{\rho(\rho-k-1)}}{k} & S^T \left(P^T P^T\right)^{-1} S \frac{\rho-k-1}{k} \end{bmatrix}$$

$$= \begin{bmatrix} V & Z \\ Z^T W \end{bmatrix}$$

It is worth observing that the two diagonal blocks belong to different classes of matrix-valued distributions. The block $V$ has all its non-zero eigenvalues equal to $\rho/k$, whereas the block $W$ has a heavy tailed matrix-variate distribution. The matrix norm of our interest is dominated by the latter.

In the original problem, $r$ takes values in $\{1, 2, ..., d\}$, so it may be also worth noting that in the resulting two terms it is sufficient to consider $r \in \{1, ..., \rho\}$ and $r \in \{1, ..., \bar{\rho}\}$ respectively, since for $r > \rho$ we have $\|A_1 V A_1\| = \|V\|$, and likewise for $r > \bar{\rho}$ we have $\|A_2 W A_2\| = \|W\|$. 

7
3.3. Main result

We are now ready to prove the following:

**Theorem 8** Assume $3 \leq \rho < d$, $k$ finite, $k < \rho - 1$, and there exists $c, \eta > 0$ constants such that $k \left( \bar{\rho} \right)^{\frac{1}{2}} \left( \frac{\bar{\rho} + \nu}{\bar{\rho} + \nu + t^{1+\eta}} \right) \leq c, \forall \bar{\rho} \geq c \cdot \bar{\rho}$. Then $R^T (RM_0 R^T)^{-1} R \cdot \begin{bmatrix} k/\rho \cdot I_\rho & 0 \\ 0 & k/(\rho - k - 1) \cdot I_{\bar{\rho}} \end{bmatrix}$ satisfies the MSR with $(c, \eta)$.

The proof is given in Section 5.4.

**Remark 9** It can be shown\(^1\) that the constant $c$ required in Theorem 8 exists for any choice of $\eta$, provided that $\nu > a \log(\bar{\rho}) + a$ for some constant $a$.

Therefore, recalling that $\nu = \rho - k + 1$, and $\bar{\rho} = d - \rho$, a sufficient condition for MSR in our case is that $\rho - k + 1 \geq \Omega(\log(d - \rho))$.

In practice, if $M$ was a singular covariance estimate, then $\rho$ is always no larger than the sample size. In fact, $\rho$ can be small for a number of reasons, e.g. because of small sample size, or multiple collinearities in the data, or because the support of the relevant features is in a small dimensional subspace. If $\rho$ is too small relative to $d$ then we cannot guarantee the MSR condition. However, we see that it is enough for the range space dimension $\rho$ to be at least logarithmic in the null-space dimension $(d - \rho)$ for the MSR condition to hold.

In particular, a setting with exponentially many irrelevant features relative to the sample size (Ng, 2004; Kabán & Durrant, 2008) meets the condition.

Our main result, for the size of the compressive ensemble is an immediate corollary of Theorem 8, is the following.

**Corollary 10 (Required ensemble size)** Under the same conditions as in Theorem 8, $\forall \epsilon \in (0, 1)$, we have:

$$E \left[ \frac{1}{m} \sum_{i=1}^{m} R_i^T (R_i M_0 R_i^T)^{-1} R_i - E \left[ R^T (RM_0 R^T)^{-1} R \right] \right] \leq \epsilon \cdot E \left[ R^T (RM_0 R^T)^{-1} R \right]$$

$$\leq \epsilon \cdot \frac{k}{\rho - k - 1}$$

provided that the ensemble size is:

$$m \geq C_1(c, \eta) \cdot \frac{d}{\epsilon^{2+2/\eta}} \tag{25}$$

where $C_1(c, \eta)$ is an absolute constants independent of $d$.

---

4. Numerical demonstration

We give an empirical demonstration of our finding, that under the mild conditions the ensemble size \( m \) only needs to grow linearly with \( d \).

We took the fixed singular matrix \( M = M_0 \) with rank \( \rho = 50 \), generated random Gaussian matrices \( R_1, R_2, ..., R_m \) and computed:

\[
\epsilon := \frac{\mathbb{E} \left[ \left\| \frac{1}{m} \sum_{i=1}^{m} R_i^T (R_i M R_i^T)^{-1} R_i - \mathbb{E} [R^T (R M R)^{-1} R] \right\| \right]}{\left\| \mathbb{E} [R^T (R M R)^{-1} R] \right\|}
\]

The expectations in the above have analytic forms, which we used. We increased the ensemble size \( m \) progressively until \( \epsilon \) reached below a threshold, for two different choices of the threshold: \( \epsilon = 0.5 \) and \( \epsilon = 0.3 \). We varied \( k \) and \( d \), and ran 15 independent repetitions of this experiment each time. The ensemble sizes \( (m) \) on the vertical axis of Figure 1 are averages computed from the 15 independent repetitions.

From Figure 1 it is most apparent that for each \( k \) (equivalently \( \nu \)) the growth of the required ensemble size \( (m) \) follows a linear trend as a function of \( d \), as expected from Corollary 10. The best linear fits are superimposed on these figures.

We then repeated the experiment, this time generating \( M \) randomly and then fixing it. Again the rank was \( \rho = 50 \), and the condition number in the range space of \( M \) was equal to \( \kappa(M) = 1.5 \) each time. Figure 2 shows the ensemble sizes required for \( \epsilon \) to reach below 0.5. As expected from our Lemma 7 combined with Corollary 10, the required ensemble size again grows linearly with \( d \). These simulations suggest that our concentration bound is tight order-wise.

We also find it interesting to notice on these figures that the slope decreases with \( \nu = \rho - k - 1 \) through \( \eta \), but it increases with the same through \( c_{MSR} \). There is clearly a tradeoff in choosing \( \nu \) – which is indeed the users’ choice via setting \( k \) – in agreement with the previous empirical observation in (Durrant & Kabán, 2015), namely that setting \( k \) around the middle of its allowed range works well (and hence the assumption we made that \( \rho/k \) is bounded is quite natural).

5. Proofs

5.1. Proof of Lemma 1

We use the following representation (see e.g. Lin (1972)): If \( y \sim \mathcal{N}_\rho(0, \Sigma) \) and \( s^2 \sim \chi^2\_\nu \) independent of \( y \), then \( \frac{y \sqrt{\Sigma}}{s} \sim T_\nu(0, \Sigma, \nu) \).

Take \( y \sim \mathcal{N}(0, I_d) \) a standard Gaussian vector and \( u \sim \mathcal{N}(0, I_\nu) \) – so \( \|u\|^2 \sim \chi^2_\nu \). We have:

\[
\Pr \{ \|x\|^2 > t \} = \Pr \left\{ \frac{\|y\|^2}{\|u\|^2} > \frac{t}{\nu} \right\} = \Pr \left\{ \|y\|^2 > \frac{t}{\nu} \|u\|^2 \right\} = \Pr \left\{ \exp \left( \lambda \|y\|^2 \right) \exp \left( -\frac{t}{\nu} \lambda \|u\|^2 \right) > 1 \right\}, \forall \lambda > 0
\]
Figure 1: Numerical experiment demonstrating that the required ensemble size grows linearly in $d$, for $\epsilon = 0.5$, and $\epsilon = 0.3$. We used $M = M_0$ with $\rho = 50$.

Figure 2: Similar experiment, now using a generic singular matrix $M$ of rank $\rho = 50$ (generated randomly and then fixed) with condition number $\kappa(M) = 1.5$ in its range space. As suggested by Lemma 7, the required ensemble size remains linear in $d$. 
\[
\begin{align*}
\leq & \ E \left[ \exp(\lambda \|y\|^2) \right] \ E \left[ \exp \left( -\frac{t}{\nu} \lambda \|u\|^2 \right) \right], \forall \lambda > 0 \\
= & \ (1 - 2\lambda)^{-\frac{d}{2}} \left( 1 + 2\lambda \frac{t}{\nu} \right)^{-\frac{\nu}{2}} \forall \lambda \in (0, 1/2)
\end{align*}
\]

We optimise the bound in \( \lambda \) by solving the stationary equation:
\[
\frac{\partial}{\partial \lambda} = \frac{2\lambda t(d/\nu + 1) - t + d}{(1 - 2\lambda)^{d/2}(1 + 2\lambda t/\nu)^{\nu/2+1}} = 0
\]
which gives
\[
\lambda = \frac{t - d}{2t(d/\nu + 1)}
\]
This is indeed in the interval \((0, 1)\) as required, since \( t > d \). After plugging back, the RHS of eq. (30) becomes:
\[
\begin{align*}
&\left( 1 - \frac{t - d}{t(d/\nu + 1)} \right)^{-d/2} \left( 1 + \frac{t - d}{\nu(d/\nu + 1)} \right)^{-\nu/2} \\
= & \left( \frac{d(t + \nu)}{t(d + \nu)} \right)^{-d/2} \left( \frac{t + \nu}{d + \nu} \right)^{-\nu/2} \\
= & \left( \frac{d}{t} \right)^{-d/2} \left( \frac{d + \nu}{t + \nu} \right)^{\nu+d/2}
\end{align*}
\]

5.2. Proof of Lemma 6
Recall that \( A \) is a projection matrix. Here, \( U \) will be shorthand for \( U_{(\Lambda)} \). Let us rewrite the matrix norm of interest as the following:
\[
\|AU A\| = \|B^T(BB^T)^{-1}BU B^T(BB^T)^{-1}B\| \\
= \|(BB^T)^{-1/2}BU B^T(BB^T)^{-1/2}\| \\
\equiv \|\mathfrak{B}U\mathfrak{B}^T\|
\]
where we introduced the notation \( \mathfrak{B} \equiv (BB^T)^{-1/2}B \).
Now, decompose the \( r \times d \) matrix \( B \) as a sum of two matrices of which the first matrix contains the first \( \rho \) columns of \( B \) and zeros in its last \( \bar{\rho} \) columns, and the second matrix has zeros in its first \( \rho \) columns followed by the remaining \( \bar{\rho} \) columns of \( B \).
\[
B = \begin{bmatrix} B_1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & B_2 \end{bmatrix}
\]
\[
= (B_1B_1^T)^{1/2} \begin{bmatrix} \mathfrak{B}_1 & 0 \end{bmatrix} + (B_2B_2^T)^{1/2} \begin{bmatrix} 0 & \mathfrak{B}_2 \end{bmatrix}
\]
where we used the notations \( \mathfrak{B}_i \equiv (B_iB_i^T)^{-1/2}B_i, i \in \{1, 2\} \). That is, we ortho-normalized the nonzero sub-matrices. Using these, we now construct a new positive semi-definite matrix that helps us reduce the problem to the diagonal blocks, namely the following \( 2r \times 2r \) matrix:
\[
\tilde{U} \equiv \begin{bmatrix} \mathfrak{B}_1 & 0 \\ 0 & \mathfrak{B}_2 \end{bmatrix} \cdot \begin{bmatrix} V_{(\Lambda)}^T \\ Z_{(\Lambda)}^T \end{bmatrix} \cdot \begin{bmatrix} \mathfrak{B}_1^T & 0 \\ 0 & \mathfrak{B}_2^T \end{bmatrix}
\]
This is clearly positive semidefinite since the \( d \times d \) matrix in the middle was assumed to be positive semidefinite.

By definition, the matrix norm of interest, eq. (34) is:

\[
\|AU\| = \max_{x \in \mathbb{R}^r} \frac{x^T (BB^T)^{-1/2} \left[ (B_1B_1^T)^{1/2} \right]^T \left[ (B_1B_1^T)^{1/2} \right] (BB^T)^{-1/2} x}{x^T x}
\]

(37)

and by observing that

\[
(BB^T)^{-1/2} \left[ (B_1B_1^T)^{1/2} \right]^T \left[ (B_1B_1^T)^{1/2} \right] (BB^T)^{-1/2} = (BB^T)^{-1/2} (B_1B_1^T + B_2B_2^T) (BB^T)^{-1/2}
\]

\[
= (BB^T)^{-1/2} BB^T (BB^T)^{-1/2} = I_r
\]

and denoting \( y = \left[ (B_1B_1^T)^{1/2} \right] (BB^T)^{-1/2} x \), we have:

\[
eq \text{eq.}(37) \leq \max_{y \in \mathbb{R}^r} \frac{y^T \tilde{U} y}{y^T y} = \|\tilde{U}\|
\]

(38)

This inequality holds because \( y \) takes values in a larger space than \( x \).

We are now ready to split up the original matrix norm of interest. By definition, and by construction, we have:

\[
eq \text{eq.}(38) = \|\tilde{U}\|
\]

\[
= \lambda_{\max} \left[ \begin{array}{ccc}
\mathcal{B}_1 V_{(\Delta)}^T & \mathcal{B}_1 Z_{(\Delta)} & \mathcal{B}_2 W_{(\Delta)}^T \\
\mathcal{B}_2 Z_{(\Delta)} & \mathcal{B}_2 W_{(\Delta)} & \mathcal{B}_2 W_{(\Delta)}^T
\end{array} \right]
\]

\[
\leq \|\mathcal{B}_1 V_{(\Delta)}^T\| + \|\mathcal{B}_2 W_{(\Delta)}\| + \|\mathcal{B}_2 W_{(\Delta)}^T\|
\]

\[
= \|B_1B_1^T\|^{-1/2} B_1V_{(\Delta)}^T B_1^T (B_1B_1^T)^{-1/2}
\]

\[
+ \|B_2B_2^T\|^{-1/2} B_2W_{(\Delta)}^T (B_2B_2^T)^{-1/2}
\]

\[
\leq \|A_1V_{(\Delta)}A_1\| + \|A_2W_{(\Delta)}A_2\|
\]

(39)

where \( A_i \equiv B_i(B_iB_i^T)^{-1/2} B_i, i \in \{1, 2\} \) are projections in \( \mathbb{R}^{\rho} \) and \( \mathbb{R}^\rho \) respectively. The inequality (39) is easy to check\(^2\).

5.3. Proof of Lemma 7

Recall that \( \Lambda \) is the \( \rho \times \rho \) diagonal matrix of non-zero eigenvalues of \( M \). We have \( \Lambda = \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} \).

\(^2\) joriki (http://math.stackexchange.com/users/6622/joriki), Largest eigenvalue of a positive semi-definite matrix is less than or equal to sum of largest eigenvalues of its diagonal blocks, URL (version: 2012-05-14): http://math.stackexchange.com/q/144963
Now, using the Rayleigh quotient inequality combined with Poincaré inequality (Theorem 4.2.2., Corr. 4.3.16 in (Horn & Johnson, 1985)) we have the following bounds: \( \forall x \in \mathbb{R}^p, \)

\[
x^T P^T (P \Delta P^T)^{-1} P x = x^T P^T (P P^T)^{-1/2} [(P P^T)^{-1/2} P \Delta \cdot P^T (P P^T)^{-1/2}]^{-1} (P P^T)^{-1/2} P x
\]

\[
\leq \frac{x^T P^T (P P^T)^{-1} P x}{\lambda_{\text{max}}(\Delta)}
\]

Likewise,

\[
x^T P^T (P \Delta P^T)^{-1} P x \geq \frac{x^T P^T (P P^T)^{-1} P x}{\lambda_{\text{min}}(\Delta)}
\]

By similar arguments, \( \forall x \in \mathbb{R}^p, \)

\[
\frac{x^T P^T (P P^T)^{-1} P x}{\lambda_{\text{min}}(\Delta)} \leq \frac{x^T S^T (P P^T)^{-1} S x}{\lambda_{\text{min}}(\Delta)} \leq \frac{x^T S^T (P P^T)^{-1} S x}{\lambda_{\text{max}}(\Delta)}
\]

In consequence, for the upper diagonal block we have:

\[
\| A_1 \mathbb{E}[P^T (P \Delta P^T)^{-1} P]^{-1/2} \cdot P^T (P \Delta P^T)^{-1} P \cdot \mathbb{E}[P^T (P \Delta P^T)^{-1} P]^{-1/2} A_1 \| \\
\leq \frac{1}{\lambda_{\text{min}}(\Delta)} \| \mathbb{E}[P^T (P \Delta P^T)^{-1} P]^{-1} \| \cdot \| A_1 P^T (P P^T)^{-1} P A_1 \|
\]

\[
\leq \frac{\lambda_{\text{max}}(\Delta)}{\lambda_{\text{min}}(\Delta)} \| \mathbb{E}[P^T (P P^T)^{-1} P]^{-1} \| \cdot \| A_1 P^T (P P^T)^{-1} P A_1 \|
\]

\[
= \kappa(\Delta) \frac{\rho}{k} \| A_1 P^T (P P^T)^{-1} P A_1 \| = \kappa(\Delta) \frac{\rho}{k} \| A_1 V A_1 \|
\]

Likewise, for the lower diagonal block we have:

\[
\| A_2 \mathbb{E}[S^T (P \Delta P^T)^{-1} S]^{-1/2} \cdot S^T (P \Delta P^T)^{-1} S \cdot \mathbb{E}[S^T (P \Delta P^T)^{-1} S]^{-1/2} A_2 \|
\]

\[
\leq \frac{\lambda_{\text{max}}(\Delta)}{\lambda_{\text{min}}(\Delta)} \frac{\rho - k - 1}{k} \| A_2 S^T (P P^T)^{-1} S A_2 \|
\]

\[
= \kappa(\Delta) \frac{\rho - k - 1}{k} \| A_2 W A_2 \|
\]

which concludes the proof. \( \Box \)

5.4. Proof of Theorem 8

Take any projection matrix in \( \mathbb{R}^d, \) and \( t \geq c \cdot r \) where \( r \) is the rank of \( A. \) By Lemma 6,

\[
\Pr \{ \| A U A \| \geq t \} \leq \Pr \{ \| A_1 V A_1 \| + \| A_2 W A_2 \| \geq t \} \leq \Pr \{ \| A_1 V A_1 \| \geq t/2 \} + \Pr \{ \| A_2 W A_2 \| \geq t/2 \}
\]

where \( A_1 \) and \( A_2 \) are rank \( r \) projection matrices in \( \mathbb{R}^p \) and \( \mathbb{R}^p \) respectively.

The matrix \( V \equiv P^T (P P^T)^{-1} P \cdot \frac{\rho}{k} \) satisfies MSR trivially, with any choice of \( \eta > 0, \) e.g. by Proposition 8.5 in (Youssef, 2013). It remains to ensure that MSR holds for \( W. \) Recall
that $W \equiv S^T (P P^T)^{-1} S \cdot \frac{\rho - k - 1}{k}$, and we have $E[W] = I_\rho$. $A_2$ is a projection matrix in $\mathbb{R}^\rho$, of rank $r \in \{1, \ldots, \rho\}$. So $A_2$ has the form $A_2 = B_2^T (B_2 B_2^T)^{-1} B_2$ for some full row-rank matrix $B_2$ of size $r \times \rho$.

Observe that we can absorb $A_2$ into $S$, since:

$$
\lambda_{\text{max}}(A_2 S^T (P P^T)^{-1} S A_2) = \lambda_{\text{max}}((B_2 B_2^T)^{-1/2} B_2 S^T (P P^T)^{-1} S B_2 (B_2 B_2^T)^{-1/2})
$$

where $Q := S B_2^T (B_2 B_2^T)^{-1/2}$, and since $S$ has i.i.d. standard Gaussian entries, the $k \times r$ matrix $Q$ also has i.i.d. standard Gaussian entries. Hence, $J := (P P^T)^{-1/2} Q$ has a matrix-variate $T$-distribution. In particular, $J \cdot \sqrt{(\nu - 2)/k} \sim T_{k \times \rho}(0, I_k, I_\rho, \nu)$ with $\nu = \rho - k + 1$. Hence, as a consequence of Lemma 1, i.e. eq. (12) we have for all $t > r$:

$$
\Pr \{|A_2 W A_2\| \geq t\} \leq k \cdot \left(\frac{t}{r}\right)^{\frac{r}{\nu}} \cdot \left(\frac{r + \nu}{t + \nu}\right)^{\frac{\nu + r}{2}} \leq k \cdot \left(\frac{t}{\rho}\right)^{\frac{r}{\nu}} \cdot \left(\frac{\rho + \nu}{t + \nu}\right)^{\frac{\nu + 2}{2}}
$$

Finally, we need to require that the r.h.s. of eq. (43) is upper bounded by $c \cdot t^{-1-\eta}$ for some $c > 0$ constant, and this is ensured by the condition in the statement of the Theorem.

6. Conclusions and future work

We quantified the Monte Carlo error of a compressive ensemble based covariance regularisation scheme, and determined how many independent copies of the random matrix are needed for the sample average of random matrices to get sufficiently close to the matrix expectation. Under mild assumptions we found that the ensemble size only needs to grow linearly with the dimension of the positive semi-definite input matrix. Order-wise this is tight. The main technical issue was to isolate and control the spectral norm of a heavy-tailed sub-matrix induced by the singularity of the input covariance matrix.

An interesting avenue for future work would be to extend the approach presented here beyond the use of Gaussian random projections in the ensemble members. In particular, we can show that the multivariate $t$-distribution with $\nu$ degrees of freedom belongs to the family of $\frac{-1}{\nu}$-concave distributions (see Borell, 1975; Vempala, 2009) for definitions), which suggests that such extension may be feasible as long as the lower diagonal block belongs in this family.

Acknowledgements

Thanks are due to Bob Durrant and Olivier Guédon for an enlightening discussion at Institut Henri Poincaré. Thanks to Bob also for proof-reading and thoughtful comments that improved the paper. Thanks also to Yaakov Baruch for promptly providing a proof for Remark 9 on MathOverflow. This work is funded by the EPSRC Fellowship EP/P004245/1.

References


