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 apriori. 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

advances 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 consructed 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:

$$\mathrm{ic}_k(M) = E[R^T(RMR^T)^{-1}R] \tag{1}$$

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), $\mathrm{ic}_k(M)$ is always non-singular, even if M was singular. The spectral properties of $\mathrm{ic}_k(M)$ 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,

$$\hat{\mathrm{ic}}_k(M) = \frac{1}{m}\sum_{i=1}^{m} R_i^T(R_iMR_i^T)^{-1}R_i \tag{2}$$

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

$$\mathrm{E}_R[\|\mathrm{ic}_k(M) - \hat{\mathrm{ic}}_k(M)\|] \leq \epsilon? \tag{3}$$

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 $m = \mathcal{O}(d \log d)$, but for many distributions $m = \mathcal{O}(d)$ 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 $m \in \mathcal{O}(d)$. 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 $m \in \mathcal{O}(d)$ 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 $PP^T \sim \mathcal{W}(\rho, I_k)$ 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:

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

where $\Gamma_p(a) \equiv \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma\left(a + \frac{1-i}{2}\right)$ 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:

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

where $\lambda_{\max}$ 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:

$$\begin{aligned}\Pr\left\{\lambda_{\max}(Q^T(PP^T)^{-1}Q) \cdot \frac{\rho - k - 1}{k} \geq t\right\} &= \Pr\left\{\lambda_{\max}(J^TJ)\frac{\nu - 2}{k} \geq t\right\} \\ &= \Pr\left\{\lambda_{\max}(JJ^T)(\nu - 2) \geq tk\right\}\end{aligned}$$

$$\begin{aligned} &\leq \Pr\left\{\mathrm{Tr}(JJ^T)(\nu-2) \geq tk\right\} \\ &= \Pr\left\{\sum_{j=1}^{k} \|J_j\|^2 \cdot (\nu-2) \geq tk\right\} \end{aligned} \tag{7}$$

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 ν. In particular, J_j above is distributed as a multivariate t with ν 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.

Lemma 1 (Chernoff-type bound on square norm of t distributed random vectors) *Let $x \sim T_d(0, I_d, \nu)$. Then $\forall t > d$,*

$$Pr\left\{\|x\|^2 > t\right\} \leq \left(\frac{d}{t}\right)^{-\frac{d}{2}} \left(\frac{d+\nu}{t+\nu}\right)^{\frac{\nu+d}{2}} \tag{8}$$

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. For $t > d$, we have:*

$$\lim_{\nu\to\infty} \left(\frac{d}{t}\right)^{-\frac{d}{2}} \left(\frac{d+\nu}{t+\nu}\right)^{\frac{\nu+d}{2}} = \left(\frac{d}{t}\right)^{-\frac{d}{2}} \exp\left(-\frac{t-d}{2}\right) \geq Pr\left\{\|y\|^2 > t\right\} \tag{9}$$

For finite ν, the following bound highlights that the r.h.s. in Lemma 1 tightens with increasing ν. That is, concentration is better for higher degrees of freedom. This agrees with the intuition that a heavier tail (smaller ν) implies more mass spread-out in the tails, hence concentration becomes weaker.

Remark 3 *For $t > d$ the following holds:*

$$\left(\frac{d}{t}\right)^{-\frac{d}{2}} \left(\frac{d+\nu}{t+\nu}\right)^{\frac{\nu+d}{2}} \leq \left(\frac{t}{d}\right)^{\frac{d}{2}} \exp\left(-\frac{t-d}{2} \cdot \frac{d+\nu}{t+\nu}\right) \tag{10}$$

We apply our Lemma 1 to the r-dimensional random vector $J_j\sqrt{\nu-2} \sim T_r(0, I_r, \nu)$, so we can further bound the right hand side (RHS) of eq. (7) for all $t > c \cdot r$, where $c > 1$ is a constant. Since k is finite, we have:

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

$$\leq k \cdot \left(\frac{t}{r}\right)^{\frac{r}{2}} \cdot \left(\frac{r+\nu}{t+\nu}\right)^{\frac{\nu+r}{2}} \tag{12}$$

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 $\exists \eta, c_{MSR} > 0$ constants s.t.*

$$Pr\{\|AUA\| \geq t\} \quad \leq \quad \frac{c_{MSR}}{t^{1+\eta}}, \forall t \geq c_{MSR} \cdot rank(A), \forall A \text{ orthogonal projection in } \mathbb{R}^d$$

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 η 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 $\hat{\text{ic}}_k$ of eq. (2) as the following, where an appropriate transformation is included in order to satisfy $\mathrm{E}[U_{(M)}] = I_d$:

$$U_{(M)} \equiv \mathrm{E}[R^T(RMR^T)^{-1}R]^{-1/2} \cdot R^T(RMR^T)^{-1}R \cdot \mathrm{E}[R^T(RMR^T)^{-1}R]^{-1/2} \tag{14}$$

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

$$U_{(M)} \sim L \cdot U_{(\Lambda)} \cdot L^T \tag{15}$$

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 Λ. Indeed,

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

and U_Λ satisfies $\mathrm{E}[U_{(\Lambda)}] = I_d$.

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.

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 $r \in \{1, ..., d\}$. 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 $\Pr\left\{\|AU_{(\Lambda)}A\| \geq t\right\}$ for all $t > c_{MSR} \cdot r$.

Denote by $\underline{\Lambda}$ the $\rho \times \rho$ diagonal matrix of the non-zero eigenvalues of M, and it will be useful to split R as $R = \left[\begin{array}{cc} P & S \end{array}\right]$ into the $k \times \rho$ matrix P and the $k \times \bar{\rho}$ matrix S, where $\bar{\rho} = d - \rho$.

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

$$U_{(\Lambda)} = \left[\begin{array}{cc} V_{(\underline{\Lambda})} & Z_{(\underline{\Lambda})} \\ Z^T_{(\underline{\Lambda})} & W_{(\underline{\Lambda})} \end{array}\right] \tag{17}$$

where

$$V_{(\underline{\Lambda})} = \mathrm{E}[P^T(P\underline{\Lambda}P^T)^{-1}P]^{-1/2} \cdot P^T(P\underline{\Lambda}P^T)^{-1}P \cdot \mathrm{E}[P^T(P\underline{\Lambda}P^T)^{-1}P]^{-1/2} \tag{18}$$

$$W_{(\underline{\Lambda})} = \mathrm{E}[S^T(P\underline{\Lambda}P^T)^{-1}S]^{-1/2} \cdot S^T(P\underline{\Lambda}P^T)^{-1}S \cdot \mathrm{E}[S^T(P\underline{\Lambda}P^T)^{-1}S]^{-1/2} \tag{19}$$

$$Z_{(\underline{\Lambda})} = \mathrm{E}[P^T(P\underline{\Lambda}P^T)^{-1}P]^{-1/2} \cdot P^T(P\underline{\Lambda}P^T)^{-1}S \cdot \mathrm{E}[S^T(P\underline{\Lambda}P^T)^{-1}S]^{-1/2} \tag{20}$$

Then we can split the problem by showing the following:

Lemma 6 *Denote by $\underline{\Lambda}$ the $\rho \times \rho$ diagonal matrix of the non-zero eigenvalues of M, and A is a rank r projection matrix in $\mathbb{R}^d$. For U defined as above, we have:*

$$\|AU_{(\Lambda)}A\| \leq \|A_1 V_{(\underline{\Lambda})} A_1\| + \|A_2 W_{(\underline{\Lambda})} A_2\| \tag{21}$$

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, constructed as the following: Let B be s.t. $A = B^T(BB^T)B$. Decompose the $r \times d$ matrix B as a sum of two matrices of which the first matrix contains the first ρ columns of B and zeros in its last $\bar{\rho}$ columns, and the second matrix has zeros in its first ρ columns followed by the remaining $\bar{\rho}$ columns of B – that is, $B = \left[\begin{array}{cc} B_1 & 0 \end{array}\right] + \left[\begin{array}{cc} 0 & B_2 \end{array}\right]$. Now, $A_i := B_i(B_iB_i^T)^{-1/2}B_i, i \in \{1, 2\}$.

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 $\kappa(\underline{\Lambda}) \equiv \lambda_{\max}(M)/\lambda_{\min\neq 0}(M) = \lambda_{\max}(\underline{\Lambda})/\lambda_{\min}(\underline{\Lambda})$ 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 $\underline{\Lambda} = I_\rho$. Generalisation to other M then presents no difficulty using the following Lemma 7.

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

$$\|A_1 V_{(\underline{\Lambda})} A_1\| \leq \kappa(\underline{\Lambda}) \cdot \|A_1 V(I_\rho) A_1\| \tag{22}$$

$$\|A_2 W_{(\underline{\Lambda})} A_2\| \leq \kappa(\underline{\Lambda}) \cdot \|A_2 W(I_\rho) A_2\| \tag{23}$$

where I_ρ is the ρ-dimensional identity matrix.

The advantage is that the expectation that appears in $\mathrm{ic}_k(M_0)$ has a closed form in this case:

$$\mathrm{ic}_k(M_0) = \begin{bmatrix} \frac{k}{\rho} \cdot I_\rho & 0 \\ 0 & \frac{k}{\rho-k-1} I_{\bar{\rho}} \end{bmatrix} \tag{24}$$

This can be seen easily by computing:

$$\begin{aligned} \mathrm{E}[V_{(I_\rho)}] &= \frac{k}{\rho} I_\rho \\ \mathrm{E}[W_{(I_\rho)}] &= \frac{k}{\rho-k-1} I_\rho \\ \mathrm{E}[Z_{(I_\rho)}] &= 0 \end{aligned}$$

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 $V_{(\underline{\Lambda})}$ satisfies the MSR with η and $c_{MSR} \cdot (\kappa(\underline{\Lambda}))^{1+\eta}$. Likewise, if $W_{(I_\rho)}$ satisfies the MSR with $\eta', c'_{MSR} > 0$, then $W_{(\underline{\Lambda})}$ satisfies the MSR with η' and $c'_{MSR} \cdot (\kappa(\underline{\Lambda}))^{1+\eta'}$.

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

$$\begin{aligned} U &= \begin{bmatrix} P^T(PP^T)^{-1}P \cdot \frac{\rho}{k} & P^T(PP^T)^{-1}S \cdot \frac{\sqrt{\rho(\rho-k-1)}}{k} \\ S^T(PP^T)^{-1}P \cdot \frac{\sqrt{\rho(\rho-k-1)}}{k} & S^T(PP^T)^{-1}S \cdot \frac{\rho-k-1}{k} \end{bmatrix} \\ &\equiv \begin{bmatrix} V & Z \\ Z^T & W \end{bmatrix} \end{aligned}$$

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 ρ/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 $\|A_2 W A_2\| = \|W\|$.

3.3. Main result

Using the preparatory work in the previous section, we can prove the following:

Theorem 8 *Assume* $3 \leq \rho < d$, k *finite,* $k < \rho - 1$, $\nu \equiv \rho - k + 1$, *and there exists* $c, \eta > 0$ *constants such that* $k \left(\frac{t}{\bar{\rho}}\right)^{\frac{\bar{\rho}}{2}} \left(\frac{\bar{\rho}+\nu}{t+\nu}\right)^{\frac{\bar{\rho}+\nu}{2}} t^{1+\eta} \leq c, \forall t \geq c \cdot \bar{\rho}$. *Then* $R^T(RM_0R^T)^{-1}R \cdot \begin{bmatrix} k/\rho \cdot I_\rho & 0 \\ 0 & k/(\rho - k - 1) \cdot I_{\bar{\rho}} \end{bmatrix}$ *satisfies the MSR with* (c, η).

Remark 9 *It can be shown*[1] *that the constant* c *required in Theorem 8 exists* for any *choice of* η, *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 ρ is always no larger than the sample size. In fact, ρ 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 ρ 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 ρ 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:*

$$E\left[\| \frac{1}{m} \sum_{i=1}^{m} R_i^T (R_i M_0 R_i^T)^{-1} R_i - E\left[R^T (RM_0R^T)^{-1}R\right] \|\right] \leq \epsilon \cdot \| E\left[R^T (RM_0R^T)^{-1}R\right] \| = \epsilon \cdot \frac{k}{\rho - k - 1}$$

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.

The proofs are provided in the supplementary matrial[2]. The next section provides a numerical demonstration of our main finding.

1. Yaakov Baruch (https://mathoverflow.net/users/2480/yaakov-baruch), Implausible inequality?, URL (version: 2017-08-20): https://mathoverflow.net/q/279128
2. http://www.cs.bham.ac.uk/~axk/64-proofs.pdf

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:

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

The expectations in the above have analytic forms, which we used. We increased the ensemble size m progressively until ϵ 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 ν) 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 ϵ 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 η, but it increases with the same through c_{MSR}. There is clearly a tradeoff in choosing ν – 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 ρ/k is bounded is quite natural).

5. 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.

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 ν degrees of freedom belongs to the family of $-1/\nu$-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.

A more distant extension of interest would be to consider other ensemble combinations in the context of machine learning, for instance constrained mixture ensembles (Nabney et al., 2005), especially in models where theoretical guarantees are scarce.

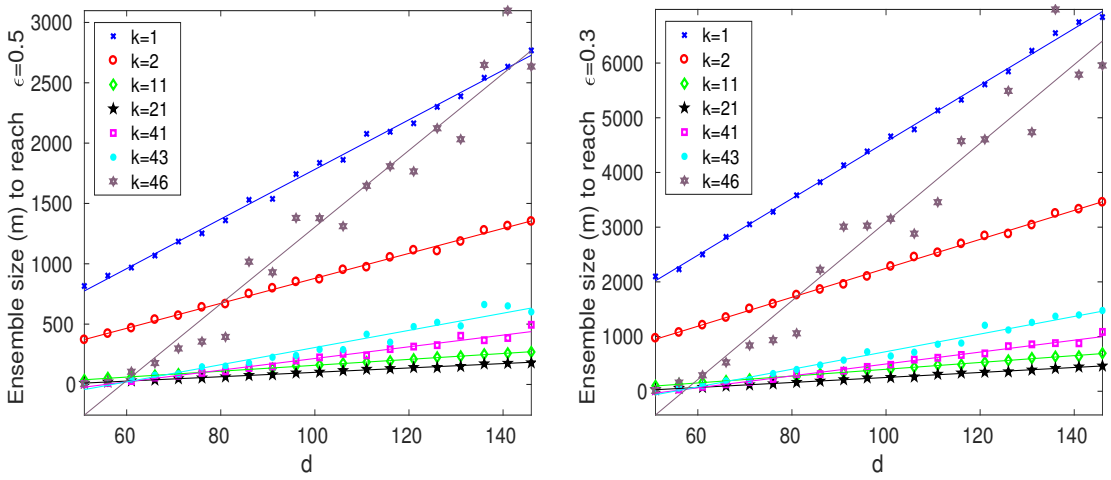

Figure 1: Numerical experiment demonstrating that the required ensemble size grows linearly in d, for $\epsilon = 0.5$, and $\epsilon = 0.3$. We used $M = M_0$ with $\rho = 50$.

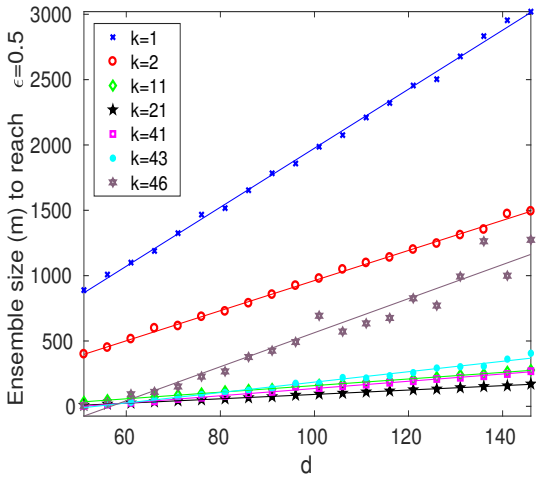

Figure 2: Similar experiment, now using a generic singular matrix M of rank $\rho = 50$ (generated randomly and then fixed) with condition number $\kappa(M) = 1.5$ in its range space. As suggested by Lemma 7, the required ensemble size remains linear in d.

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

R. Adamczak, O. Guédon, R. Latala, K. Oleszkiewicz, A.E. Litvak, A. Pajor, N. Tomczak-Jaegermann. Moment estimates for convex measures. Electronic J. of Probability 17(101):1-19, 2012.

R. Ahlswede, A. Winter, Strong converse for identification via quantum channels, IEEE Trans. Information Theory 48:568-579, 2002.

C. Borell. Convex set functions in d-space, Periodica Math. Hungarica 6:111-136, 1975.

K. Chandrasekaran, A. Deshpande, S. Vempala. Sampling s-concave functions. In Proc. of 13th Intl. Workshop on Randomization and Computation, 5687:420-433, 2009.

J.A. Daz-García, R. Gutiérrez-Jáimez. Matricvariate and matrix multivariate T distributions and associated distributions. Metrika 75(7):963-976, 2012.

R.J. Durrant, A. Kabán. Random projections as regularizers: Learning a linear discriminant from fewer observations than dimensions. Machine Learning 99(2):257-286, 2015.

A.K. Gupta, D.K. Nagar. Matrix variate distributions. CRC Press, 1999.

T. Ideker, R. Sharan. Protein networks in disease. Genome research 18(4):644-652, 2008.

A. Kabán, R.J. Durrant. Learning with $\ell_{q<1}$ vs. ℓ_1-norm regularization with exponentially many irrelevant features. In Proc. of 19th European Conference on Machine Learning (ECML), LNAI 5211:580-596, 2008.

N. Krämer, J. Schäfer, A. Boulesteix. Regularized estimation of large-scale gene association networks using graphical Gaussian models. BMC Bioinformatics 10:384, 2009.

O. Ledoit, M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. J. of Multivariate Analysis 88(2):365-411, 2004.

T. Marzetta, G. Tucci, S. Simon. A random matrix theoretic approach to handling singular covariance estimates. IEEE Trans. on Information Theory 57(9):6256-6271, 2011.

N. Meinshausen, P. Bühlmann. High-dimensional graphs and variable selection with the lasso. Annals of Statistics 1436-1462, 2006.

I. Nabney, Y. Sun, P. Tiňo, A. Kabán. Semisupervised learning of hierarchical latent trait models for data visualization. IEEE Trans. on Knowledge and Data Engineering 17(3):384-400, 2005.

A. Ng. Feature selection, ℓ_1 vs. ℓ_2 regularization, and rotational invariance. In Proc. of 21-st International Conference on Machine Learning (ICML), 2004.

M. Rudelson. Random vectors in the isotropic position. J. of Functional Analysis 164(1):60-72, 1999.

G.A. Thanei, C. Heinze, N. Meinshausen. Random projections for large-scale regression, arXiv:1701.05325 [math.ST], 2017.

P. Youssef. Estimating the covariance of random matrices. Electronic J. of Probability 18:107, 2013.

M. Yuan, Y. Lin. Model selection and estimation in the Gaussian graphical model. Biometrika, 19-35, 2007.