Sufficient ensemble size for random matrix theory-based handling of singular covariance matrices

Kaban, Ata

DOI:
10.1142/S0219530520400072

Document Version
Peer reviewed version

Citation for published version (Harvard):
Kaban, A 2020, 'Sufficient ensemble size for random matrix theory-based handling of singular covariance matrices', *Analysis and Applications*, vol. 18, no. 5, pp. 929-950. https://doi.org/10.1142/S0219530520400072

Sufficient Ensemble Size for Random Matrix Theory based Handling of Singular Covariance Matrices

Ata Kabán

School of Computer Science, University of Birmingham, Edgbaston, B15 2TT, Birmingham, United Kingdom
A.Kaban@cs.bham.ac.uk

Received (Day Month Year)
Revised (Day Month Year)

Singular covariance matrices are frequently encountered in both machine learning and optimization problems, most commonly due to high dimensionality of data and insufficient sample sizes. Among many methods of regularization, here we focus on a relatively recent random matrix theoretic approach, the idea of which is to create well-conditioned approximations of a singular covariance matrix and its inverse by taking the expectation of its random projections. We are interested in the error of a Monte Carlo implementation of this approach, which allows subsequent parallel processing in low dimensions in practice. We find that $\mathcal{O}(d)$ random projections, where d is the size of the original matrix, are sufficient for the Monte Carlo error to become negligible, in the sense of expected spectral norm difference, for both covariance and inverse covariance approximation, in the latter case under mild assumptions.

Keywords: Singular covariance; precision matrix; curse of dimensionality; random projections; Monte Carlo error.

Mathematics Subject Classification 2010: 68T99, 15B52, 15A15

1. Introduction

Dealing with singular covariance matrices, and obtaining well-conditioned, invertible approximations thereof, represent common issues in many high dimensional learning and optimization settings, where the available sample size is too small relative to the feature dimension of the data. In machine learning, examples include classification and clustering with multivariate Gaussians, least squares regression, and Gaussian graphical models [30,24]. In large scale black-box optimization, similar problems are encountered by the class of model-building algorithms known as Estimation of Distribution Algorithms [19], which are particularly prone to the curse of dimensionality on high dimensional search spaces [16].

Many methods have been proposed, the problem is closely related to regularization [7,10], which has been extensively studied. For covariance matrices, a common idea is to have some restriction on the number of free parameters that describe the covariance structure. Two major branches of methods include rotation-sensitive

methods, for example methods based on sparsity or structured sparsity restrictions – these assume that only few features correlate with each other – and rotation-invariant methods, such as the Ledoit-Wolf estimator [20], ridge regularization, and more recently proposed random projection ensembles [22].

This paper is concerned with the latter approach, which is computationally attractive and only requires cheaply collected random projections of the data. In addition, the lack of a-priori structural assumptions also means that such methods remain appropriate when there are no known or justifiable structural assumptions to exploit. For instance, gene or protein association networks often present complex and dense interactions between many genes or proteins at some stage of disease development [18,13], making sparsity assumptions unjustified.

In this vein, Marzetta et al. [22] proposed two general-purpose approaches that take a given non-random positive semi-definite singular covariance matrix and use its random projections to construct a non-singular approximation of it, or of its inverse, defined in the form of matrix expectations.

The authors [22] conducted a detailed theoretical analysis on these matrix expectations. However, in practice a Monte Carlo average would be employed instead – that is, a finite ensemble that averages estimates from multiple random projections. To give some examples, the covariance approximation ensemble scheme is encountered in optimization heuristics [16] where its role is to drive the search for a global optimum within a high dimensional search space. The inverse covariance approximation scheme is encountered in machine learning for Fisher discriminant analysis in high dimensional / small sample settings [9], or in learning an ensemble of compressive OLS regressors [28]. This paper is concerned with the question of how large the finite ensemble needs to be so that the error of the Monte Carlo average and its expectation is below a user-specified threshold in terms of the expected spectral norm difference.

In formal terms, given a non-random $d \times d$ singular, positive semi-definite, rank $\rho < d$ matrix M, and an integer k, we consider the following covariance and inverse covariance approximators:

$$\mathrm{cov}_k(M) = E_R[R^T RMR^T R] \tag{1.1}$$

$$\mathrm{cov}_k^-(M) = E_R[R^T (RMR^T)^{-1} R] \tag{1.2}$$

where R is a random $k \times d$, matrix with i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ entries. Following the literature, R is called a random projection matrix, and the interesting case is when $k < \rho - 1$. The original definitions in [22] employed Haar distributed random matrices rather than Gaussian, however when d is large the rows of R are nearly orthogonal anyway, due to concentration of measure, so in practice orthogonalization of the rows of R may be omitted to save computation time. Also observe that in the case of $\mathrm{cov}_k^-(M)$, eq. (1.2) is not affected by whether R has orthogonal rows or not.

Let $R_1, R_2, ..., R_m$ be independent copies of R. We are interested in the following

matrix averages:

$$\hat{\mathrm{cov}}_k(M) = \frac{1}{m}\sum_{i=1}^{m} R_i^T R_i M R_i^T R_i \tag{1.3}$$

$$\hat{\mathrm{cov}}_k^-(M) = \frac{1}{m}\sum_{i=1}^{m} R_i^T (R_i M R_i^T)^{-1} R_i \tag{1.4}$$

The question we study is how large m needs to be so that $\mathrm{E}_R[\|\mathrm{cov}_k(M) - \hat{\mathrm{cov}}_k(M)\|]$, and $\mathrm{E}_R[\|\mathrm{cov}_k^-(M) - \hat{\mathrm{cov}}_k^-(M)\|]$ respectively, are below some pre-defined threshold?

In [15] we presented initial findings about the inverse covariance approximator, eq. (1.4). This version goes into more depth and detail, and contains additional results about the covariance approximator, eq. (1.3). The latter is able to explain previous empirical observations about the sufficient ensemble size [16] that eluded previous analysis.

1.1. *Context and Summary of Main Results*

In covariance estimation, it is known from the work of [26] that, for general distributions with support on the sphere of radius $\sqrt{d}$ the required sample size is[a] $m = \mathcal{O}(d \log d)$, 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 of the latter kind [1].

Relatively recently, work by [29] extended such results to the matrix-covariance setting – that is, instead of random vectors consider random matrices and their covariance – and gave some generic conditions under which $m = \mathcal{O}(d)$. This order has been known for a long time for sums of certain well-behaved matrix distributions, such as sub-Gaussians [3]. However, the classical Ahlswede & Winter matrix concentration inequalities [3] employed on $\hat{\mathrm{cov}}_k$ in only lead to an estimate of $m = \mathcal{O}(d \log d)$, while practical experience suggested the conjecture of $m = \mathcal{O}(d)$ [16]. Moreover, in the case of $\hat{\mathrm{cov}}_k^-$, the presence of matrix inverses combined with the singularity of M give rise to a heavy tailed sub-matrix, and it is far from obvious whether an ensemble size of $m = \mathcal{O}(d)$ could possibly suffice under realistically reasonable assumptions.

We summarize below our main results. These will be formally stated in Theorems 4.1 and 5.1 respectively. Throughout our analysis, k and ρ are fixed integers, $k < \rho - 1$, and M is a fixed $d \times d$ positive semi-definite matrix of rank $\rho < d$ with condition number in its range space, $\kappa(M) \equiv \lambda_{\max}(M)/\lambda_\rho(M)$, upper bounded independently of d. The ambient dimension d is typically large. We will use the notations $\rho \equiv \mathrm{rank}(M)$, and $\bar{\rho} \equiv d - \rho \geq 1$, and $\|\cdot\|$ with a matrix argument will denote the spectral norm.

[a]The uniform distribution on canonical basis vectors in $\mathbb{R}^d$ is an illustrative example where the necessity of $\mathcal{O}(d \log d)$ points to obtain a non-singular covriance w.h.p. follows from the coupon collector problem.

- For any $\epsilon \in (0,1)$, in order to ensure $\mathrm{E}_R\left[\|\hat{\mathrm{cov}}_k(M) - \mathrm{cov}_k(M)\|\right] \leq \epsilon \cdot \|\mathrm{cov}_k(M)\|$, it is sufficient to take $m = \mathcal{O}(d)$.
- Suppose that $3 \leq \rho < d$, $\rho > k+1$, and $\rho - k + 1 > a\log(\bar{\rho}) + a$ for some constant $a > 0$. For any $\epsilon \in (0,1)$, in order to ensure $\mathrm{E}_R\left[\|\hat{\mathrm{cov}}_k^-(M) - \mathrm{cov}_k^-(M)\|\right] \leq \epsilon \cdot \|\mathrm{cov}_k^-(M)\|$ it is sufficient to take $m = \mathcal{O}(d)$.

The requirement that appears in the second statement, $\rho - k + 1 > a\log(\bar{\rho}) + a$ is rather mild. Essentially it says that M must have rank at least logarithmic in its null-space dimension $\bar{\rho}$. For instance, if M arises from a sample covariance, then ρ is always no larger than the sample size; thus, roughly speaking, a setting with exponentially many irrelevant features relative to the sample size can still satisfy this requirement.

2. Tools

This section lists and develops some re-usable analytic tools, which will be employed for proving our results. The next two subsections present techniques that help reduce the problem from the ensemble level to the individual matrix level. The remaining two subsections contain tools to deal with the latter.

2.1. *A specific result from random matrix theory*

The following result due to [29] gives sufficient conditions for a finite average of low-rank covariance matrices to approach their expectation with $\mathcal{O}(d)$ (as opposed to $\mathcal{O}(d\log d)$) independent terms.

Definition 2.1 ([29]). A positive semi-definite random matrix $U = U^T$ of dimension $d \times d$ and $\mathrm{E}[U] = I_d$ satisfies the matrix strong regularity (MSR) condition if $\exists \eta, c_{MSR} > 0$ constants such that,

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

For our purposes, the random matrix U will be the generic term of $\hat{\mathrm{cov}}_k$ or $\hat{\mathrm{cov}}_k^-$ subjected to an isotropic transformation, and will be defined in Section 3, eq. (3.1). The projection matrix A is deterministic and should not be confused with the random projections.

Theorem 2.1 ([29]). *Let U be a $d \times d$ positive semi-definite random 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:*

$$E[\|\frac{1}{m}\sum_{i=1}^{m} U_i - I_d\|] \leq \epsilon \tag{2.1}$$

where C_1 is a constant that depends only on η and c_{MSR}.

2.2. *Splitting device*

The setting of Theorem 2.1 closely resembles our problem at a high level, however the structure of the matrices of our interest will need different treatment for different sub-matrices. As we shall see later in the analysis, this is due to the singularity of M that induces different distributions on its range space and its null space respectively.

We write the generic term U as the following:

$$U = \begin{bmatrix} V & Z \\ Z^T & W \end{bmatrix} \tag{2.2}$$

where V and W are $\rho \times \rho$ and $\bar{\rho} \times \bar{\rho}$ positive semi-definite sub-matrices respectively.

With the aim to prove a MSR condition, we take an arbitrary $d \times d$ 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 $\Pr\{\|AUA\| \geq t\}$ for all $t > c_{MSR} \cdot r$. The following lemma implies that we can split this problem and it is sufficient to have the MSR condition separately on the block-diagonal sub-matrices of U.

Lemma 2.1. *For U defined in eq. (2.2), there exists a $\rho \times \rho$ projection matrix A_1 and a $\bar{\rho} \times \bar{\rho}$ projection matrix A_2, both of rank r in $\mathbb{R}^{\rho}$ and $\mathbb{R}^{\bar{\rho}}$ respectively, such that:*

$$\|AUA\| \leq \|A_1 V A_1\| + \|A_2 W A_2\| \tag{2.3}$$

We note that, in the original problem, r takes values in $\{1, 2, ..., d\}$ – however, 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 can choose A_1 to have $\|A_1 V A_1\| = \|V\|$, and likewise A_2 for $r > \bar{\rho}$ to have $\|A_2 W A_2\| = \|W\|$. Hence, Lemma 2.1 implies that if both V and W satisfy the MSR condition then U satisfies MSR.

Proof. [Proof of Lemma 2.1] Recall that A is a projection matrix. Let us rewrite the matrix norm of interest as the following:

$$\begin{aligned} \|AUA\| &= \|B^T(BB^T)^{-1}BUB^T(BB^T)^{-1}B\| \\ &= \|(BB^T)^{-1/2}BUB^T(BB^T)^{-1/2}\| \\ &\equiv \|\mathfrak{B}U\mathfrak{B}^T\| \end{aligned} \tag{2.4}$$

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 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.

$$\begin{aligned} B &= \begin{bmatrix} B_1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & B_2 \end{bmatrix} \\ &= (B_1 B_1^T)^{1/2} \begin{bmatrix} \mathfrak{B}_1 & 0 \end{bmatrix} + (B_2 B_2^T)^{1/2} \begin{bmatrix} 0 & \mathfrak{B}_2 \end{bmatrix} \end{aligned} \tag{2.5}$$

where we used the notations $\mathfrak{B}_i \equiv (B_i B_i^T)^{-1/2} B_i, i \in \{1,2\}$. That is, we orthonormalized the nonzero sub-matrices. Using these, we now construct a new positive semi-definite matrix of size $2r \times 2r$:

$$\tilde{U} \equiv \begin{bmatrix} \mathfrak{B}_1 & 0 \\ 0 & \mathfrak{B}_2 \end{bmatrix} \cdot \begin{bmatrix} V & Z \\ Z^T & W \end{bmatrix} \cdot \begin{bmatrix} \mathfrak{B}_1^T & 0 \\ 0 & \mathfrak{B}_2^T \end{bmatrix} \tag{2.6}$$

This is clearly positive semi-definite since the $d \times d$ matrix in the middle was assumed to be positive semi-definite.

By definition, the matrix norm of interest, eq. (2.4) is:

$$\|AUA\| = \max_{x \in \mathbb{R}^r, x \neq 0} \frac{x^T (BB^T)^{-1/2} \begin{bmatrix} (B_1 B_1^T)^{1/2} \\ (B_2 B_2^T)^{1/2} \end{bmatrix}^T \cdot \tilde{U} \cdot \begin{bmatrix} (B_1 B_1^T)^{1/2} \\ (B_2 B_2^T)^{1/2} \end{bmatrix} (BB^T)^{-1/2} x}{x^T x} \tag{2.7}$$

and by observing that

$$\begin{aligned} (BB^T)^{-1/2} \begin{bmatrix} (B_1 B_1^T)^{1/2} \\ (B_2 B_2^T)^{1/2} \end{bmatrix}^T \cdot \begin{bmatrix} (B_1 B_1^T)^{1/2} \\ (B_2 B_2^T)^{1/2} \end{bmatrix} (BB^T)^{-1/2} \\ = (BB^T)^{-1/2} \left(B_1 B_1^T + B_2 B_2^T \right) (BB^T)^{-1/2} \\ = (BB^T)^{-1/2} BB^T (BB^T)^{-1/2} \\ = I_r \end{aligned}$$

and denoting $y = \begin{bmatrix} (B_1 B_1^T)^{1/2} \\ (B_2 B_2^T)^{1/2} \end{bmatrix} (BB^T)^{-1/2} x$, we have:

$$\text{eq.}(2.7) \ \leq \max_{y \in \mathbb{R}^{2r}, y \neq 0} \frac{y^T \tilde{U} y}{y^T y} = \|\tilde{U}\|. \tag{2.8}$$

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:

$$\begin{aligned} \text{eq.}(2.8) &= \|\tilde{U}\| \\ &= \left\| \begin{bmatrix} \mathfrak{B}_1 V \mathfrak{B}_1^T & \mathfrak{B}_1 Z \mathfrak{B}_2^T \\ \mathfrak{B}_2 Z^T \mathfrak{B}_1^T & \mathfrak{B}_2 W \mathfrak{B}_2^T \end{bmatrix} \right\| \\ &\leq \|\mathfrak{B}_1 V \mathfrak{B}_1^T\| + \|\mathfrak{B}_2 W \mathfrak{B}_2^T\| \\ &= \|(B_1 B_1^T)^{-1/2} B_1 V B_1^T (B_1 B_1^T)^{-1/2}\| + \|(B_2 B_2^T)^{-1/2} B_2 W B_2^T (B_2 B_2^T)^{-1/2}\| \\ &\leq \|A_1 V A_1\| + \|A_2 W A_2\| \end{aligned} \tag{2.9}$$

where $A_i \equiv B_i (B_i B_i^T)^{-1/2} B_i, i \in \{1,2\}$ are projections in $\mathbb{R}^\rho$ and $\mathbb{R}^{\bar{\rho}}$ respectively. The inequality (2.9) is not difficult to check, see e.g. [14]. □

2.3. *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 from [11], 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 degrees of freedom

$$\nu = \rho - k + 1. \tag{2.10}$$

Here, and throughout this paper, we refer to the parameterization from [11], so the $k \times r$ matrix J has the following probability density:

$$p(J) = \frac{\Gamma_k\left(\frac{\nu+k+r-1}{2}\right)}{\pi^{kr}\Gamma_k\left(\frac{\nu+k-1}{2}\right)} \det(I_k + JJ^T)^{-\frac{\nu+k+r-1}{2}} \tag{2.11}$$

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 [11].

The goal of this section is to obtain a polynomially decaying upper bound on the largest eigenvalue of $J^TJ = Q^T(PP^T)^{-1}Q$:

$$\Pr\left\{\|Q^T(PP^T)^{-1}Q\| \cdot \frac{\rho - k - 1}{k} \geq t\right\} \leq ? \tag{2.12}$$

where the multiplier on the l.h.s. ensures that the scaled positive semi-definite matrix is isotropic.

We should note that the matrix-variate T is different from a multivariate t vector re-shaped into a matrix [8]. 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:

$$\begin{aligned}
\Pr\left\{\|Q^T(PP^T)^{-1}Q\| \cdot \frac{\rho - k - 1}{k} \geq t\right\} &= \Pr\left\{\|J^TJ\| \cdot \frac{\nu - 2}{k} \geq t\right\} \\
&= \Pr\left\{\|JJ^T\| \cdot (\nu - 2) \geq tk\right\} \\
&\leq \Pr\left\{\mathrm{Tr}(JJ^T)(\nu - 2) \geq tk\right\} \\
&= \Pr\left\{\sum_{j=1}^{k} \|J_j\|^2 \cdot (\nu - 2) \geq tk\right\}
\end{aligned} \tag{2.13}$$

where J_j denotes the j-th row of J, and the vector norm in the last line is the L_2 norm.

By Theorem 4.3.9 in [11], all marginal distributions of the rows (and columns) of J are 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 parameterization we are using, the pdf of this random vector is given by plugging $k = 1$ into eq.

(2.11). 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.

The following lemma bounds the squared norm of a multivariate t-distributed random vector.

Lemma 2.2 (Chernoff-type bound on square norm of t distributed random vectors). *Let $x \sim T_d(0, I_d, \nu)$. Then $\forall t > d$,*

$$Pr\left\{\|x\|^2 > t\right\} \leq \left(\frac{d}{t}\right)^{-\frac{d}{2}} \left(\frac{d+\nu}{t+\nu}\right)^{\frac{\nu+d}{2}} \tag{2.14}$$

Proof. [Proof of Lemma 2.2] We use the following representation of the multivariate t-distribution (see e.g. [21]): If $y \sim \mathcal{N}_p(0, \Sigma)$ and $s^2 \sim \chi^2_\nu$ independent of y, then $\frac{y\sqrt{\nu}}{s} \sim T_p(0, \Sigma, \nu)$.

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

$$\begin{aligned}
\Pr\left\{\|x\|^2 > t\right\} &= \Pr\left\{\frac{\|y\|^2}{\|u\|^2} > \frac{t}{\nu}\right\} && (2.15)\\
&= \Pr\left\{\|y\|^2 > \frac{t}{\nu}\|u\|^2\right\} && (2.16)\\
&= \Pr\left\{\exp\left(\lambda\|y\|^2\right)\exp\left(-\frac{t}{\nu}\lambda\|u\|^2\right) > 1\right\}, \forall\lambda > 0 && (2.17)\\
&\leq \mathrm{E}\left[\exp(\lambda\|y\|^2)\right]\mathrm{E}\left[\exp\left(-\frac{t}{\nu}\lambda\|u\|^2\right)\right], \forall\lambda > 0 && (2.18)\\
&= (1-2\lambda)^{-\frac{d}{2}}\left(1+2\lambda\frac{t}{\nu}\right)^{-\frac{\nu}{2}} \forall\lambda \in (0, 1/2) && (2.19)
\end{aligned}$$

We optimize the bound in λ 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

$$\lambda = \frac{t-d}{2t(d/\nu+1)}$$

Since $t > d$, this is in the interval $(0, 1/2)$ as required, for any positive value of ν.

After plugging back, the RHS of eq. (2.19) becomes:

$$\begin{aligned}
&\left(1 - \frac{t-d}{t(d/\nu+1)}\right)^{-d/2}\left(1 + \frac{t-d}{\nu(d/\nu+1)}\right)^{-\nu/2}\\
&= \left(\frac{d(t+\nu)}{t(d+\nu)}\right)^{-d/2}\left(\frac{t+\nu}{d+\nu}\right)^{-\nu/2}\\
&= \left(\frac{d}{t}\right)^{-\frac{d}{2}}\left(\frac{d+\nu}{t+\nu}\right)^{\frac{\nu+d}{2}}
\end{aligned}$$

□

Remark 2.1. Lemma 2.2 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:

$$\lim_{\nu\to\infty} \left(\frac{d}{t}\right)^{-\frac{d}{2}} \left(\frac{d+\nu}{t+\nu}\right)^{\frac{\nu+d}{2}} = \left(\frac{d}{t}\right)^{-\frac{d}{2}} \exp\left(-\frac{t-d}{2}\right) \geq \Pr\left\{\|y\|^2 > t\right\} \quad (2.20)$$

where $y \sim \mathcal{N}(0, I_d)$

Remark 2.2. For finite ν, the r.h.s. in Lemma 2.2 tightens with increasing ν, which agrees with the intuition that concentration is better with higher degrees of freedom. More formally, it is straightforward to see that we can bound:

$$\left(\frac{d}{t}\right)^{-\frac{d}{2}} \left(\frac{d+\nu}{t+\nu}\right)^{\frac{\nu+d}{2}} \leq \left(\frac{t}{d}\right)^{\frac{d}{2}} \exp\left(-\frac{t-d}{2} \cdot \frac{d+\nu}{t+\nu}\right) \quad (2.21)$$

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

Since k is finite, using Lemma 2.2, 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. (2.13) for all $t > c \cdot r$, where $c > 1$ is a constant, as the following:

$$\text{eq. (2.13)} \leq \sum_{j=1}^{k} \Pr\left\{\|J_j \cdot \sqrt{\nu-2}\|^2 \geq t\right\} \quad (2.22)$$

$$\leq k \cdot \left(\frac{t}{r}\right)^{\frac{r}{2}} \cdot \left(\frac{r+\nu}{t+\nu}\right)^{\frac{\nu+r}{2}} \quad (2.23)$$

2.4. *Upper bound on the norm of log-concave random vectors*

A large family of distributions, known as logarithmically concave distributions exhibits concentration that is stronger than what is required for MSR. Proposition 8.5 in [29] showed MSR for random matrices whose vectorized form is a log-concave random vector, based on Paouris' inequality, which gives a tail bound on the norm of any log-concave random vector.

The definition below is in fact Borell's characterization of this class of distributions [5]. For more background and useful properties see e.g. [27].

Definition 2.2. A random (vector) variable having a density $p()$ is said to be log-concave if the function $-\log p()$ is convex.

The following inequality is due to Paouris [25,2].

Theorem 2.2 ([25]). *If X is an isotropic log-concave random vector in $\mathbb{R}^d$, then there exists an absolute constant $c > 0$ s.t. for any $\epsilon \geq 1$,*

$$Pr\left\{\|X\| \geq c\epsilon\sqrt{d}\right\} \leq \exp(-\epsilon\sqrt{d}) \quad (2.24)$$

3. Reductions for $\hat{\text{cov}}_k^{\pm}$

We shall use the notation $\hat{\text{cov}}_k^{\pm}$ (and $\text{cov}_k^{\pm}$) to refer to collectively to the covariance approximator $\hat{\text{cov}}_k$ (and cov_k) and the inverse covariance approximator $\hat{\text{cov}}_k^{-}$ (and cov_k^{-}). Further, denote by $U_{(M)}^{\pm}$ the isotropic transformation of the generic term of the matrix sum in $\hat{\text{cov}}_k^{\pm}(M)$, from eqs. (1.3)-(1.4), as the following:

$$U_{(M)}^{\pm} \equiv \text{E}[R^T(RMR^T)^{\pm 1}R]^{-1/2} \cdot R^T(RMR^T)^{\pm 1}R \cdot \text{E}[R^T(RMR^T)^{\pm 1}R]^{-1/2} \quad (3.1)$$

These two random matrices given in eq. (3.1), i.e. $U_{(M)}$ and $U_{(M)}^{-}$ will play the role of the matrix U that appeared in Definition 2.1. It is straightforward to check that both satisfy $\text{E}[U_{(M)}^{\pm}] = I_d$.

We shall observe a series of properties and reductions leading on to establishing that MSR holds for $U_{(M)}^{\pm}$.

Lemma 3.1 (M can be assumed diagonal w.l.o.g.). *Let $M = L\Lambda L^T$ be the singular value decomposition of M, so Λ is the $d \times d$ diagonal matrix of the non-negative eigenvalues of M, and $LL^T = L^TL = I_d$. We have:*

$$E[\|\frac{1}{m}\sum_{i=1}^{m} U_{i(M)}^{\pm} - I_d\|] = E[\|\frac{1}{m}\sum_{i=1}^{m} U_{i(\Lambda)}^{\pm} - I_d\|] \quad (3.2)$$

Proof. We will refer to a generic term of the sum by dropping the index i. Since R has i.i.d. Gaussian entries, it has the same distribution as RL. We can check that $U_{(M)}^{\pm}$ has the same distribution as:

$$U_{(M)}^{\pm} \sim L \cdot U_{(\Lambda)}^{\pm} \cdot L^T \quad (3.3)$$

Indeed, since RL has the same distribution as R, we have

$$\begin{aligned}
\text{E}[R^T(RMR^T)^{\pm 1}R]^{-1/2} &= \left(L\text{E}[L^TR^T(RL\Lambda L^TR^T)^{\pm 1}RL]L^T\right)^{-1/2} && (3.4)\\
&= \left(L\text{E}[R^T(R\Lambda R^T)^{\pm 1}R]L^T\right)^{-1/2} && (3.5)\\
&= L\text{E}[R^T(R\Lambda R^T)^{\pm 1}R]^{-1/2}L^T && (3.6)
\end{aligned}$$

where in the last line we used the fact that, for diagonal Λ, the matrix $\text{E}[R^T(R\Lambda R^T)^{\pm 1}R]$ is diagonal [16,9,22], in other words $\text{E}[R^T(RMR^T)^{\pm 1}R]$ has the same eigenvectors as M.

Plugging this back into the definition of $U_{(M)}^{\pm}$ confirms eq. (3.3), and we have:

$$\begin{aligned}
\text{E}[\|\frac{1}{m}\sum_{i=1}^{m} U_{i(M)}^{\pm} - I_d\|] &= \text{E}[\|L(\frac{1}{m}\sum_{i=1}^{m} U_{i(\Lambda)}^{\pm} - I_d)L^T\|]\\
&= \text{E}[\|\frac{1}{m}\sum_{i=1}^{m} U_{i(\Lambda)}^{\pm} - I_d\|]
\end{aligned}$$

□

Next, we shall therefore aim to obtain a result of the form of eq. (2.1) for terms $U_i^{\pm}$ of the form $U_{(\Lambda)}^{\pm}$, in other words, we can identify M with Λ without loss of generality (w.l.o.g.). Note also that $U_{(\Lambda)}^{\pm}$ is isotropic, i.e. it satisfies $\text{E}[U_{(\Lambda)}^{\pm}] = I_d$.

Take a $d \times d$ projection matrix of rank $r \in \{1, ..., d\}$. We shall now use our splitting device, Lemma 2.1. Denote by $\underline{\Lambda}$ the $\rho \times \rho$ diagonal matrix of the non-zero eigenvalues of M, and we 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 $\bar{\rho} = d - \rho$. We can express $U^{\pm}_{(\Lambda)}$ as the following:

$$U^{\pm}_{(\Lambda)} = \begin{bmatrix} V^{\pm}_{(\underline{\Lambda})} & Z^{\pm}_{(\underline{\Lambda})} \\ Z^{\pm T}_{(\underline{\Lambda})} & W^{\pm}_{(\underline{\Lambda})} \end{bmatrix} \tag{3.7}$$

where

$$V^{\pm}_{(\underline{\Lambda})} = \mathrm{E}[P^T(P\underline{\Lambda}P^T)^{\pm 1}P]^{-1/2} \cdot P^T(P\underline{\Lambda}P^T)^{\pm 1}P \cdot \mathrm{E}[P^T(P\underline{\Lambda}P^T)^{\pm 1}P]^{-1/2} \tag{3.8}$$

$$W^{\pm 1}_{(\underline{\Lambda})} = \mathrm{E}[S^T(P\underline{\Lambda}P^T)^{\pm 1}S]^{-1/2} \cdot S^T(P\underline{\Lambda}P^T)^{\pm 1}S \cdot \mathrm{E}[S^T(P\underline{\Lambda}P^T)^{\pm}S]^{-1/2} \tag{3.9}$$

$$Z^{\pm}_{(\underline{\Lambda})} = \mathrm{E}[P^T(P\underline{\Lambda}P^T)^{\pm 1}P]^{-1/2} \cdot P^T(P\underline{\Lambda}P^T)^{\pm 1}S \cdot \mathrm{E}[S^T(P\underline{\Lambda}P^T)^{\pm 1}S]^{-1/2} \tag{3.10}$$

Therefore, by Lemma 2.1 we have

$$\|AU^{\pm}_{(\Lambda)}A\| \leq \|A_1V^{\pm}_{(\underline{\Lambda})}A_1\| + \|A_2W^{\pm}_{(\underline{\Lambda})}A_2\| \tag{3.11}$$

Before proceeding to bound each term on the r.h.s., one more simplification will be handy. We shall assume that $\kappa(\underline{\Lambda}) \equiv \|\underline{\Lambda}\| \cdot \|\underline{\Lambda}^{-1}\|$ is bounded by some constant independent of d. This is reasonable, since this is the condition number of the non-random $\rho \times \rho$ matrix $\underline{\Lambda}$, and $\rho < d$.

Lemma 3.2 (Taking $\underline{\Lambda} = I_\rho$ only changes the constants). *Assume that $\kappa(\underline{\Lambda})$ is bounded by a constant independent of d. Then,*

$$\|A_1V^{\pm}_{(\underline{\Lambda})}A_1\| \leq \kappa(\underline{\Lambda}) \cdot \|A_1V^{\pm}_{(I_\rho)}A_1\| \tag{3.12}$$

$$\|A_2W^{\pm}_{(\underline{\Lambda})}A_2\| \leq \kappa(\underline{\Lambda}) \cdot \|A_2W^{\pm}_{(I_\rho)}A_2\| \tag{3.13}$$

where I_ρ is the ρ-dimensional identity matrix.

Proof. [Proof of Lemma 3.2] Recall that $\underline{\Lambda}$ is the $\rho \times \rho$ diagonal matrix of non-zero eigenvalues of M, so $\Lambda = \begin{bmatrix} \underline{\Lambda} & 0 \\ 0 & 0 \end{bmatrix}$.

Using the Rayleigh quotient inequality combined with Poincaré inequality (Theorem 4.2.2., Corr. 4.3.16 in [12]) we have the following bounds: $\forall x \in \mathbb{R}^\rho$,

$$\begin{aligned} x^TP^T(P\underline{\Lambda}P^T)^{\pm 1}Px &= x^TP^T(PP^T)^{-\frac{1}{2}}[(PP^T)^{\pm\frac{1}{2}}P\underline{\Lambda} \cdot P^T(PP^T)^{\pm\frac{1}{2}}]^{\pm 1}(PP^T)^{-\frac{1}{2}}Px \\ &\leq x^TP^T(PP^T)^{\pm 1}Px \cdot \|\underline{\Lambda}^{\pm 1}\| \end{aligned}$$

Likewise,

$$x^TP^T(P\underline{\Lambda}P^T)^{\pm 1}Px \geq x^TP^T(PP^T)^{\pm 1}Px \cdot \|\underline{\Lambda}^{\mp 1}\|.$$

In consequence, for the upper diagonal block we have:

$$\begin{aligned} &\|A_1\mathrm{E}[P^T(P\underline{\Lambda}P^T)^{\pm 1}P]^{-\frac{1}{2}} \cdot P^T(P\underline{\Lambda}P^T)^{\pm 1}P \cdot \mathrm{E}[P^T(P\underline{\Lambda}P^T)^{\pm 1}P]^{-\frac{1}{2}}A_1\| \\ &\leq \|\mathrm{E}[P^T(P\underline{\Lambda}P^T)^{\pm 1}P]^{-1}\| \cdot \|A_1P^T(PP^T)^{\pm 1}PA_1\| \cdot \|\underline{\Lambda}^{\pm 1}\| \\ &\leq \|\mathrm{E}[P^T(PP^T)^{\pm 1}P]^{-1}\| \cdot \|A_1P^T(PP^T)^{\pm 1}PA_1\| \cdot \|\underline{\Lambda}^{\pm 1}\| \cdot \|\underline{\Lambda}^{\mp 1}\| \\ &= \kappa(\underline{\Lambda})\|A_1V_{(I_\rho)}A_1\| \end{aligned}$$

In the last line we used the fact that $\mathrm{E}[P^T(PP^T)^{\pm 1}P]$ is a spherical matrix.

By the same arguments, for the lower diagonal block we have:

$$\begin{aligned}\|A_2\mathrm{E}[S^T(P\underline{\Lambda}P^T)^{\pm 1}S]^{-\frac{1}{2}} \cdot S^T(P\underline{\Lambda}P^T)^{\pm 1}S \cdot \mathrm{E}[S^T(P\underline{\Lambda}P^T)^{\pm 1}S]^{-\frac{1}{2}}A_2\| \\ \leq \|\mathrm{E}[S^T(PP^T)^{\pm 1}S]^{-1}\| \cdot \|A_2S^T(PP^T)^{\pm 1}SA_2\| \cdot \|\underline{\Lambda}^{\pm 1}\| \cdot \|\underline{\Lambda}^{\mp 1}\| \\ = \kappa(\underline{\Lambda})\|A_2W_{(I_\rho)}A_2\|\end{aligned}$$

since $\mathrm{E}[S^T(PP^T)^{\pm 1}S]$ is also spherical. This concludes the proof. □

The advantage of Lemma 3.2 is that it allows us to work with a matrix expectation that has a closed form. Let $M_0 \equiv \begin{bmatrix} I_\rho & 0 \\ 0 & 0 \end{bmatrix}$. The expectation that appears in $\mathrm{cov}_k^-(M_0)$ has a closed form expression (this would not be the case with a generic diagonal matrix argument). In addition, the expectations in both approximators $\mathrm{cov}_k^\pm(M_0)$ are spherical matrices. This will come in handy in the proof, as well as in the remaining analysis. In particular, straightforward computation gives:

$$\mathrm{E}[V_{(I_\rho)}^+] = \sigma^4 k(\rho + k + 1)I_\rho; \quad \mathrm{E}[W_{(I_\rho)}^+] = \sigma^4 \rho k I_{\bar{\rho}}; \quad \mathrm{E}[Z_{(I_\rho)}^+] = 0 \tag{3.14}$$

$$\mathrm{E}[V_{(I_\rho)}^-] = \frac{k}{\rho}I_\rho; \quad \mathrm{E}[W_{(I_\rho)}^-] = \frac{k}{\rho - k - 1}I_{\bar{\rho}}; \quad \mathrm{E}[Z_{(I_\rho)}^-] = 0 \tag{3.15}$$

Lemma 3.2 implies that, if $V_{(I_\rho)}^\pm$ satisfies the MSR with $\eta, c_{MSR} > 0$, then $V_{(\underline{\Lambda})}^\pm$ satisfies the MSR with η and $c_{MSR} \cdot (\kappa(\underline{\Lambda}))^{1+\eta}$. Likewise, if $W_{(I_\rho)}^\pm$ satisfies the MSR with $\eta', c'_{MSR} > 0$, then $W_{(\underline{\Lambda})}^\pm$ satisfies the MSR with η' and $c'_{MSR} \cdot (\kappa(\underline{\Lambda}))^{1+\eta'}$.

In the sequel, with the choice $M := M_0$ we will omit the lower index (I_ρ) from our notations of $U^\pm, V^\pm, W^\pm, Z^\pm$.

The remaining sections complete the analysis separately for $\hat{\mathrm{cov}}_k$ and $\hat{\mathrm{cov}}_k^-$ respectively. As the sign associated with the approximator will be clear from the section titles, we may omit the upper indexes.

4. Main Result for $\hat{\mathrm{cov}}_k$

Theorem 4.1 (Sufficient ensemble size for $\hat{\mathrm{cov}}_k$). *Let M be a $d \times d$ rank $\rho < d$ positive semi-definite matrix having $\kappa(M) \equiv \lambda_{\max}(M)/\lambda_\rho(M)$ bounded above independently of d. For any $\epsilon \in (0,1)$, and any choice of $\eta > 0$, there exist constants $c > 0$, and $C_1(c,\eta) > 0$, such that, taking an ensemble of size $m \geq C_1(c,\eta) \cdot \frac{d}{\epsilon^{2+2/\eta}}$ ensures that:*

$$E_R\left[\|\hat{cov}_k(M) - cov_k(M)\|\right] \leq \epsilon \cdot \|cov_k(M)\| \tag{4.1}$$

Proof. [Proof of Theorem 4.1] We shall make use of the notations and results developed so far. We work with the isotropic transformation of the generic term of $\hat{\mathrm{cov}}_k$, which is $U_{(M)}^+$ (defined earlier in eq. (3.1)), and for the rest of this subsection we omit the upper index. The plan is to show that this matrix satisfies the

MSR condition with some $c, \eta > 0$. This will then imply, by Theorem 2.1, that $\mathrm{E}_R\left[\|\frac{1}{m}\sum_{i=1}^{M} U_{i,(M)} - I_d\|\right] \leq \epsilon$, and rearranging gives the form stated in eq. (4.1).

To this end, by Lemmas 3.1 and 3.2, it is sufficient to show MSR for the simpler matrix $U \equiv U_{(I_\rho)}$. Take any projection matrix in $\mathbb{R}^d$, and $t \geq c \cdot r$ where r is the rank of A. By Lemma 2.1,

$$\Pr\{\|AUA\| \geq t\} \leq \Pr\{\|A_1 V A_1\| + \|A_2 W A_2\| \geq t\} \tag{4.2}$$

$$\leq \Pr\{\|A_1 V A_1\| \geq t/2\} + \Pr\{\|A_2 W A_2\| \geq t/2\} \tag{4.3}$$

where A_1 and A_2 are rank r projection matrices in $\mathbb{R}^\rho$ and $\mathbb{R}^{\bar{\rho}}$ respectively.

It now remains to show that both terms satisfy MSR. Lemma 4.1 below shows this is indeed the case, for any choices of $\eta, \eta' > 0$. Hence, by Lemma 2.1, U satisfies MSR. This completes the proof of the theorem. □

Lemma 4.1. *For any choices of $\eta, \eta' > 0$, there exist constants c_{MSR}, c'_{MSR} independently of d s.t. W and V from eq. (4.3) satisfy the MSR condition with (c_{MSR}, η), and (c'_{MSR}, η') respectively.*

Proof. [Proof of Lemma 4.1] Recall, by eq. (3.14) we have:

$$\mathrm{E}[R^T(RM_0R^T)R] = \sigma^4 \begin{bmatrix} (k^2+k+\rho k)I_\rho & 0 \\ 0 & \rho k I_{\bar{\rho}} \end{bmatrix} \tag{4.4}$$

Therefore,

$$U = \sigma^{-4} \begin{bmatrix} P^T(PP^T)P \cdot \frac{1}{k(\rho+k+1)} & P^T(PP^T)S \cdot \frac{1}{k\sqrt{\rho(\rho+k+1)}} \\ S^T(PP^T)P \cdot \frac{1}{k\sqrt{\rho(\rho+k+1)}} & S^T(PP^T)S \cdot \frac{1}{\rho k} \end{bmatrix} \tag{4.5}$$

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

We need a polynomially decaying upper bounds on the following tail probabilities, corresponding to MSR of the block-diagonal sub-matrices V and W respectively.

$$\Pr\left\{\|A_1 P^T(PP^T)PA_1\| \cdot \frac{1}{\sigma^4 k(\rho+k+1)} \geq t\right\} \leq ? \tag{4.6}$$

$$\Pr\left\{\|A_2 S^T(PP^T)SA_2\| \cdot \frac{1}{\sigma^4 \rho k} \geq t\right\} \leq ? \tag{4.7}$$

where A_1, A_2 are rank-r projection matrices. Both V and W are isotropic, as we have $\mathrm{E}[V] = I_\rho$, and $\mathrm{E}[W] = I_{\bar{\rho}}$.

We start with W, as this is the matrix whose dimensions $\bar{\rho} \times \bar{\rho} = (d-\rho) \times (d-\rho)$ depend on d. A_2 is a projection matrix in $\mathbb{R}^{\bar{\rho}}$, of rank $r \in \{1, ..., \bar{\rho}\}$, so it must have the form $A_2 = B_2^T(B_2B_2^T)^{-1}B_2$ for some full row-rank matrix B_2 of size $r \times \bar{\rho}$. Therefore we can absorb A_2 into S:

$$\|A_2 S^T(PP^T)SA_2\| = \|(B_2B_2^T)^{-1/2}B_2S^T(PP^T)SB_2^T(B_2B_2^T)^{-1/2}\|$$

$$= \|Q^T(PP^T)Q\| \tag{4.8}$$

where $Q := SB_2^T(B_2B_2^T)^{-1/2}$, and since S has i.i.d. Gaussian entries, the $k \times r$ matrix Q also has i.i.d. Gaussian entries with unchanged variance. We have:

$$\Pr\left\{\|Q^T(PP^T)Q\| \cdot \frac{1}{\sigma^4\rho k} \geq t\right\} \leq \Pr\left\{\sum_{i=1}^{\rho} P_iQQ^TP_i \geq \sigma^4\rho kt\right\} \tag{4.9}$$

$$\leq \rho \cdot \Pr\left\{\|Q^Tp\|^2 \geq \sigma^4 kt\right\} \tag{4.10}$$

where p has the distribution of a generic column of P.

Clearly, the vector Q^Tp has dependent entries. However, notice that its distribution $Q^Tp \sim GAL_r(2I_r, 0, k/2)$ follows a generalized Laplace distribution, of the form described in [23,17], and that it has covariance matrix is $\mathrm{E}[Q^Tpp^TQ] = \sigma^4 kI_r$. Therefore, $Q^Tp/\sigma^2\sqrt{k}$ is isotropic log-concave; this can be checked e.g. by checking that the negative log of the density function is convex. A simpler alternative is to deduce it from the equivalent representation given in [23], together with known properties of log-concave distributions: Take another, r-dimensional standard Gaussian vector $z \sim \mathcal{N}(0, I_r)$ independent of p; then the distribution of Q^Tp is the same as that of $\|p\| \cdot z$. Both the chi-distribution and the Gaussian belong to the log-concave family of distributions, and the product of independent log-concave distributions is also log-concave [27].

Therefore, by Theorem 2.2, there exists a constant $c > 0$ s.t. for any $\epsilon \geq 1$ we have:

$$\Pr\left\{\|Q^Tp\|^2/(\sigma^4 k) \geq c^2\epsilon^2 r\right\} \leq \exp(-\epsilon\sqrt{r}) \tag{4.11}$$

Rearranging, this is equivalent to the r.h.s. of eq. (4.10) being bounded as:

$$\rho \cdot \Pr\left\{\|Q^Tp\|^2 \geq \sigma^4 kt\right\} \leq \rho\exp(-\sqrt{t}/c) \tag{4.12}$$

independently of r, for any $t \geq c^2 r$, where c is the absolute constant from Paouris' inequality.

This exponential inequality is stronger than the polynomial inequality that we need to conclude the MSR condition of the random matrix $S^T(PP^T)S$. It implies the existence of the required constant c_{MSR} for any choice of $\eta > 0$, so the MSR holds.

Moving on to the upper diagonal block matrix, A_1 is a projection matrix, so $\|A_1\| = 1$, and we have:

$$\Pr\left\{\|A_1(P^TP)^2A_1\| \geq \sigma^4 k(\rho + k + 1)t\right\}$$

$$\leq \Pr\left\{\sum_{i=1}^{\rho} \|P_i\|^2 \geq \sigma^2\sqrt{k(\rho + k + 1)t}\right\} \tag{4.13}$$

As the columns P_i are statistically independent, $\sum_{i=1}^{\rho} \|P_i\|^2 \sim \chi^2(\rho k)$ is chi-square

distributed with ρk degrees of freedom. Applying the Chernoff bound we get:

$$\begin{aligned}\text{eq. (4.13)} &\leq \left(\frac{\sigma^2\sqrt{k(\rho+k+1)t}}{\rho k}\right)^{\frac{\rho k}{2}} \exp\left(-\frac{\sqrt{t}}{2}\sigma^2\sqrt{k(\rho+k+1)}-\rho k\right)\\ &\leq \exp\left(-\frac{1}{8}\left[\sigma^2\sqrt{\frac{k(\rho+k+1)t}{\rho k}}-\sqrt{\rho k}\right]^2\right)\end{aligned}$$

The last line applied the inequality $\log(1+\epsilon) \leq \epsilon - \epsilon^2/2$. The exponential factor dominates, and since $k\rho$ is constant, there exist $c'_{MSR}, \eta' > 0$ constants s.t. for $t \geq c'_{MSR} r$ the r.h.s. be bounded by $c'_{MSR} t^{-(1+\eta')}$. □

5. Main Result for $\hat{\text{cov}}_k^-(M)$

We shall now deal with the inverse covariance, or precision matrix approximator.

Theorem 5.1 (Sufficient ensemble size for $\hat{\text{cov}}_k^-$). *Let M be a $d \times d$ rank $\rho < d$ positive semi-definite matrix having $\kappa(M) \equiv \lambda_{\max}(M)/\lambda_\rho(M)$ bounded above independently of d. Suppose that $3 \leq \rho < d$, $\rho > k+1$. For any $\epsilon \in (0,1)$, and any choice of $\eta > 0$, there exist constants $a, c > 0$, and $C_2(c,\eta) > 0$, such that, if $\rho - k + 1 > a\log(\bar{\rho}) + a$, then taking an ensemble of size $m \geq C_2(c,\eta) \cdot \frac{d}{\epsilon^{2+2/\eta}}$ ensures that:*

$$E_R\left[\|\hat{cov}_k^-(M) - cov_k^-(M)\|\right] \leq \epsilon \cdot \|cov_k^-(M)\| \tag{5.1}$$

The high level proof strategy is similar to that seen before, however the two sub-matrices involved in the analysis will turn out to belong to different families of matrix distributions, and consequently require different treatment. In particular, the sub-matrix that corresponds to the null-space of M is outside the log-concave family of distributions, and instead it follows a matrix-variate T distribution. We will employ the bounds we developed in Section 2.3 to show that it satisfies the MSR condition under the stated assumptions.

Proof. [Proof of Theorem 5.1] The isotropic transformation of the generic term of $\hat{\text{cov}}_k^-$ is $U^-_{(M)}$, defined in eq. (3.1). Next, we shall establish the MSR condition for this matrix under the conditions in the theorem statement. Then, Theorem 2.1, and rearranging yield eq. (5.1) as required.

As before, by Lemmas 3.1 and 3.2, it is sufficient to show MSR for the simpler matrix $U^- \equiv U^-_{(I_\rho)}$. To this end, take an arbitrary projection matrix in $\mathbb{R}^d$, and $t \geq c \cdot r$ where r is the rank of A. By Lemma 2.1, we have:

$$\begin{aligned}\Pr\left\{\|AU^-A\| \geq t\right\} &\leq \Pr\left\{\|A_1V^-A_1\| + \|A_2W^-A_2\| \geq t\right\} &(5.2)\\ &\leq \Pr\left\{\|A_1V^-A_1\| \geq t/2\right\} + \Pr\left\{\|A_2W^-A_2\| \geq t/2\right\} &(5.3)\end{aligned}$$

where A_1 and A_2 are rank r projection matrices in $\mathbb{R}^\rho$ and $\mathbb{R}^{\bar{\rho}}$ respectively.

Recalling the form of these matrices from eq. (3.15), we have:

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

Observe that the two diagonal blocks belong to different classes of matrix-valued distributions. The block V^- has all of 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. Also note that W^- is isotropic, as $\mathrm{E}[W^-] = I_{\bar{\rho}}$, but it is not log-concave.

Let us deal with the easy submatrix first. The matrix $V^- \equiv P^T(PP^T)^{-1}P \cdot \frac{\rho}{k}$ satisfies MSR trivially, with any choice of $\eta > 0$, e.g. by Proposition 8.5 in [29]. It remains to ensure that MSR holds for W^-.

As before, A_2 is a projection matrix in $\mathbb{R}^{\bar{\rho}}$, of rank $r \in \{1, ..., \bar{\rho}\}$, necessarily of the form $B_2^T(B_2B_2^T)^{-1}B_2$ for some full row-rank matrix B_2 of size $r \times \bar{\rho}$. We can absorb A_2 into S, since:

$$\begin{aligned} \|A_2S^T(PP^T)^{-1}SA_2\| &= \|(B_2B_2^T)^{-1/2}B_2S^T(PP^T)^{-1}SB_2^T(B_2B_2^T)^{-1/2}\| \\ &= \|Q^T(PP^T)^{-1}Q\| \end{aligned} \tag{5.4}$$

where $Q := SB_2^T(B_2B_2^T)^{-1/2}$, 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, $J := (PP^T)^{-1/2}Q$ has a matrix-variate T-distribution. In particular, $J \cdot \sqrt{(\nu-2)/k} \sim T_{k\times\bar{\rho}}(0, I_k, I_{\bar{\rho}}, \nu)$ with $\nu = \rho - k + 1$. Hence, as a consequence of Lemma 2.2, i.e. eq. (2.23) we have for all $t > r$:

$$\Pr\left\{\|A_2W^-A_2\| \geq t\right\} \leq k \cdot \left(\frac{t}{r}\right)^{\frac{r}{2}} \cdot \left(\frac{r+\nu}{t+\nu}\right)^{\frac{\nu+r}{2}} \leq k \cdot \left(\frac{t}{\bar{\rho}}\right)^{\frac{\bar{\rho}}{2}} \cdot \left(\frac{\bar{\rho}+\nu}{t+\nu}\right)^{\frac{\nu+\bar{\rho}}{2}} \tag{5.5}$$

Finally, for this to imply MSR, we need to require that the r.h.s. of eq. (5.5) is upper bounded by $c \cdot t^{-1-\eta}$ for some $c > 0$ constant. The following remark shows that this is indeed the case under the conditions in the theorem statement.

Remark 5.1. There exists a constant $c > 0$ for any choice of $\eta > 0$ s.t.

$$\left(\frac{t}{\bar{\rho}}\right)^{\frac{\bar{\rho}}{2}} \cdot \left(\frac{\bar{\rho}+\nu}{t+\nu}\right)^{\frac{\nu+\bar{\rho}}{2}} \leq c \cdot t^{-1-\eta},$$

provided that $\nu > a\log(\bar{\rho}) + a$ for some constant a.

The proof of this remark is due to [4] and is reproduced in the Appendix for completeness.

As k is fixed, it can be absorbed into c. Therefore, recalling that $\nu = \rho - k + 1$, and $\bar{\rho} = d - \rho$, a sufficient condition for MSR in our case is that $\rho - k + 1 \geq \Omega(\log(d-\rho))$. The proof is complete. □

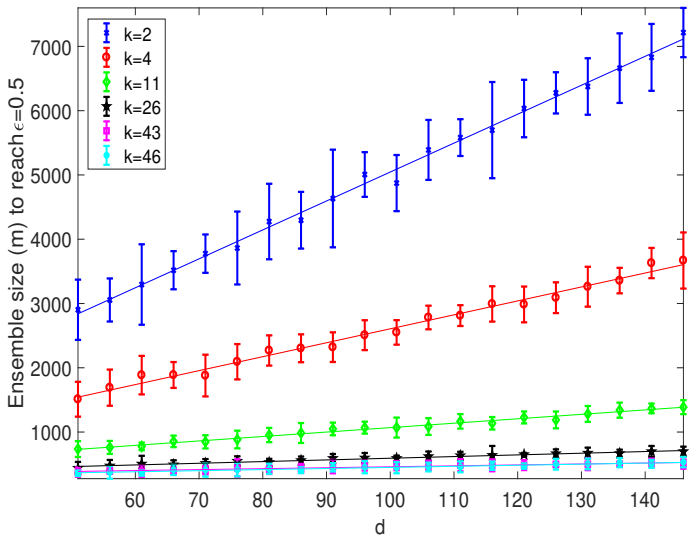

Fig. 1. Numerical experiment for the covariance approximator $\hat{\text{cov}}_k(M)$ where M has rank $\rho = 50$. We observe the required ensemble size grows linearly with d, for all choices of k tested.

6. Numerical demonstration

In this section we give an empirical demonstration of our findings, namely that the ensemble size m only needs to grow linearly with d for the random projection based finite ensemble covariance and inverse covariance approximators to become close to their expectations.

We took a fixed singular matrix M of rank $\rho = 50$, and generated independent random Gaussian matrices $R_1, R_2, ..., R_m$ (with variance of entries set to $1/k$) and computed the following:

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

We increased the ensemble size m progressively until ϵ reached below a pre-defined threshold. We varied d, and ran 15 independent repetitions of this experiment for several values of k.

Figures 1 and 2 present the ensemble sizes required, for $\hat{\text{cov}}_k$ and $\hat{\text{cov}}_k^-$ respectively. The ensemble sizes (m) on the vertical axis in these figures represent averages computed from the 15 independent repetitions, and the error bars depict the inter-quartile range. The best linear fits are also superimposed. We should note that the absolute magnitudes on the vertical axes are not comparable between the plots of $\hat{\text{cov}}_k$ and $\hat{\text{cov}}_k^-$, since the construction of these matrices is quite different.

We see from Figures 1 and 2, as expected from our theoretical results, that for each choice of k, the required ensemble size displays a growth that is linear in d.

These simulations suggest that our concentration bound is tight order-wise, and

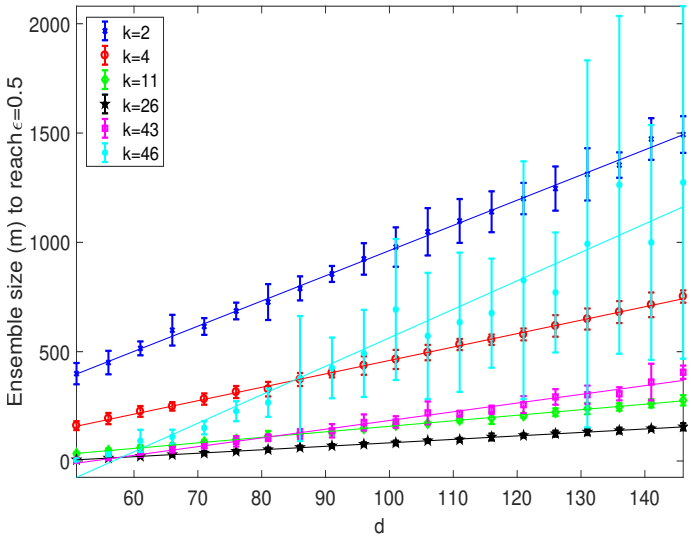

Fig. 2. Numerical experiment for the inverse covariance approximator $\hat{\text{cov}}_k^-(M)$, where M has rank $\rho = 50$. Again, we observe for all choices of k tested, the required ensemble size grows linearly with d.

provides useful guidance on setting the ensemble size in practice, especially when one is interested to monitor various performance metrics as dimension increases, to test the scalability of algorithms. Based on our results, in practice one may therefore set m to a constant multiple of d, where the choice of the magnitude of this constant depends on the available computational resources. Indeed, a constant as low as 1 or 2 was observed to work well in the continuous optimization application described in [16].

In addition, although not addressed by our theoretical analysis, the numerics suggest that a choice of k around the middle of its allowed range produces lower error and lower slope for $\hat{\text{cov}}_k^-$, in agreement with previous empirical experiences in classification [9]. Too low values of k induce high error due to over-smoothing, while too high values of k display large variability as the degree of freedom ν of the heavy tailed random sub-matrix becomes low, implying weaker concentration properties. On the other hand, for $\hat{\text{cov}}_k$, low values of k induce high error for the same reason of over-smoothing, but higher values of k remain stable. This was intuitively expected, as the matrix distribution is better behaved, no heavy-tailed sub-matrix is involved. In this case, the problem of setting of k in practice may also depend on the purpose, as higher values of k are computationally more expensive.

7. Conclusions and future work

We quantified the Monte Carlo error of two random matrix theory based approaches that deal with singular covariance matrices. From this we deduced the number of in-

dependent random projections that ensure that the finite ensemble gets sufficiently close to the associated matrix expectation. We found that the ensemble size only needs to grow linearly with the dimension of the positive semi-definite input matrix, both in the case of covariance approximation and in the case of inverse covariance approximation, in the latter case under mild assumptions.

Further work of interest includes the question of what is the optimal choice of k. As already noted in [22], this is difficult to answer. For instance, if the singular covariance M is a low-sample estimate of a non-singular true covariance, the approximation error between the latter and the converged ensemble (as considered in [22]), the Monte Carlo error of the finite ensemble (as considered in this work), and the availability of computational resources in practice all contribute to this choice. Nevertheless, a better understanding of each component offers some guidance.

Another worthwhile avenue for further research is to extend the analysis to non-Gaussian random projections. In particular, we can show that the multivariate t-distribution with ν degrees of freedom belongs to the family of $-1/\nu$-concave distributions (definitions may be found in e.g. [5,6]), which suggests this may be feasible.

Acknowledgements

We thank Olivier Guédon and Bob Durrant for an insightful discussion at the Institut Henri Poincaré, which sparked this work. Thanks go also to the MathOverflow community for the contribution [4], and the anonymous referees for their suggestions that improved the presentation. This work is funded by EPSRC under Fellowship grant EP/P004245/1.

References

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

[2] R. Adamczak, R. Latala, A.E. Litvak, K. Oleszkiewicz, A. Pajor, N. Tomcsak-Jaegermann. A short proof of Paouris' inequality. Canadian Mathematical Bulletin, 57(1):3-8, 2014.

[3] R. Ahlswede, A. Winter, Strong converse for identification via quantum channels, IEEE Transactions on Information Theory 48:568-579, 2002.

[4] Y. Baruch. Answer to 'Implausible inequality?', MathOverflow (version: 2017-08-20), URL: https://mathoverflow.net/q/279128

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

[6] K. Chandrasekaran, A. Deshpande, S. Vempala. Sampling s-concave functions. Proceedings of the 13th International Workshop on Randomization and Computation, 5687:420-433, 2009.

[7] P.L. Combettes, S. Salzo, S. Villa. Regularized learning schemes in feature Banach spaces. Analysis and Applications. 16(1): 1-54, 2018.

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

[9] 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.
[10] Z-C. Guo, D-H. Xiang, X. Guo, D-X. Zhou. Thresholded spectral algorithms for sparse approximations. Analysis and Applications 15(3):433-455, 2017.
[11] A.K. Gupta, D.K. Nagar. Matrix Variate Distributions. CRC Press, 1999.
[12] R. Horn, C. Johnson. Matrix Analysis, Cambridge Univ. Press, 1985.
[13] T. Ideker, R. Sharan. Protein networks in disease. Genome research 18(4):644-652, 2008.
[14] Joriki. Largest eigenvalue of a positive semi-definite matrix is less than or equal to sum of largest eigenvalues of its diagonal blocks, Mathematics StackExchange, URL (version: 2012-05-14): http://math.stackexchange.com/q/144963
[15] A. Kabán. On compressive ensemble induced regularization: How close is the finite ensemble precision matrix to the infinite ensemble? Proceedings of the 28th International Conference on Algorithmic Learning Theory (ALT 2017), pp. 617-628, 2017.
[16] A. Kabán, J. Bootkrajang, R.J Durrant. Toward large-scale continuous EDA: A random matrix theory perspective, Evolutionary Computation, 24:2, pp. 255-291, 2016.
[17] S. Kotz, T.J. Kozubowski, K. Podgorski. The Laplace Distribution and Generalizations. Birkhauser. pp.229–245. ISBN 0817641661, 2001.
[18] N. Krämer, J. Schäfer, A. Boulesteix. Regularized estimation of large-scale gene association networks using graphical Gaussian models. BMC Bioinformatics 10:384, 2009.
[19] P. Larrañaga and J. A. Lozano. Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation. Kluwer, 2002.
[20] O. Ledoit, M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88(2):365-411, 2004.
[21] P-E. Lin. Some characterisations of the multivariate t distribution. Journal of Multivariate Analysis 2:339-344, 1972.
[22] T. Marzetta, G. Tucci, S. Simon. A random matrix theoretic approach to handling singular covariance estimates. IEEE Transactions on Information Theory 57(9):6256-6271, 2011.
[23] P-A. Mattei. Multiplying a Gaussian matrix by a Gaussian vector. Statistics and Probability Letters 128: 67-70, 2017.
[24] N. Meinshausen, P. Bühlmann. High-dimensional graphs and variable selection with the lasso. Annals of Statistics 1436-1462, 2006.
[25] G. Paouris. Concentration of mass on convex bodies. Geometric Functional Analysis, 16:1021-1049, 2006.
[26] M. Rudelson. Random vectors in the isotropic position. Journal of Functional Analysis 164(1):60-72, 1999.
[27] A. Saumard, J. A. Wellner. Log-concavity and strong log-concavity: A review. Statistics Surveys 8:45-114, 2014.
[28] G.A. Thanei, C. Heinze, N. Meinshausen. Random Projections for Large-Scale Regression. In: Ahmed S. (eds) Big and Complex Data Analysis. Contributions to Statistics. Springer, Cham, 2017.
[29] P. Youssef. Estimating the covariance of random matrices. Electronic Journal of Probability 18:107, 2013.
[30] M. Yuan, Y. Lin. Model selection and estimation in the Gaussian graphical model. Biometrika, 19-35, 2007.

Appendix

The following proof of Remark 5.1 is reproduced from MathOverflow [4]. Let $t, r, \nu \geq 1$, of which $\bar{\rho}$ can be unbounded, $\nu = \rho - k + 1$ can be constrained to be larger than some threshold that may depend on $\bar{\rho}$ in some mild way, such as logarithmically. We need to show that: $\exists c, \eta > 0$ constants independent of $\bar{\rho}$ s.t. $\forall t > \bar{\rho} \cdot c$,

$$L(t) \stackrel{\text{def}}{=} \left(\frac{t}{\bar{\rho}}\right)^{\bar{\rho}/2} \left(\frac{\bar{\rho}+\nu}{t+\nu}\right)^{(\bar{\rho}+\nu)/2} \cdot c^{-1} \cdot t^{-1-\eta} \stackrel{?}{\leq} 1$$

In fact, this will be shown to hold for *any* choice of $\eta > 0$, for some constants c and a, if $\nu > a \log(\bar{\rho}) + a$.

Taking logarithms and derivatives w.r.t. t, the unique maximum of $L(t)$ is found at $t_{\max} = \frac{\nu(\bar{\rho}+2+2\eta)}{\nu-2-2\eta}$. If $a \geq 3 + 2\eta$ then $\frac{t_{\max}}{\bar{\rho}} \leq (3+2\eta)^2$; so $t_{\max}$ falls outside the range $t > c\bar{\rho}$ if $c > (3+2\eta)^2$, and one only needs to verify $L(t) \leq 1$ for $t = c\bar{\rho}$:

$$L(c\bar{\rho}) = c^{\bar{\rho}/2} \cdot \left(\frac{\bar{\rho}+\nu}{c\bar{\rho}+\nu}\right)^{\bar{\rho}/2+\nu/2} \cdot c^{\eta}\bar{\rho}^{1+\eta} \stackrel{?}{\leq} 1$$

Case 1.
We use of the following inequality: $(1 - \frac{y}{x+y})^x < \sqrt{3}^{-y}$ if $x \geq y/2 > 0$. Therefore: $\left(\frac{\bar{\rho}+\nu}{c\bar{\rho}+\nu}\right)^{\bar{\rho}/2+\nu/2} = \left(1 - \frac{(c-1)\bar{\rho}/2}{c\bar{\rho}/2+\nu/2}\right)^{\bar{\rho}/2+\nu/2} < \sqrt{3}^{-(c-1)\bar{\rho}/2}$, and hence:
$L(c\bar{\rho}) = c^{\bar{\rho}/2} \cdot \left(\frac{\bar{\rho}+\nu}{c\bar{\rho}+\nu}\right)^{\bar{\rho}/2+\nu/2} \cdot (c\bar{\rho})^{1+\eta} c^{-1} < c^{\bar{\rho}/2} \cdot \sqrt{3}^{-(c-1)\bar{\rho}/2} \cdot c^{\eta}\bar{\rho}^{1+\eta}$

Noting that $\frac{c}{\sqrt{3}^{c-1}} \to 0$ as $c \to \infty$, so given η, the inequality holds for every $\bar{\rho}$ if c is large enough, provided $x \geq y/2$, that is $\nu \geq (c-3)\bar{\rho}/2$.

Case 2.
Now assume $a \log(\bar{\rho}) + a < \nu < (c-3)\bar{\rho}/2$ for some a. Rewrite:

$$L = c^{-\nu/2} \cdot \left(\frac{\bar{\rho}+\nu}{\bar{\rho}+\nu/c}\right)^{\bar{\rho}/2+\nu/2} \cdot c^{\eta}\bar{\rho}^{1+\eta} \stackrel{?}{\leq} 1$$

Now, we make use of the inequality: $(1 + \frac{x}{y})^y < e^x$ if $x, y > 0$. Therefore:

$$\begin{aligned}
\left(\frac{\bar{\rho}+\nu}{\bar{\rho}+\nu/c}\right)^{\bar{\rho}/2+\nu/2} &= \left(\frac{\bar{\rho}+\nu}{\bar{\rho}+\nu/c}\right)^{\bar{\rho}/2+\nu/(2c)} \cdot \left(\frac{\bar{\rho}+\nu}{\bar{\rho}+\nu/c}\right)^{(1-1/c)\nu/2} \\
&= \left(1 + \frac{(1-1/c)\nu/2}{\bar{\rho}/2+\nu/(2c)}\right)^{\bar{\rho}/2+\nu/(2c)} \cdot \left(\frac{\bar{\rho}+\nu}{\bar{\rho}+\nu/c}\right)^{(1-1/c)\nu/2} \\
&< e^{(1-1/c)\nu/2} \cdot \left(\frac{\bar{\rho}+\nu}{\bar{\rho}+\nu/c}\right)^{(1-1/c)\nu/2}
\end{aligned}$$

Now it suffices to prove:
$c^{-\nu/2} \cdot \left(e \cdot \frac{\bar{\rho}+\nu}{\bar{\rho}+\nu/c}\right)^{(1-1/c)\nu/2} \cdot c^{\eta}\bar{\rho}^{1+\eta} \stackrel{?}{\leq} 1.$
Raising to the power of $2/\nu$, and rearranging, yields:
$e^{1-1/c} \cdot \frac{\bar{\rho}+\nu}{c\bar{\rho}+\nu} \cdot \left(\frac{\bar{\rho}+\nu/c}{\bar{\rho}+\nu}\right)^{1/c} \cdot c^{\eta 2/\nu} \cdot \bar{\rho}^{(2+2\eta)/\nu} \stackrel{?}{\leq} 1.$

Since $\nu < (c-3)\bar{\rho}/2 \implies \frac{\bar{\rho}+\nu}{c\bar{\rho}+\nu} < 1/3$, and since $\frac{\bar{\rho}+\nu/c}{\bar{\rho}+\nu} < 1$, it is sufficient to prove $\frac{e^{1-1/c}}{3} \cdot c^{2\eta/\nu} \cdot \bar{\rho}^{(2+2\eta)/\nu} \stackrel{?}{\leq} 1$

Finally, $\nu > a \log(\bar{\rho}) + a \implies \bar{\rho}^{(2+2\eta)/\nu} < e^{(2+2\eta)/a}$. One can now choose a such that $\frac{e}{3} \cdot c^{2\eta/a} \cdot e^{(2+2\eta)/a} < 1$.