I show analytically that the average of \( k \) bootstrapped correlation matrices rapidly becomes positive-definite as \( k \) increases, which provides a simple approach to regularize singular Pearson correlation matrices. If \( n \) is the number of objects and \( t \) the number of features, the averaged correlation matrix is almost surely positive-definite if \( k > e^{-1} \frac{n}{t} \simeq 1.58 \frac{n}{t} \) in the limit of large \( t \) and \( n \). The probability of obtaining a positive-definite correlation matrix with \( k \) bootstraps is also derived for finite \( n \) and \( t \). Finally, I demonstrate that the number of required bootstraps is always smaller than \( n \). This method is particularly relevant in fields where \( n \) is orders of magnitude larger than the size of data points \( t \), e.g., in finance, genetics, social science, or image processing.

**Keywords** Correlation · Regularization · High-Dimensionality

1 Introduction

Correlation and covariance matrices are fundamental dependence estimators in statistical inference. Their use includes risk minimization in finance [1], analysis of functional genomics [2], or image processing [3]. However, when the number of objects under study (\( n \)) exceeds the number of available data points (\( t \)), these matrices cannot be inverted. As a result, many standard inference methods cannot be applied directly. To overcome this issue, a large literature on eigenvalue regularization has been devoted to this issue over the last decades. The most relevant ones are the Ledoit-Wolf linear shrinkage [4] and the more recent non-linear shrinkage [5]. These methods, apart from regularizing singular correlation matrices, attempt to reduce the noise effect due to finite sample size. In addition, Ref. [6] proposes a recursive algorithm that aims to find the most similar positive-definite matrix to an initial problematic matrix that is not positive-definite. Similarly to the proposed method, this approach does not try to denoise the target matrix but corrects the eigenvalue distribution by removing the non-positive eigenvalues.

In this work, I propose a simple alternative approach based on bootstrap resampling to regularize correlation matrices with \( z > 0 \) zero degenerate eigenvalues. In particular, I prove that the probability to obtain a positive defined matrix from the average of \( k \) bootstrap resampling scenarios converges rapidly with respect to \( k \) to one provided that \( k \) is larger than \( e^{-1} \frac{n}{t} \).

2 The Bootstrap Average Correlation Matrix

Let \( X \in \mathbb{R}^{n \times t} \) be the data matrix and \( C \in \mathbb{R}^{n \times n} \) its Pearson correlation matrix. We assume that no column or row of \( X \) is a linear combination of the others; this implies that \( C \) has rank \( r = \min\{n, t - 1\} \). Let \( X^{(b)} \in \mathbb{R}^{n \times t} \) be a bootstrap copy of \( X \) obtained by sample replacement of the columns of \( X \), and \( C^{(b)} \) its correlation matrix. A generic element of \( X^{(b)} \) is \( x^{(b)}_{ij} = x_{ih^{(b)}} \), where \( h^{(b)} \) is a vector of dimension \( t \) obtained by random sampling with replacement of the elements of vector \( (1, 2, \cdots, t) \).
This paper derives an approximate expression of the probability that the smallest eigenvalue $\lambda_0$ of the correlation matrix $\langle C \rangle := k^{-1} \sum_{i=0}^{k} C^{(i)}$ is larger than zero as a function of the number of bootstrap copies. The minimum number of bootstrap copies $k^+$, that guarantees $\langle C \rangle$ to be positive-definite within a chosen confidence level, shows a real transition in the large-system limit, defined here as $n, t \to \infty$ at fixed $q$.

### 3 The Distribution of the Number of Null Eigenvalues

The first step is to obtain a probability distribution of the number of zero eigenvalues $z_b$ of a given bootstrap correlation matrix $C^{(b)}$. One has

$$z_b = \max\{n + 1 - u_b, 0\}, \tag{1}$$

where $u_b$ is the number of unique column indices sampled from $X$ in the $b$-bootstrap copy. The exact probability distribution of $u_b$ is known to be \[7\]

$$P(u_b) = \frac{S_2(t, u_b) t!}{t^t (t-u_b)!} \tag{2}$$

where $S_2(t, u_b)$ is the Sterling number of the second kind. Such a distribution has mean and variance

$$\mu(t) = t \left[ 1 - \left( 1 - \frac{1}{t} \right)^t \right] \tag{3}$$

$$\sigma^2(t) = t \left( 1 - \frac{1}{t} \right)^t + t^2 \left( 1 - \frac{1}{t} \right) \left( 1 - \frac{2}{t} \right) - t^2 \left( 1 - \frac{1}{t} \right)^{2t}. \tag{4}$$

In the limit of large $t$, eqs (3) become

$$\mu(t) \approx \left( 1 - \frac{1}{e} \right) t + \frac{1}{2e}$$

$$\sigma^2(t) \approx \left( \frac{e - 2}{e^2} \right) t + \frac{3 - e}{2e^2}. \tag{4}$$

Furthermore, it is worth noticing that the deviation of the empirical $P(u_b)$ from a normal $\mathcal{N}(\mu(t), \sigma(t))$ is negligible for even for moderately large $t$ \[7\], as reported in the right-hand side plot of Fig. [4]

If we consider a condition characterized by an abundance of expected zero eigenvalues, i.e., $n \gg t$, then the probability distribution of $z_b$ according to Eq. (1) can be approximate by a Normal distribution

$$P(z_b) \approx \mathcal{N}(n + 1 - \mu(t), \sigma(t)). \tag{5}$$

Now that the distribution of the zero eigenvalues for the single bootstrap copy is known, we can answer the original question, and consider $k$ bootstrap copies of $X$ such that $\langle C \rangle := k^{-1} \sum_{i=1}^{k} C^{(i)}$. 

\[\text{Figure 1:} \text{ The left plot shows the exact (Eq. (3)) and approximate (eq. (4)) $t$-dependence of the first two moments of $P(u_b)$ distribution of Eq. (2). The right plot shows the approximate Normal Cumulative Distribution Function (CDF) against the observed CDF obtained with $10^4$ random sampling for every integer value of $u_b \in [0, t]$.}\]
A bi-dimensional mapping of the values of $k$ according to Eq. (5), the distribution of $X_t$ with the constraint to be orthogonal with $V_j$ such that $\dim(V_i \cap V_j) \geq 1$, then $w C^{(j)} w' > 0$; and thus $w (C) w' > 0$ for every vector $w$ that lies in $V_i$ or $V_j$.

It is important to point out that eigenvectors associated to $z_i$ zero eigenvalues can be assumed to be “randomly” chosen with the constraint to be orthogonal with $V_j^\perp$, the space defined by the eigenvectors associated with the $n - z_i$ non-zero eigenvalues; this because they do not carry any information about the correlation matrix $C^{(i)}$ since they explain zero variance. Therefore every rotation of the basis of $V_i$ constrained to be orthogonal with $V_j^\perp$ will produce exactly the same matrix $C^{(i)}$. In the $k = 2$ case, the probability that $\dim(V_1 \cap V_2) \geq 1$ will be approximately 1 if $z_1 + z_2 \leq n$ and 0 otherwise. It is possible to visualize this relationship easily in a three-dimensional space, i.e., $n = 3$. In case of two random straight lines, that have dimensions $z_1 = 1$ and $z_2 = 1$, the probability that they intersect in a straight line is almost zero since they must be coincident; differently, if we consider two random planes $z_1 = 2$ and $z_2 = 2$ they will intersect in a straight line almost surely apart from only configurations in which they are parallels. The above-discussed approximation, in the case of the spectral decomposition, is valid if the probability that the orthogonal spaces $V_i^\perp$ and $V_j^\perp$ defined from the $n - z_1$ and $n - z_2$ non-zero eigenvalues perfectly overlap is negligible. In a bootstrap resampling, when $t$ is sufficiently large, this probability is approximately zero, as this requires to sample the same column indices of $X$ for both bootstrap realizations, in other words, $C^{(1)} = C^{(2)}$.

More generally, for $k$ bootstrap copies, every hyper-plane $V_i$ will verify $\dim(V_i \cap V_j) \geq 1$ for at least one $j \neq i$ with probability 1 if

$$\zeta := \sum_{i=1}^k z_i \leq (k - 1) n.$$  

If the above inequality holds, then $\langle C \rangle$ has no zero eigenvalue. From Eq. (6), one can derive an upper bound for the number of bootstrap copies required. In fact, even if all bootstrap correlations have $n - 1$ null eigenvalues, no more than $k = n$ bootstrap copies are necessary to obtain a positive define matrix $\langle C \rangle$.

According to Eq. (5), the distribution of $\zeta$ can be approximated by a sum of $k$ identical normal distributions that converges to

$$\mathcal{P}(\zeta) \approx \Phi \left( \frac{\zeta}{\sqrt{k} \sigma(t)} \right).$$  

Therefore, the probability that the smallest eigenvalue $\lambda_0$ of $\langle C \rangle$ is larger than zero can be obtained from the cumulative distribution function of $\mathcal{P}(\zeta)$ estimated at $(k - 1) n$, that is

$$\mathcal{P}(\lambda_0 > 0) \approx \mathcal{P}(\zeta \leq (k - 1) n) = \int_{-\infty}^{(k-1)n} \mathcal{P}(\zeta) d\zeta \approx \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{\mu(t) - 1}{\sigma(t)} \right) \right].$$  

The above equation suggests to set a threshold $\alpha$ such that $\mathcal{P}(\lambda_0 > 0) > 1 - \alpha$, i.e., $1 - \text{erf}(\alpha) = \alpha$ (for example, $\alpha \approx 1.82$ for $\alpha = 0.01$). One can then define the number of bootstraps required to achieve $\mathcal{P}(\lambda_0 > 0) > 1 - \alpha$ by setting the argument of the erf function to $a$, which gives

$$k^+(a) \approx \frac{a^2 \sigma^2(t) + [\mu(t) - 1]n + \sqrt{a^4 \sigma^4(t) + 2a^2 \sigma^2(t) [\mu(t) - 1]n}}{[\mu(t) - 1]^2}.$$

A bi-dimensional mapping of the values of $k^+(a)$ with $a = 1.82$ as function of $n$ and $t$, shown in Fig. 2, left, shows that the number of bootstrap copies $k^+$ required to have a positive defined $\langle C \rangle$ is quite small, at least for not too extreme values of $q = n/t$.

To have a rough estimate of the transition point $k^+$ for $\mathcal{P}(\lambda_0 > 0) \approx 1$ in the limit of large $n$, we can substitute $t = n/q$, and compute the $k^+$ of the inflection point of the error function, obtained for the argument of erf equals to zero

$$k^+ = \frac{2enq}{2(e - 1)n + (1 - 2e)q}.$$
The large-system limit of the first derivative slope at the inflection point if the error function diverges to infinity

\[
\lim_{n \to \infty} \frac{d}{dk} \frac{k[2(e - 1)n - 2eq + q] - 2enq}{2q\sqrt{k \left[ \frac{2(e - 2)n}{q} - e + 3 \right]}} \bigg|_{k=k^*} = \infty. \tag{11}
\]

This means that \( \mathcal{P}(\lambda_0 > 0) \) has a real transition in the large-system limit. The value of the inflection point of Eq. (10), in the large-system limit, converges to

\[
\lim_{n \to \infty} k^* = \lim_{n \to \infty} k^+(a) = \frac{e}{(1 - e)}q \approx 1.58 \frac{n}{t}. \tag{12}
\]

The right-hand side of Fig. 2 shows that this approximation can provide a quite accurate estimation of the magnitude of \( k^+ \) even for \( n \) small when \( q \) is not extremely large.

In summary, both the approximate distribution of \( \mathcal{P}(\lambda_0 > 0) \) of Eq. (8) and the bound limits \( k^+ \) of Eqs. (9) show a very good agreement with the observations, reported in Fig. 3.
4 Discussion

I have shown that the average correlation matrix of $k$ bootstrap copies converges to a positive-defined matrix for $k$ much smaller than the order of the matrix. Such a matrix can be used in many applications which require to invert $C$, such as risk optimization. An extensive comparative analysis of the performance of these approaches will be addressed in future works.

Acknowledgements

I thank prof. Damien Challet for helpful support and discussions. This publication stems from a partnership between CentraleSupélec and BNP Paribas.

References

[1] Harry Markowitz. *Portfolio selection: Efficient diversification of investments*, volume 16. John Wiley New York, 1959.

[2] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[3] Santiago Velasco-Forero, Marcus Chen, Alvina Goh, and Sze Kim Pang. Comparative analysis of covariance matrix estimation for anomaly detection in hyperspectral images. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):1061–1073, 2015.

[4] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[5] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *The Review of Financial Studies*, 30(12):4349–4388, 2017.

[6] Nicholas J Higham. Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, 22(3):329–343, 2002.

[7] Alex F Mendelson, Maria A Zuluaga, Brian F Hutton, and Sébastien Ourselin. What is the distribution of the number of unique original items in a bootstrap sample? *arXiv preprint arXiv:1602.05822*, 2016.