Optimal covariance cleaning for heavy-tailed distributions: Insights from information theory

Christian Bongiorno [1,*] and Marco Berritta [2]

[1] *Université Paris-Saclay, CentraleSupélec, Laboratoire de Mathématiques et Informatique pour la Complexité et les Systèmes, 91192 Gif-sur-Yvette, France*

[2] *Department of Physics and Astronomy, University of Exeter, Stocker Road, Exeter EX4 4QL, United Kingdom*

(Received 27 April 2023; accepted 31 October 2023; published 27 November 2023)

In optimal covariance cleaning theory, minimizing the Frobenius norm between the true population covariance matrix and a rotational invariant estimator is a key step. This estimator can be obtained asymptotically for large covariance matrices, without knowledge of the true covariance matrix. In this study, we demonstrate that this minimization problem is equivalent to minimizing the loss of information between the true population covariance and the rotational invariant estimator for normal multivariate variables. However, for Student's t distributions, the minimal Frobenius norm does not necessarily minimize the information loss in finite-sized matrices. Nevertheless, such deviations vanish in the asymptotic regime of large matrices, which might extend the applicability of random matrix theory results to Student's t distributions. These distributions are characterized by heavy tails and are frequently encountered in real-world applications such as finance, turbulence, or nuclear physics. Therefore, our work establishes a connection between statistical random matrix theory and estimation theory in physics, which is predominantly based on information theory.

DOI: 10.1103/PhysRevE.108.054133

I. INTRODUCTION

In today's data-rich environment, the central role of multivariate analysis across various fields has become increasingly evident, driven by advancements in computational capabilities and data availability. As a result, there is a growing need to estimate large covariance matrices. However, this process becomes challenging when the number of variables n is large relative to the number of observations t, a phenomenon known as the "curse of dimensionality." In such cases, the covariance matrix becomes increasingly noisy and can even turn nonpositive definite when $n > t$. It is often the case that we must operate under conditions where n is approximately equal to t due to factors such as nonstationarity, placing intrinsic limits on our data collection capacity.

Several techniques have been proposed to refine covariance estimations in the "curse of dimensionality" regime. To grasp the rationale behind these techniques, it is essential to delve into foundational statistical theory. When considering a single pair of time series, the sample covariance estimator is unbiased, implying the minimization of the mean-squared error (MSE) and maximization of likelihood. This is a well-accepted premise. However, complications arise when extending this logic to combined estimators involving multiple time series, such as a covariance matrix. The traditional understanding of "unbiased" may shift based on our objective. If the goal is to minimize the MSE over all matrix elements, i.e., the Frobenius norm, then the sample estimator is no longer unbiased with respect to this combined loss. This discrepancy is highlighted by the Stein paradox [1,2]. It might appear contradictory at first, but it is not upon closer inspection. Even in the multivariate context, the sample estimator remains unbiased regarding individual losses. In simpler terms, for every matrix element, the true expectation aligns with the sample estimator. Yet, when t is small, the variances surrounding these expectation values are not negligible. As a result, an alternative estimator can leverage variance information to minimize the overall MSE, embodied by the Frobenius norm.

Thus, the primary objective of covariance cleaning theory is to introduce a multivariate estimator optimized for a combined loss. In this context, a frequently employed strategy is the correction of eigenvalues of the sample covariance matrix. The earliest of such methods, rooted in random matrix theory (RMT), was proposed in Ref. [3]. The fundamental proposition of this method is the convergence of the eigenvalue distribution of a random noise time series to the Marchenko-Pastur distribution when $n, t \to \infty$ with $q = n/t > 1$ finite. Here, n refers to the number of variables or dimensions in the data, whereas t represents the number of observations or data points. Real-world matrices, like those representing daily returns in financial markets, partially conform to the Marchenko-Pastur distribution but also exhibit outlier eigenvalues. In the method from Ref. [3], all eigenvalues λ that are less than a threshold $\lambda_{\max}$ are treated as sample noise fluctuations and are filtered out. Here, $\lambda_{\max}$ is the maximum eigenvalue predicted by the Marchenko-Pastur distribution. This value serves as a boundary between the bulk of the distribution, representing noise, and the outliers, potentially representing a signal.

However, this approach has been replaced by more effective techniques, proving that the sample eigenvalues that belong to the bulk of the Marchenko-Pastur carry relevant

*Corresponding author: christian.bongiorno@centralesupelec.fr

information. Starting from this observation, a large variety of approaches to address the optimal eigenvalue correction were proposed [4]. Needless to say, behind any idea of optimality, a proper covariance estimator target must be defined. The case of the former cited method is the Oracle estimator [5]. In a simple world, the Oracle estimator is the optimal correction that minimizes the Frobenius norm distance of the unknown population matrix. As the population matrix is unknown, assumptions like large sample size limit and stationarity are required to have an asymptotic approximation for such estimators. The latter is, of course, one of the strongest, as shown in Ref. [6].

In this work, we explore a different loss function for the target matrix. We question whether the current state-of-the-art literature holds true when one aims to minimize the information lost by the estimator when approximating the population matrix. To quantify this information loss, we employ the Kullback-Leibler (KL) divergence [7]. For two probability measures P and Q over a set $\mathcal{X}$, the KL divergence is defined as

$$\mathrm{K}(P||Q) := \int_{\mathcal{X}} \left(\log \frac{P}{Q}\right) dP. \tag{1}$$

This measure, a cornerstone of information theory, provides a quantification of the informational discrepancy between two probability distributions. It essentially measures the "informational cost" of approximating one distribution (in our case, the population matrix) with another (the estimator).

The use of the KL divergence is not confined to information theory. It has also found extensive application in various physical contexts. Indeed, the task of estimating physical quantities from noisy measurements is pervasive in physics, and methods that revolve around minimizing the KL divergence to extract information about a system have been widely employed. For example, it has been used to study spin-glasses [8], in high-energy physics [9], to extract information from gravitational waves detection [10], and to estimate the electron density in solid-state systems [11]. It has also been deployed in a host of other contexts [12,13]. Interestingly, the quantum extension of the KL divergence—the quantum relative entropy and its associated quantity, the quantum Fisher information [14]—has played a pivotal role in quantum metrology [15]. This is not surprising, considering that one of the primary objectives of quantum metrology is to estimate a physical quantity based on a finite set of noisy measurements, a problem that our work also addresses. In general, the quantum relative entropy is central to several concepts in quantum mechanics. For instance, quantities such as entanglement or purity are often estimated using the trace distance or the fidelity, both of which are related to the KL divergence [16]. In such cases, the loss of information is typically quantified using the Fisher information metric. However, the Fisher information metric's locality precludes its use in scenarios where very little is known beforehand [17]. Furthermore, in quantum mechanics, covariance matrices are instrumental in characterizing the entanglement properties of multipartite systems [18]. Specifically, the covariance matrix can be leveraged to construct the logarithmic negativity, a measure of entanglement [19,20].

The problem of estimating physical quantities from noisy measurements when the underlying probability distribution is a fat-tailed distribution, such as the Student's t distribution, is common in various fields of physics [21–24] and finance [25–27]. Specifically, in finance, daily returns are well approximated using the Student's t distribution [28], although, at larger time scales, a slow convergence toward a Gaussian distribution is observed [29]. In practical applications, considering larger timescales can be problematic due to factors like nonstationarity and the potential for missing significant market events. Consequently, daily returns, with their rich informational content, remain a focus for many financial analysts and researchers. Furthermore, leveraging results from information theory, Ref. [30] proved that fat tails significantly affect the largest eigenvalue, differing notably from the Gaussian case. This shows that fat tails are an important feature that cannot be overlooked.

In this context, the application of the Kullback-Leibler (KL) divergence for correlation matrix filtering for Normal multivariate variables was first introduced in Ref. [31]. A preliminary attempt to generalize this to multivariate heavy-tailed distributions was made in Ref. [32], although it considered only a simple single factor model with heavy tails. To date, a closed expression of the KL divergence for multivariate Student's t distributions remains elusive, with the work in Ref. [33] providing a closed expression under the quasi-normality assumption, and only for a Student's t distributed probability.

In our work, we delve deeper into this problem. We explore the relationship between estimators based on the minimal Frobenius norm and the KL divergence, and how they behave differently for fat-tailed distributions. We highlight that, while the target estimators for the minimal Frobenius norm and KL coincide for normal multivariate variables, they diverge in the finite n regime for fat-tailed distributions. Consequently, we develop a numerical approach for optimal correction for the Student's t distribution and derive asymptotically the limiting equation for the KL divergence in the Student's t case in the large n (thermodynamic) limit. In doing so, our work illuminates the interplay between the choice of the estimator and the underlying distribution, offering valuable insights into the estimation problem in situations where a Gaussian assumption is not suitable.

II. PROBLEM STATEMENT

The main goal of the filtering methods is to find the best rotational invariant estimator (RIE) for the population correlation matrix $\mathbf{C}$. The RIE is defined as

$$\boldsymbol{\Xi}(\boldsymbol{\Lambda}) := \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}', \tag{2}$$

where $\mathbf{V} \in O(n)$ are given, and they are not eigenvectors of $\mathbf{C}$ but, in general, are the eigenvectors of the sample correlation matrix. The standards correlation cleaning approaches rely on finding the eigenvalue matrix $\boldsymbol{\Lambda}_F$ that minimizes the Frobenius norm distance with the population matrix $||\boldsymbol{\Xi}(\boldsymbol{\Lambda}) - \mathbf{C}||_F$. Such estimator, when the population matrix $\mathbf{C}$ is known is called Oracle, and it can be obtained [5] from

$$\boldsymbol{\Lambda}_F = (\mathbf{V}'\mathbf{C}\mathbf{V})_d, \tag{3}$$

where the operator $(\bullet)_d$ sets to zero the out-of-diagonal elements.

The optimal RIE that minimizes the KL divergence assuming that the population matrix is known, must be derived from the multivariate form of the KL,

$$\mathrm{K}[\mathbf{C}||\mathbf{\Xi}(\mathbf{\Lambda})] := \mathcal{E}\left[\log\left(\frac{\mathcal{P}(x;\mathbf{C})}{\mathcal{P}(x;\mathbf{\Xi}(\mathbf{\Lambda}))}\right)\right]_{\mathcal{P}(x;\mathbf{C})}. \tag{4}$$

Here, the operator $\mathcal{E}$ stands for the expectation. In this formulation, the KL divergence represents the expected information loss when we use $\mathcal{P}(x;\mathbf{\Xi}(\mathbf{\Lambda}))$ as an approximation for data that are actually distributed according to $\mathcal{P}(x;\mathbf{C})$. More specifically, it calculates the expected value of the logarithmic ratio of the probability density functions (PDFs) with covariance matrices $\mathbf{C}$ and $\mathbf{\Xi}(\mathbf{\Lambda})$ for a multivariate random variable with zero means and population covariance matrix $\mathbf{C}$.

III. MULTIVARIATE GAUSSIAN

For multivariate Gaussian variables with zero means, the KL divergence is already well-established [31] and can be expressed as follows:

$$\mathrm{K}[\mathbf{C}||\mathbf{\Xi}(\mathbf{\Lambda})] = \frac{1}{2}\left[\mathrm{Tr}[\mathbf{\Xi}(\mathbf{\Lambda})^{-1}\mathbf{C}] - n + \log\left(\frac{|\mathbf{\Xi}(\mathbf{\Lambda})|}{|\mathbf{C}|}\right)\right]. \tag{5}$$

Interestingly, for Gaussian multivariate variables, the eigenvalues that minimize the Frobenius norm also minimize the KL divergence, i.e., $\mathbf{\Lambda}_F = \mathbf{\Lambda}_{\mathrm{KL}}$. The former result can be obtained by solving

$$\begin{aligned}\partial_{\lambda_k}\mathrm{K}[\mathbf{C}||\mathbf{\Xi}(\mathbf{\Lambda})] &= \frac{1}{2}\left(\frac{1}{\lambda_k} - \frac{1}{\lambda_k^2}\mathrm{Tr}[\mathbf{v}_k\mathbf{v}_k'\mathbf{C}]\right)\\ &= \frac{1}{2}\left(\frac{1}{\lambda_k} - \frac{1}{\lambda_k^2}\mathbf{v}_k'\mathbf{C}\mathbf{v}_k\right) = 0.\end{aligned} \tag{6}$$

That leads to Eq. (3).

IV. MULTIVARIATE *T* STUDENT

If both distributions are instead two multivariate t-student random variables, then the computation is more challenging and requires tailored approximations. The PDF for a standardized t student of n random variables is

$$\mathcal{P}(\mathbf{x};\mathbf{C},\nu) = \frac{\Gamma[(\nu+n)/2]}{\Gamma(\nu/2)\nu^{n/2}\pi^{n/2}|\mathbf{C}|^{1/2}}\left(1+\frac{1}{\nu}\mathbf{x}'\mathbf{C}^{-1}\mathbf{x}\right)^{-\frac{n+\nu}{2}}. \tag{7}$$

Our first step is to illustrate the difference between the Gaussian and the t-student cases in a low sample size (n) scenario. For this, we have developed a Monte Carlo methodology, which can be found in the referenced repository [34], that calculates the expected value of multiple random realizations of Eq. (4) for random variables drawn from the distribution given in Eq. (7). In the top panel of Fig. 1, we depict the difference between the minimum of the KL divergence and the Frobenius norm. In this particular example, we have used $n = 2$, which means that the correlation's eigenvalue has just one degree of freedom, given that they must sum up to n. This figure also includes a numerical approximation of the KL divergence obtained through an integral

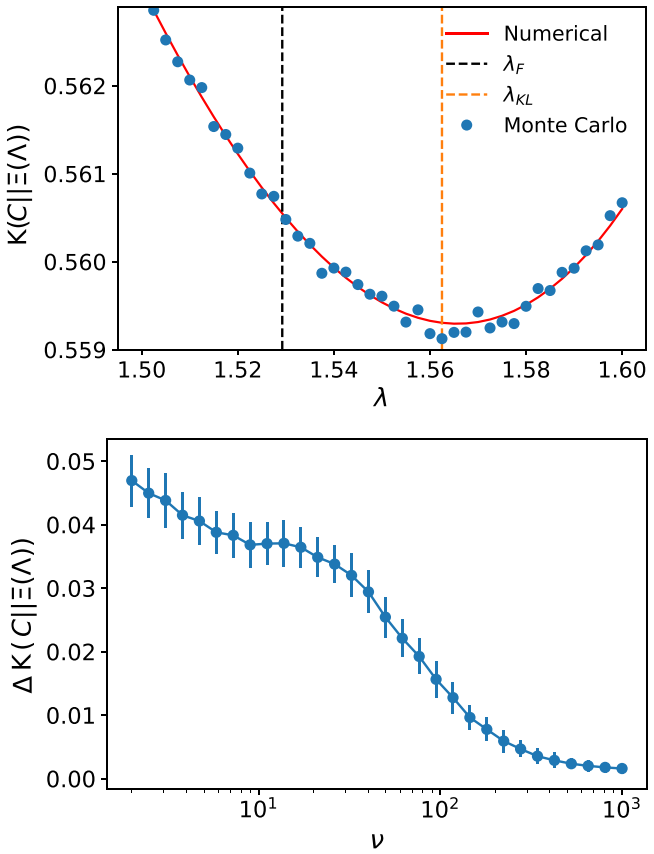

FIG. 1. The top panel shows the KL for a matrix with $n = 2$ and a Student's t distribution with $\nu = 4$, the plot reports the Monte Carlo estimation and the numerical quadrature obtained integrating the two variables in sufficiently large interval ($[-100, 100]$). The plot shows eigenvalues for the minimal KL and Frobenius norms, highlighting significant differences. The bottom panel shows the difference in KL between $\mathbf{\Xi}(\mathbf{\Lambda}_F)$ and $\mathbf{\Xi}(\mathbf{\Lambda}_{\mathrm{KL}})$ for the Student's t distribution with parameter ν of a synthetic covariance matrix $\mathbf{C}$ with $n = 30$ and an eigenvalue distribution that comes from a geometric progression with exponent 1.8. The eigenvector $\mathbf{V}$ is obtained by applying a small random rotation in terms of Euler angles to the original eigenvalues of $\mathbf{C}$. The rotation of $\mathbf{C}$'s eigenvectors is done to avoid trivial solutions, namely $\lambda = 1$. The optimal $\mathbf{\Lambda}_{\mathrm{KL}}$ is obtained with a Monte Carlo computed with 10 000 samples. The plot indicates the mean and the standard deviation of 100 runs.

quadrature method. However, this method becomes impractical as n increases. For larger values of n, we can obtain a numerical approximation of ΔK by using a combination of sequential least-squares programming (SLQP) minimization and the Monte Carlo approach. This numerical approximation accentuates the discrepancy with the Frobenius norm estimator as ν approaches 2, as demonstrated in Fig. 1. This discrepancy, even in just one example, refutes the assumption that both metrics yield the same minimum.

A. Asymptotic derivation

Given that a numerical discrepancy is observed for small n, our aim is to derive an asymptotic approximation of the KL for two Student's t distributions in the large n limit. This involves

calculating the expected value of

$$\log\left(\frac{\mathcal{P}(\mathbf{x};\mathbf{C},\nu)}{\mathcal{P}(\mathbf{x};\boldsymbol{\Xi}(\boldsymbol{\Lambda}),\nu)}\right) = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Xi}(\boldsymbol{\Lambda})|}{|\mathbf{C}|} + (n+\nu)\log\left(\frac{1+\frac{1}{\nu}\mathbf{x}'\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})\mathbf{x}}{1+\frac{1}{\nu}\mathbf{x}'\mathbf{C}^{-1}\mathbf{x}}\right)\right], \tag{8}$$

with $x \sim \mathcal{P}(\mathbf{x};\mathbf{C},\nu)$. To compute the expected value of the first term, we leverage the linearity of the expectation operator and examine each logarithm of the ratio separately. The last term poses a greater challenge. One might consider applying the replica trick [35,36]. The replica trick is based on the equation $\mathcal{E}[\log(Z(\mathbf{x}))] = \lim_{\eta\to 0}\frac{\mathcal{E}[Z(\mathbf{x})^\eta]-1}{\eta}$ where in our case Z is the argument of the logarithm in the last term. Unfortunately, the analytical calculations of the η dependence of $\mathcal{E}[Z(\mathbf{x})^\eta]$, where $\mathbf{x}$ is distributed according to a Student's t distribution, results extremely challenging. Moreover, they may not even be analytically calculable. Therefore, an alternative approach is required. To address this, we consider the first-order two-variable Taylor expansion of the variance of the argument of the logarithm. This expansion is centered around the expected values of the quadratic bilinear forms $\mathbf{x}'\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})\mathbf{x}$ and $\mathbf{x}'\mathbf{C}^{-1}\mathbf{x}$, which we denote as A and B, respectively, and $c = 1/\nu$. The detailed computations for the variance, the covariance, and the expectation in the multivariate t distribution case are provided in Ref. [37]. This approach leads us to

$$\begin{aligned}\mathcal{V}\left[\frac{1+cA}{1+cB}\right] \lesssim \tilde{\mathcal{V}}\left[\frac{1+cA}{1+cB}\right] &= c^2\left[\frac{\mathcal{V}[A]}{(1+c\mathcal{E}[B])^2} - 2\frac{1+c\mathcal{E}[A]}{(1+c\mathcal{E}[B])^3}\mathcal{C}[A,B] + \frac{(1+c\mathcal{E}[A])^2}{(1+c\mathcal{E}[B])^4}\mathcal{V}[B]\right] \\ &= \frac{2(\nu-2)n\Big(\nu - n\overline{\mathrm{Tr}[\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})]}^2 + \overline{\mathrm{Tr}[\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})]}(\nu+n-2) - 2(\nu-2)\overline{\mathrm{Tr}[\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})]} - 2\Big)}{(\nu-4)(\nu+n-2)^3}\end{aligned} \tag{9}$$

where we extracted the system size dependence from the traces $\mathrm{Tr}[\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})] = n\overline{\mathrm{Tr}[\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})]}$ and $\mathrm{Tr}[\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})] = n\overline{\mathrm{Tr}[\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})]}$.
We have numerically observed that this approximation overestimates the actual variance. By recognizing that

$$\lim_{n\to\infty}\tilde{\mathcal{V}}\left[\frac{1+\frac{1}{\nu}\mathbf{x}'\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})\mathbf{x}}{1+\frac{1}{\nu}\mathbf{x}'\mathbf{C}^{-1}\mathbf{x}}\right] = 0, \tag{10}$$

we can also infer that the true variance approaches zero in the same limit. Consequently, in high dimensions, the distribution of the ratio converges to a δ distribution centered at their expected values. This simplification facilitates the computation of the KL in the large n limit. The KL divergence can then be expressed as

$$\begin{aligned}&\mathrm{K}[\mathbf{C}||\boldsymbol{\Xi}(\boldsymbol{\Lambda})] \\ &\approx \frac{1}{2}\left\{\log\frac{|\boldsymbol{\Xi}(\boldsymbol{\Lambda})|}{|\mathbf{C}|} + (n+\nu)\right. \\ &\left.\times\left[\log\left(1+\frac{n\overline{\mathrm{Tr}[\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})]}}{\nu-2}\right) - \log\left(1+\frac{n}{\nu-2}\right)\right]\right\}.\end{aligned} \tag{11}$$

Then the normalized KL can be derived with an asymptotic limit

$$\begin{aligned}\overline{\mathrm{K}[\mathbf{C}||\boldsymbol{\Xi}(\boldsymbol{\Lambda})]} &= \lim_{n\to\infty}\frac{\mathrm{K}[\mathbf{C}||\boldsymbol{\Xi}(\boldsymbol{\Lambda})]}{n} \\ &= \tfrac{1}{2}(\overline{\log|\boldsymbol{\Xi}(\boldsymbol{\Lambda})|} - \overline{\log|\mathbf{C}|} + \log\overline{\mathrm{Tr}[\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})]}),\end{aligned} \tag{12}$$

and it is independent of ν. From the former equation, it is possible to obtain the Gaussian case with the limit

$$\begin{aligned}&\lim_{\nu\to\infty}\frac{\mathrm{K}[\mathbf{C}||\boldsymbol{\Xi}(\boldsymbol{\Lambda})]}{n} \\ &= \tfrac{1}{2}(\overline{\log|\boldsymbol{\Xi}(\boldsymbol{\Lambda})|} - \overline{\log|\mathbf{C}|} + \overline{\mathrm{Tr}[\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})]} - 1).\end{aligned} \tag{13}$$

The derivative of the normalized KL of Eq. (12) in the eigenvalues by expressing n again, that in the large limit is approximately

$$\partial_{\lambda_k}\overline{\mathrm{K}[\mathbf{C}||\boldsymbol{\Xi}(\boldsymbol{\Lambda})]} \approx \frac{1}{2}\left(\frac{1}{n\lambda_k} - \frac{1}{\lambda_k^2}\frac{\mathbf{v}_k'\mathbf{C}\mathbf{v}_k}{\mathrm{Tr}[\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})]}\right), \tag{14}$$

which is equal to zero in the $\Lambda_F = (\mathbf{V}'\mathbf{C}\mathbf{V})_d$ since $\mathrm{Tr}[\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}_F^{-1})] = n$. As a result, the former equation has the same zeros of Eq. (6). Proving the equivalence of the minimum for the Gaussian and t-student cases in the large system limit.

Another interesting observation is that in the n large limit also the normalized KL of Eq. (12) in the Λ_F is equal for the Gaussian and student's t cases to

$$\overline{\mathrm{K}[\mathbf{C}||\boldsymbol{\Xi}(\boldsymbol{\Lambda}_\mathbf{F})]} = \tfrac{1}{2}(\overline{\log|\boldsymbol{\Xi}(\boldsymbol{\Lambda}_\mathbf{F})|} - \overline{\log|\mathbf{C}|}). \tag{15}$$

Finally, if ν is not a negligible fraction of n, then we could write $\nu = hn$ with h finite and nonzero. Then,

$$\begin{aligned}&\lim_{n\to\infty}\frac{\mathrm{K}[\mathbf{C}||\boldsymbol{\Xi}(\boldsymbol{\Lambda})]}{n} \\ &= \tfrac{1}{2}[\overline{\log|\boldsymbol{\Xi}(\boldsymbol{\Lambda})|} - \overline{\log|\mathbf{C}|} + (1+h)\log(h+\overline{\mathrm{Tr}[\mathbf{C}\boldsymbol{\Xi}(\boldsymbol{\Lambda}^{-1})]}) \\ &\quad - (1+h)\log(1+h)].\end{aligned} \tag{16}$$

The former equation highlights that a discrepancy between the Normal and Student' t case is observed for small h, while the discrepancy disappears for $h\to\infty$ or if the equation is computed in Λ_F.

In Fig. 2 we show that our approximation converges pretty well to the asymptotic expectations for moderately large $n = 1000$. In particular, on the top plot, we confirm that the deviation from the Normal expectations is observed also for very large values of ν, whenever the ratio $h = \nu/n$ is not large. In the bottom plot, we confirm Eq. (15), in fact, all estimates converge to the same values independently from the value of h.

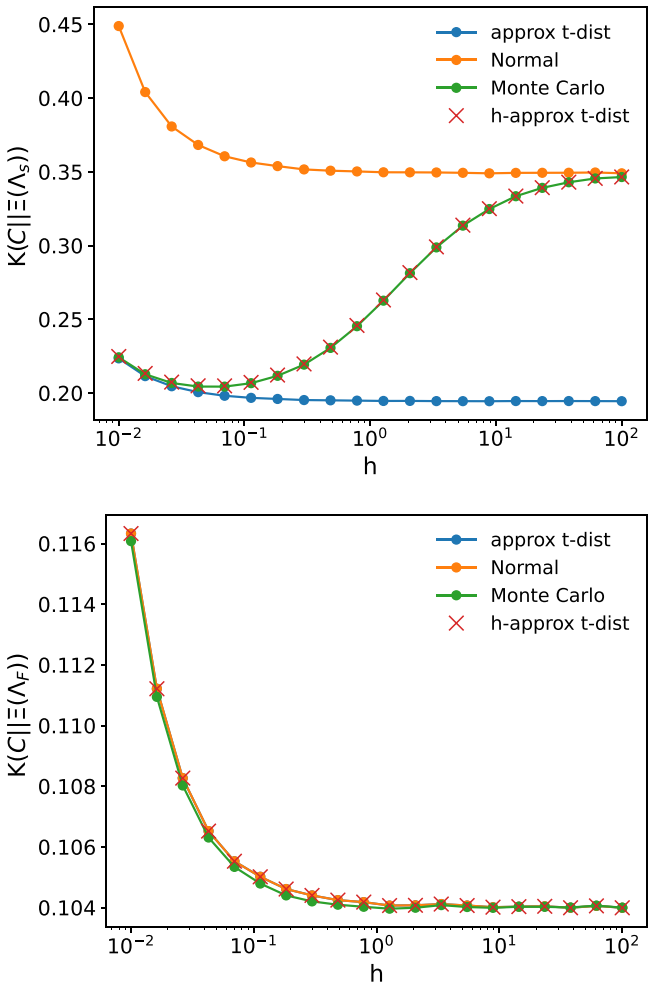

FIG. 2. The top panel shows simulations with $n = 1000$ of the KL between the population and the sample covariance matrix as a function of h; the bottom plot shows the KL between the population matrix and the Oracle estimator. The different lines represent, Eq. (12) (blue), Eq. (5) (orange), the numerical Monte Carlo (green), and Eq. (16) (red). The values are averaged over 100 independent runs.

V. DISCUSSION

In summary, we have demonstrated that the key step of optimal covariance cleaning theory, namely minimizing the Frobenius norm between the true population covariance matrix and a rotational invariant estimator, is equivalent to minimizing the loss of information between the true population covariance and the rotational invariant estimator for normal multivariate variables. We have shown that this equivalence does not necessarily hold for Student's t distributions in finite-sized matrices, but that it holds asymptotically for large matrices. This result might help to extend the applicability of random matrix theory to Student's t distributions, which are commonly encountered in real-world applications such as finance.

Our work contributes to reinforcing the connection between statistical random matrix theory and estimation theory in physics, within the framework of information theory. The use of information theory has been instrumental in the development of a wide range of physical theories and models, such as the maximum entropy principle, which has been used to derive equilibrium thermodynamics from information theory. Our findings suggest that information theory can also provide valuable insights in the field of optimal covariance cleaning theory, which has important applications in statistical data analysis, signal processing, and machine learning.

In future work, it would be interesting to explore the applicability of our results to other heavy-tailed distributions and to investigate whether the use of alternative metrics for quantifying the loss of information, such as the Kullback-Leibler divergence, could lead to improved performance in finite-sized matrices. Additionally, it would be valuable to study the performance of optimal covariance cleaning in the presence of missing or incomplete data, which is a common issue in many real-world applications. Overall, our work highlights the potential of combining concepts from information theory and random matrix theory to develop more robust and accurate statistical methods for analyzing complex data sets.

[1] W. James and C. Stein, Estimation with quadratic loss, in *Breakthroughs in Statistics: Foundations and Basic Theory* (Springer, Berlin, 1992), pp. 443–460

[2] O. Ledoit and M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivariate Anal. **88**, 365 (2004).

[3] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Noise dressing of financial correlation matrices, Phys. Rev. Lett. **83**, 1467 (1999).

[4] O. Ledoit and S. Péché, Eigenvectors of some large sample covariance matrix ensembles, Probab. Theory Relat. Fields **151**, 233 (2011).

[5] J. Bun, R. Allez, J.-P. Bouchaud, and M. Potters, Rotational invariant estimator for general noisy matrices, IEEE Trans. Inf. Theory **62**, 7475 (2016).

[6] C. Bongiorno, D. Challet, and G. Loeper, Filtering time-dependent covariance matrices using time-independent eigenvalues, J. Stat. Mech. (2023) 023402.

[7] J. M. Joyce, Kullback-Leibler divergence, in *International Encyclopedia of Statistical Science* (Springer, Berlin, 2011), pp. 720–722.

[8] L. Leuzzi, G. Parisi, F. Ricci-Tersenghi, and J. J. Ruiz-Lorenzo, Dilute one-dimensional spin glasses with power law decaying interactions, Phys. Rev. Lett. **101**, 107203 (2008).

[9] R. Gambhir, B. Nachman, and J. Thaler, Learning uncertainties the frequentist way: Calibration and correlation in high energy physics, Phys. Rev. Lett. **129**, 082001 (2022).

[10] A. J. K. Chua and M. Vallisneri, Learning Bayesian posteriors with neural networks for gravitational-wave inference, Phys. Rev. Lett. **124**, 041102 (2020).

[11] R. Y. de Vries, W. J. Briels, and D. Feil, Critical analysis of nonnuclear electron-density maxima and the maximum entropy method, Phys. Rev. Lett. **77**, 1719 (1996).
[12] M. T. DiMario and F. E. Becerra, Single-shot non-Gaussian measurements for optical phase estimation, Phys. Rev. Lett. **125**, 120505 (2020).
[13] D. Everett *et al.* (JETSCAPE Collaboration), Phenomenological constraints on the transport properties of QCD matter with data-driven model averaging, Phys. Rev. Lett. **126**, 242301 (2021).
[14] V. Vedral, The role of relative entropy in quantum information theory, Rev. Mod. Phys. **74**, 197 (2002).
[15] J. S. Sidhu and P. Kok, Geometric perspective on quantum parameter estimation, AVS Quantum Sci. **2**, 014701 (2020).
[16] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, UK, 2012).
[17] S. L. Braunstein and C. M. Caves, Statistical distance and the geometry of quantum states, Phys. Rev. Lett. **72**, 3439 (1994).
[18] G. Adesso, A. Serafini, and F. Illuminati, Extremal entanglement and mixedness in continuous variable systems, Phys. Rev. A **70**, 022318 (2004).
[19] M. B. Plenio, Logarithmic negativity: A full entanglement monotone that is not convex, Phys. Rev. Lett. **95**, 090503 (2005).
[20] G. Vidal and R. F. Werner, Computable measure of entanglement, Phys. Rev. A **65**, 032314 (2002).
[21] A. J. Majda and Y. Lee, Conceptual dynamical models for turbulence, Proc. Natl. Acad. Sci. USA **111**, 6548 (2014).
[22] T. S. Biró and A. Jakovác, Power-law tails from multiplicative noise, Phys. Rev. Lett. **94**, 132302 (2005).
[23] A.-L. Barabási, Scale-free networks: A decade and beyond, Science **325**, 412 (2009).
[24] G. L. Millhauser, E. E. Salpeter, and R. E. Oswald, Diffusion models of ion-channel gating and the origin of power-law distributions from single-channel recording, Proc. Natl. Acad. Sci. USA **85**, 1503 (1988).
[25] S. Thurner, J. D. Farmer, and J. Geanakoplos, Leverage causes fat tails and clustered volatility, Quantum Fin. **12**, 695 (2012).
[26] F. Lillo and R. N. Mantegna, Power-law relaxation in a complex system: Omori law after a financial market crash, Phys. Rev. E **68**, 016119 (2003).
[27] D. Sornette, J. V. Andersen, and P. Simonetti, Portfolio theory for "fat tails," Int. J. Theor. Appl. Fin. **03**, 523 (2000).
[28] J.-P. Bouchaud, Elements for a theory of financial risks, Physica A **285**, 18 (2000).
[29] L. A. Amaral, V. Plerou, P. Gopikrishnan, M. Meyer, and H. E. Stanley, The distribution of returns of stock prices, Int. J. Theor. Appl. Fin. **03**, 365 (2000).
[30] M. Filiasi, G. Livan, M. Marsili, M. Peressi, E. Vesselli, and E. Zarinelli, On the concentration of large deviations for fat tailed distributions, with application to financial data, J. Stat. Mech. (2014) P09030.
[31] M. Tumminello, F. Lillo, and R. N. Mantegna, Kullback-Leibler distance as a measure of the information filtered from multivariate data, Phys. Rev. E **76**, 031123 (2007).
[32] G. Biroli, J.-P. Bouchaud, and M. Potters, The student ensemble of correlation matrices: Eigenvalue spectrum and Kullback-Leibler entropy, Acta Phys. Pol. B **38**, 4009 (2007).
[33] J. E. Contreras-Reyes, Asymptotic form of the Kullback–Leibler divergence for multivariate asymmetric heavy-tailed distributions, Physica A **395**, 200 (2014).
[34] C. Bongiorno, KL-divergence monte carlo estimation, https://gitlab-research.centralesupelec.fr/2019bongiornc/kl-studentst (2023).
[35] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* (World Scientific, Singapore, 1987), Vol. 9.
[36] M. Potters and J.-P. Bouchaud, *A First Course in Random Matrix Theory* (Cambridge University Press, Cambridge, UK, 2020).
[37] J.-Y. Rong, Z.-F. Lu, and X.-Q. Liu, On quadratic forms of multivariate *t* distribution with applications, Commun. Stat.-Theory Methods **41**, 300 (2012).