



**Two-step estimators of high-dimensional correlation matrices**Andrés García-Medina <sup>\*</sup>*Centro de Investigación en Matemáticas, Unidad Monterrey, Av. Alianza Centro 502, PIIT 66628, Apodaca, Nuevo León, México and Consejo Nacional de Humanidades, Ciencias y Tecnologías, Av. Insurgentes Sur 1582, Col. Crédito Constructor 03940, Ciudad de México, México*Salvatore Micciché <sup>†</sup>*Dipartimento di Fisica e Chimica Emilio Segrè, Università degli Studi di Palermo, Viale delle Scienze, Ed. 18, 90128, Palermo, Italy*Rosario N. Mantegna <sup>‡</sup>*Dipartimento di Fisica e Chimica Emilio Segrè, Università degli Studi di Palermo, Viale delle Scienze, Ed. 18, 90128, Palermo, Italy and Complexity Science Hub Vienna, Josefstädter Strasse 39, 1080 Vienna, Austria*

(Received 3 January 2023; revised 28 September 2023; accepted 2 October 2023; published 23 October 2023)

We investigate block diagonal and hierarchical nested stochastic multivariate Gaussian models by studying their sample cross-correlation matrix on high dimensions. By performing numerical simulations, we compare a filtered sample cross-correlation with the population cross-correlation matrices by using several rotationally invariant estimators (RIEs) and hierarchical clustering estimators (HCEs) under several loss functions. We show that at large but finite sample size, sample cross-correlations filtered by RIE estimators are often outperformed by HCE estimators for several of the loss functions. We also show that for block models and for hierarchically nested block models, the best determination of the filtered sample cross-correlation is achieved by introducing two-step estimators combining state-of-the-art nonlinear shrinkage models with hierarchical clustering estimators.

DOI: [10.1103/PhysRevE.108.044137](https://doi.org/10.1103/PhysRevE.108.044137)**I. INTRODUCTION**

In recent years, many research areas have dealt with multivariate time series. Examples are physics, neuroscience, finance, climatology, genomics, etc. In all these research areas, investigators perform  $n$  measurements of a system characterized by  $p$  variables, obtaining an observation matrix  $\mathbf{Y}$  of dimension  $p \times n$ . After standardizing the  $p$  series of  $n$  records, one can compute the  $p \times p$  sample cross-correlation matrix  $\mathbf{E}$ . Sample cross-correlation matrices computed from a finite set of multivariate data generally differ from the population cross-correlation matrix  $\mathbf{C}$  associated with the model generating multivariate data. Since the seminal work of Marčenko and Pastur [1], many studies have considered the spectral properties of sample cross-correlation matrices and have used these theoretical results to set up a null model useful to discriminate information that can be extracted from data, i.e., information not compatible with a null model, and information hard to be distinguished from noise, i.e., a null model, in empirical data [2].

Comparing sample and population cross-correlation requires choosing a loss function, i.e., a function specifying a penalty for an incorrect estimate from the underlying statistical model. In the literature, several loss functions have been

proposed, and the choice of a specific one must be related to the specific problem considered. The most used loss functions are Frobenius loss, Stein loss, and Kullback-Leibler (KL) divergence.

In Refs. [3,4], the authors analytically demonstrated that the expected KL divergence of a sample correlation matrix concerning the population matrix does not depend on the reference model. The authors used the KL divergence to measure how informative the filtered correlation matrices are when applying spectral and hierarchical clustering techniques separately. The studies performed simulations against factor model structures [5] and empirical studies with financial time series listed on U.S. equity markets. As spectral techniques, the authors used two variations of methods known in the literature as the *clipping* technique. The clipping (also known as filtering or denoising in the econophysics community) technique was initially proposed in Refs. [6,7] and later cataloged in the family of rotationally invariant estimators (RIEs) by Ref. [2]. In particular, the clipping technique is associated with the spiked covariance matrix model [8].

RIE models for estimating the covariance matrix have been known in the mathematical statistics community since Stein [9] proposed them under the name *rotation-equivariant* estimators. His idea was to keep the eigenvectors of the sample covariance matrix while shrinking its eigenvalues. They were proposed in the classical paradigm when the number of observations is much greater than the number of variables. Ledoit and Wolf have been promoting these methods on the high-dimensional stage. In Ref. [10], they proposed an optimal

<sup>\*</sup>andres.garcia@cimat.mx<sup>†</sup>salvatore.micciche@unipa.it<sup>‡</sup>rosario.mantegna@unipa.it

linear shrinkage using random matrix theory (RMT) concepts. Later, in Ref. [11], the first nonlinear shrinkage model based on RMT and asymptotic theory was proposed. Their numerical implementation is given in Ref. [12]. On the other hand, Bun *et al.* [2] suggested a different numerical approach that is easier to implement. Both are approximations of the same model. Finally, Ledoit and Wolf gave a kernel-based solution that is essentially analytical and drastically improves the computation time by two orders of magnitude [13]. This solution is valid for general correlation structures but does not consider autocorrelations. Burda and Jarosz tackled the autocorrelation structure in a recent work [14] by using concepts of RMT and free probability.

The previous formulations led to considering nonlinear shrinkage formulas to estimate the population eigenvalues from the empirical ones and reconstructing the correlation matrices using the empirical eigenvectors. As such, they belong to the RIE family of estimators. These nonlinear shrinkage formulas are optimal with the Frobenius loss function. The problem of applying different loss functions is of current interest, as stated, for example, in Refs. [2,15], where the authors proposed quantifying the information kept by the optimal RIE compared to several estimators and metrics.

Several empirical covariance matrices present a spectrum compatible with the so-called spiked covariance matrix model [8,15], i.e., covariance matrices with an eigenvalue spectrum characterized by a few number of large isolated eigenvalues distinct from bulk eigenvalues. So-called hierarchically nested factor models, i.e., factor models with nested factors affecting distinct subgroups of elements of the systems present a spiked eigenvalue spectrum [16,17]. One of the main results of spiked covariance models is the presence of top eigenvector inconsistency [15], i.e., the observation that the angle between sample eigenvectors and the corresponding population eigenvectors have nonzero limits. This implies that the optimal choice of the nonlinear shrinkage function of eigenvalues might depend significantly on the specific loss function chosen. Top eigenvector inconsistency also suggests that filtering by RIEs using sample eigenvectors might miss some aspects of the population matrix. Another limitation of RIE methods concerns sample eigenvectors associated to small eigenvalues. They usually comprise components covering the entire set of elements, therefore presenting an eigenvector orientation quite distinct from localized eigenvectors associated with a correlated dynamics of a small group of elements.

The above observations have motivated an alternative filtering procedure of spiked correlation matrices based on hierarchical clustering. In fact, in standardized random multivariate variables with correlation matrices characterized by positive correlation coefficients, there is a one-to-one correspondence between the cophenetic matrix of a hierarchical clustering and a hierarchically nested factor model [5]. A hierarchical clustering procedure therefore provides a correlation matrix equivalent to a hierarchically nested factor model [18]. The effectiveness of filtering a correlation matrix by hierarchical clustering has been documented in several studies primarily associated with the problem of portfolio optimization in finance [19–23].

Although the subject of RIE methods has been analytically studied extensively [24], some assumptions about eigenvec-

tors might induce relevant limitations in the presence of complex systems characterized by correlation matrices with a hierarchically nested structure. Since no analytical results exist about optimal filtering by hierarchical clustering, we conduct a series of numerical experiments to evaluate the performances of different filtering methods based on RIE and on hierarchical clustering for different loss functions. Our numerical results suggest that RIE methods and hierarchical clustering methods give comparable results for systems whose population matrix is a spiked correlation matrix. We hope our results can stimulate the development of analytical results for HC filtering estimators.

Specifically, we numerically analyze block diagonal and hierarchical nested models on high dimensions and compare their behavior under several loss functions when applying RIE and hierarchical clustering estimators. We are also introducing two-step estimators that combine state-of-the-art nonlinear shrinkage models with hierarchical clustering estimators. These estimators outperform several of the most used estimators when the model of multivariate series is a block model or a hierarchically nested block model and when the statistical properties of records are Gaussian.

The paper is organized as follows. Section II describes the estimators proposed in this paper. Section III introduces the loss functions used to evaluate the difference between filtered sample correlations and population correlations when applying each estimator. Section IV gives the specifications of the models studied. Section V shows the main results obtained. Section VI analyzes and discusses the findings found.

## II. ESTIMATORS

For the sake of completeness, this section presents the estimators of the correlation matrix that we will use in our numerical analyses. These estimators can be grouped into three classes. The first ones belong to the RIE family, the second ones are of the hierarchical clustering type, and the third class combines both, which we denote as two-step estimators. It is important to emphasize that the first class of estimators is designed to deal with the estimation uncertainty inherent in the high-dimensional scenario when the number of variables is of the same order as the number of observations. The second class of estimators deals with the estimation uncertainty associated with the structure of the correlation blocks between variables. Therefore, it is focused on better detection of the financial sectors. Finally, the third class of estimators deals with both types of noise.

### A. Rotationally invariant estimators

The RIE has the property that the sample correlation matrix  $\mathbf{E}$  can be rotated by some orthogonal matrix  $\mathbf{O}$  and its estimation, denoted as  $\mathbf{\Xi}$ , must be rotated in the same direction. Therefore,  $\mathbf{\Xi}(\mathbf{E})$  can be diagonalized on the same basis as  $\mathbf{E}$  except for a fixed rotation  $\Omega$ . In this way,  $\mathbf{\Xi}(\mathbf{E})$  has the same eigenvectors as  $\mathbf{E}$  and it is possible to write

$$\mathbf{\Xi}(\mathbf{E}) = \sum_{i=1}^p \xi_i v_i v_i', \quad (1)$$

where  $v_i$  are the eigenvectors of  $\mathbf{E}$ , and  $\xi_i$  is a function of the eigenvalues  $[\lambda_j]_{j \in \{1, p\}}$  of  $\mathbf{E}$ .

The empirical correlation matrix  $\mathbf{E}$  is a trivial example that satisfies this condition. Then, a *naive estimator* is

$$\mathbf{\Xi}^{\text{naive}} = \mathbf{E}. \quad (2)$$

A classical RMT filter is proposed in Refs. [6,7] and is expressed as

$$\xi^{\text{RMT}} = \begin{cases} \bar{\lambda} & \text{if } \lambda_k < (1 + \sqrt{q})^2 \\ \lambda_k & \text{otherwise,} \end{cases} \quad (3)$$

where  $\bar{\lambda}$  represents the eigenvalues average below Marchenko-Pastur law's upper bound. Then, the estimated correlation matrix is given by

$$\mathbf{\Xi}^{\text{RMT}} = \sum_{i=1}^p \xi_i^{\text{RMT}} v_i v_i'. \quad (4)$$

A nonlinear shrinkage formula of the RIE family to estimate the unbiased covariance matrix when  $\mathbf{C}$  has a general form given by [11]

$$\xi_k^{\text{LP}} = \lim_{\epsilon \rightarrow 0^+} \frac{\lambda_k}{|1 - q + q\lambda_k G_E(\lambda_k - i\epsilon)|^2} \quad (5)$$

$$= \frac{\lambda_k}{|1 + u_k|^2} \quad (6)$$

$$= \frac{\lambda_k}{(\alpha_k + 1)^2 + \beta_k^2}, \quad (7)$$

where  $\lambda_k$  is an eigenvalue of  $\mathbf{E}$ ,  $G_E$  is the Stieltjes transform of  $\mathbf{E}$ , and since we are close to the real axis, the Sokhotski-Plemelj formula applies,

$$u_k = qT_E(\lambda_k - i0^+) = \alpha_k + i\beta_k, \quad (8)$$

where  $T_E = zG_E(z) - 1$ ,  $\alpha_k = q(\pi\lambda_k h_E(\lambda_k) - 1)$ , and  $\beta_k = q\pi\lambda_k \rho_E(\lambda_k)$ . Here,  $h_E$  denotes the Hilbert transform of  $\mathbf{E}$  and  $\rho_E$  its eigenvalue density. The corresponding estimated correlation matrix is given by the following expression:

$$\mathbf{\Xi}^{\text{LP}} = \sum_{i=1}^p \xi_i^{\text{LP}} v_i v_i'. \quad (9)$$

A recent proposal for nonlinear shrinkage expression is due to Burda and Jarosz [14], who incorporated autocorrelation into data-generating processes through a matrix  $\mathbf{A}$ . The authors gave explicit solutions for some specific models of the vector autoregressive moving average family [25]:

$$Y_{i,a} = \sum_{\beta=1}^{r_1} b_\beta Y_{i,a-\beta} + \sum_{\alpha=0}^{r_2} a_\alpha \epsilon_{i,a-\alpha}. \quad (10)$$

The key element to analytically incorporate autocorrelations is the  $\mathcal{S}$  transform of the associated matrix of coefficients  $\mathbf{A}$ , which, in principle, is not trivial to compute. Calculating  $\mathcal{S}$  requires some knowledge of the free probability [26,27]. Burda and Jarosz explicitly solved the model for  $(r_1, r_2) \in \{(1, 1), (1, 0), (2, 0), (0, 1), (0, 2)\}$ . The nonlinear shrinkage formula has the general form

$$\xi_k^{\text{BJ}} = \frac{\lambda_k \text{Im}\{1/Z_A(u_k)\}}{\text{Im}\{u_k\}}, \quad (11)$$

where  $Z_A$  is the  $Z$  transform of  $A$ . Consequently, the optimal estimator in the Frobenius sense is

$$\mathbf{\Xi}^{\text{BJ}} = \sum_{i=1}^p \xi_i^{\text{BJ}} v_i v_i'. \quad (12)$$

Note that when  $\mathbf{A} = \mathbf{I}$ , the  $\mathcal{S}$  transform of  $A$  is given by

$$S_A(t) = \frac{t+1}{tZ_A(t)} = 1 \Rightarrow Z_A(t) = \frac{t+1}{t}. \quad (13)$$

Then,

$$\xi_k^{\text{BJ}} = \frac{\lambda_k \text{Im}\{1/Z_A(u_k)\}}{\text{Im}\{u_k\}} = \frac{\lambda_k}{(\alpha_k + 1)^2 + \beta_k^2}, \quad (14)$$

and we recover Eq. (7).

In particular, the combination of parameters  $a_0 = \sqrt{1 - b_1^2}$ ,  $b_1 = e^{-1/\tau}$ ,  $a_i = b_{i-1} = 0$  (for  $i > 1$ ) represents the exponential decay model for which  $Z_A$  is known analytically [28,29],

$$Z_A(z) = \eta + \sqrt{\eta^2 - 1 + 1/z^2}, \quad (15)$$

where  $\eta = \coth(1/\tau)$ .

A further estimator proposed from a data-driven approach employs the technique known as *moving window cross-validation* (mwcv) and is denoted as the oracle estimator [30]. It is important to mention that this estimator approximates the state-of-the-art nonlinear shrinkage [31]. The expression to estimate the population eigenvalues is given by the expression

$$\xi_i^{\text{mwcv}} = \frac{1}{K} \sum_{\mu}^{K-1} (\lambda_i^{\text{train}, \mu} | \mathbf{E}^{\text{test}, \mu} | \lambda_i^{\text{train}, \mu}), \quad (16)$$

where  $K = (T_{\text{total}} - T)/T_{\text{out}}$ . The idea is to set  $T$  observations as a train and  $T_{\text{out}}$  as a test in a moving window scenario of the entire sample sequence of length  $T_{\text{total}} = KT_{\text{out}} + T$ . Here,  $\lambda_i^{\text{train}, \mu}$  represents the eigenvalues of the training sample in window  $\mu$  and  $\mathbf{E}^{\text{test}, \mu}$  the test sample covariance matrix in window  $\mu$ :

$$\mathbf{\Xi}^{\text{mwcv}} = \sum_{i=1}^p \xi_i^{\text{mwcv}} v_i v_i'. \quad (17)$$

## B. Hierarchical clustering estimators

The hierarchical clustering estimator was proposed in Ref. [3]. This estimator is based on the hierarchical clustering methods, which require a distance or dissimilarity matrix as an input. Then, we must transform the correlation matrix  $\mathbf{E}$  into a dissimilarity matrix. Here, we choose the transformation  $D_{ij} = 1 - E_{ij}$ , which satisfies the axioms of a distance measure. The clusters can generally be created through *divisive* or *agglomerative* methods. The proposed estimator considers the agglomerative strategy, which consists of four steps. The first step is to set each of the  $p$  variables in a single cluster. Next, in the second step, we search in  $D$  for the nearest (most similar) pair of variables (clusters), say  $a, b$ , and denote this distance by  $d_{ab}$ . In the third step, the clusters  $a$  and  $b$  are merged, denoted as  $(ab)$ , and the entries of  $D$  are updated by removing the rows and columns corresponding to the variables  $a$  and  $b$ . Hence the row and column regarding the new cluster  $(ab)$

distances to each of the remaining clusters are added to  $D$ . The four-step consists of repeating steps 1–3 until a single cluster is obtained, where the levels at which two clusters join together can be represented through a dendrogram. To quantify the nearest (most similar) clusters, we use the average linkage given by the agglomerative criteria [32]

$$d_{ab} = \frac{\sum_i \sum_j D_{ij}}{N_a N_b}, \quad (18)$$

where  $D_{ij}$  consider the distance between the objects  $i$  and  $j$  on clusters  $a$  and  $b$ , respectively, and  $N_a, N_b$  represent their number of items. This particular procedure, known as the average linkage clustering analysis (ALCA), enables us to compute the *cophenetic distance*  $\rho$  on the associated *dendrogram* [33], which is the distance between clusters at each hierarchical level. Finally, it is built up a dissimilarity matrix as a function of  $\rho$ :  $\mathbf{D}(\rho)$ ; and the filtered correlation matrix is obtained by  $\Xi(E)_{ij} = 1 - D_{ij}(\rho)$ . Figure 1 schematically shows the mechanism of applying the hierarchical clustering estimator under the ALCA approach to a  $4 \times 4$  matrix with a hierarchical nested structure. The example shows that the finer structures are filtered out using this procedure, and only the strongest correlation blocks are preserved.

### C. Two-step estimators

We introduce the two-step estimators, which consist of applying as a first step a RIE estimator to deal with the statistical uncertainty due to high dimensionality. Once this type of uncertainty is eliminated, a hierarchical clustering-based estimator is applied as a second step to highlight the hierarchically nested block structure. We expect that a two-step estimator presents very good performance for several loss functions because (i) a first application of a RIE reduces the error of estimation of largest eigenvalues and (ii) the second application of an appropriate filtering by hierarchical clustering can reduce the inconsistency in top eigenvectors that is unavoidably associated with RIEs. In particular, we consider the following combination of estimators because they present very good performance:

- (1) Two-step (I):  $\Xi^{\text{ALCA}}(\Xi^{\text{mwcV}}(\mathbf{E}))$ .
- (2) Two-step (II):  $\Xi^{\text{ALCA}}(\Xi^{\text{BJ}}(\mathbf{E}))$ .
- (3) Two-step (III):  $\Xi^{\text{ALCA}}(\Xi^{\text{LP}}(\mathbf{E}))$ .

### III. LOSS FUNCTIONS

We use six different loss functions to compare the effect of the different estimators on the correlation matrices. The first of these is the KL divergence.

Let  $\mathbf{A}, \mathbf{B}$  be two square matrices of dimension  $p \times p$ ; the KL divergence of Gaussian processes is given by [3]

$$K(\mathbf{A}, \mathbf{B}) = \frac{1}{2} [\ln[\det(\mathbf{B}\mathbf{A}^{-1})] + \text{Tr}(\mathbf{B}^{-1}\mathbf{A}) - p]. \quad (19)$$

We note that, under the assumption of Gaussianity,  $K(\mathbf{A}, \mathbf{B})$  is equivalent up to a factor of the commonly known inverse Stein’s loss function [34].

The second metric is the inverse KL divergence or Stein’s loss. This metric has the same expression as the KL di-

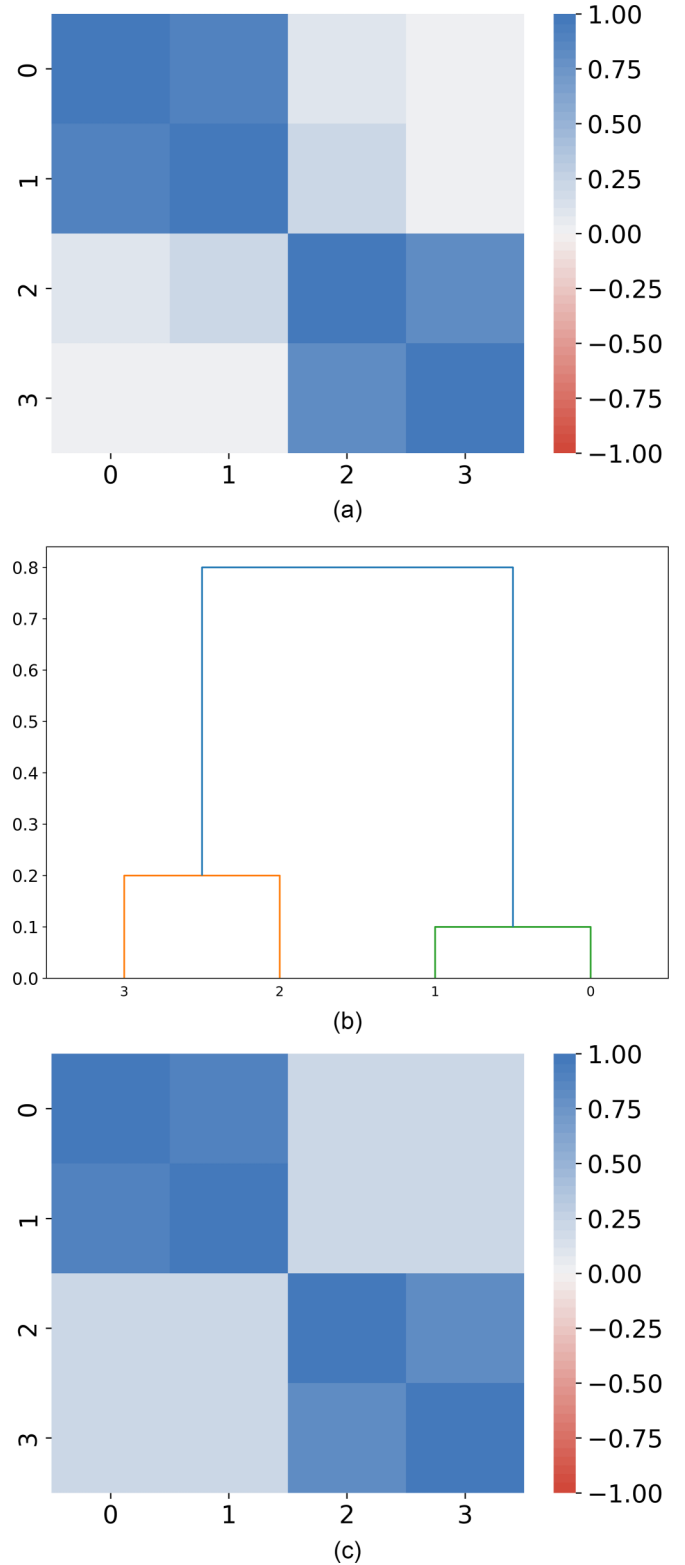


FIG. 1. The effect of applying the hierarchical clustering estimator under the ALCA approach to a  $4 \times 4$  matrix with a hierarchical nested structure. (a) Empirical matrix  $\mathbf{E}$ , (b) the associated dendrogram, and (c) the filtered correlation matrix  $\Xi(\mathbf{E})$ .

vergence given above but applied on the inverse matrices:  $K(\mathbf{A}^{-1}, \mathbf{B}^{-1})$ . It is important to mention that Stein’s loss



function  $\mathcal{L}^{\text{Stein}}$  is related to the inverse KL divergence by a scaling factor

$$\mathcal{L}^{\text{Stein}}(\mathbf{A}, \mathbf{B}) = \frac{1}{p} \text{Tr}(\mathbf{A}^{-1} \mathbf{B}) \quad (20)$$

$$-\frac{1}{p} \ln[\det(\mathbf{A}^{-1} \mathbf{B})] - 1 = \frac{2}{p} K(\mathbf{A}^{-1}, \mathbf{B}^{-1}). \quad (21)$$

This paper considers the scaled version to prevent the loss function from going to infinity with the matrix dimension, and both metrics (Stein's loss and inverse KL) are assumed to be indistinguishable. In Ref. [3], it has been shown that the expected value of the KL divergence does not depend on the specific model under the Gaussian assumption. Consider two independent sample covariance matrices  $\mathbf{E}_1, \mathbf{E}_2$  coming from the parent population  $\mathbf{C}$ ; the next scaled expectations are valid under Gaussian assumptions

$$\begin{aligned} \mathbb{E}[K(\mathbf{E}_1, \mathbf{E}_2)] &= \frac{p+1}{n-p-1}, \\ \mathbb{E}[K(\mathbf{C}, \mathbf{E})] &= \frac{1}{p} \left[ p \ln \left( \frac{2}{n} \right) + \sum_{t=n-p+1}^n \left( \frac{\Gamma'(t/2)}{\Gamma(t/2)} \right) \right. \\ &\quad \left. + \frac{p(p+1)}{n-p-1} \right], \end{aligned} \quad (22)$$

where  $\Gamma(x)$  is the usual gamma function and  $\Gamma'(x)$  is the derivative of  $\Gamma(x)$ . We have scaled the metric by  $\frac{2}{p}$  to have an exact equivalence between Stein's loss and the KL divergence, yet the original result does not consider this factor.

Moreover, the Frobenius norm is given by

$$F(\mathbf{A}, \mathbf{B}) = \frac{1}{p} \text{Tr}[(\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})']; \quad (23)$$

the corresponding inverse Frobenius is  $F(\mathbf{A}^{-1}, \mathbf{B}^{-1})$ . The Frobenius and the inverse KL divergence are designed to deal with the covariance matrix, while the inverse Frobenius and the KL divergence to the inverse covariance matrix, also known as the precision matrix.

An interesting metric in the framework of the classical portfolio theory is the minimum-variance loss function [35]:

$$\text{MV}(\mathbf{A}, \mathbf{B}) = \frac{\text{Tr}(\mathbf{B}^{-1} \mathbf{A} \mathbf{B}^{-1})/p}{[\text{Tr}(\mathbf{B}^{-1})/p]^2} - \frac{1}{\text{Tr}(\mathbf{A}^{-1})/p} \quad (24)$$

One last metric is the symmetrized Stein's loss, a combination of Stein's loss and the inverse of Stein's loss:

$$\text{SS}(\mathbf{A}, \mathbf{B}) = \frac{1}{p} \text{Tr}(\mathbf{B} \mathbf{A}^{-1} + \mathbf{A}^{-1} \mathbf{B}) - 2. \quad (25)$$

This metric pays equal attention to the problem of estimating the covariance and the precision matrix.

#### IV. MODEL

We consider a multiplicative noise model with the following structure:

$$\mathbf{Y} = \sqrt{\mathbf{C}} \mathbf{X} \sqrt{\mathbf{A}}, \quad (26)$$

$$\mathbf{E} = \frac{1}{n} \sqrt{\mathbf{C}} \mathbf{X} \mathbf{A} \mathbf{X}' \sqrt{\mathbf{C}}, \quad (27)$$

where  $\mathbf{Y}$  is the  $p \times n$  data matrix,  $\mathbf{C}$  is the  $p \times p$  population cross-correlation matrix,  $\mathbf{A}$  is the  $n \times n$  autocorrelation matrix, and  $X_{ij} \sim \mathcal{N}(0, 1)$ , that is, each element  $X_{ij}$  follows a standard Gaussian distribution. The correlation model  $\mathbf{C}$  is first constructed as follows:

$$\mathbf{L}_{kl} = \begin{cases} \gamma_{kl}, & \text{if } k = k(l), \dots, k(l + p_l) \\ 0, & \text{otherwise,} \end{cases} \quad (28)$$

where  $\mathbf{L}$  is the loading matrix of dimension  $p \times b$ ,  $k = 1, \dots, p$ ,  $l = 1, \dots, b$ ,  $p_l$  the size of each block  $l$ ,  $b$  being the number of blocks ( $b \leq p$ ), and  $\{k(l), k(l + p_l)\}$  the initial and last values of the given block  $l$ . Once defined,  $\mathbf{L}$ , the population correlation matrix  $\mathbf{C}$  is obtained simply by the expressions

$$\mathbf{Q} = \mathbf{L} \mathbf{L}', \quad (29)$$

$$C_{ij} = \delta_{ij} + Q_{ij}(1 - \delta_{ij}), \quad (30)$$

where  $\delta_{ij}$  denotes the Kronecker delta. We have considered a block diagonal and hierarchical nested block matrix structure to model  $\mathbf{C}$ . The first model comprises 12 independent diagonal blocks, while the second model is constructed with 12 overlapped diagonal blocks. In particular, we consider a homogeneous specification of the loading factors  $\gamma_{kl} = \gamma = 0.3$ . In both models, the block sizes  $p_l$  are heterogeneous as well as the initial and last values  $\{k(l), k(l + p_l)\}$ .

We analyze three different cases. The first case considers the block diagonal model with autocorrelation matrix  $\mathbf{A} = \mathbf{I}$ . The second is the hierarchical nested model with autocorrelation matrix  $\mathbf{A} = \mathbf{I}$ . And the third is the same hierarchical nested model but with autocorrelation elements of the form  $\mathbf{A}_{ij} = e^{-\frac{|i-j|}{\tau}}$ , where we have fixed  $\tau = 3$ . In other words, the first two cases represent time series without memory, while in the third case, the memory decays exponentially as a function of the separation between observations  $i, j$ .

#### V. RESULTS

We generate  $m$  realizations of multivariate time series  $\mathbf{Y}$  for each study case. Each sample matrix  $\mathbf{E}$  is computed and filtered using estimators described in Sec. II. Subsequently, the estimator's performance is measured through the six loss functions described in Sec. III. Figure 2 shows a graphical representation of the block diagonal model accompanied by a single realization of the process with  $\mathbf{A} = \mathbf{I}$  (case 1). Likewise, Fig. 3 shows a graphical representation of the hierarchical nested model accompanied by a single realization of the process with  $\mathbf{A} = \mathbf{I}$  (case 2) and with autocorrelation elements of the form  $\mathbf{A}_{ij} = e^{-|i-j|/3}$  (case 3). The realizations are made for dimensions  $p = 100$  and  $n = 200$ . It can be seen that sample matrices show statistical uncertainty because the number of observations and the number of variables is finite. The noise occurs naturally when we study cross-correlations of a large number of variables with a limited number of records. This condition is quite common in many research fields. For example, practitioners in finance prefer a high-dimensional setting, i.e.,  $p \sim n$ , to avoid nonstationary effects or structural changes in return time series of assets traded in financial markets.

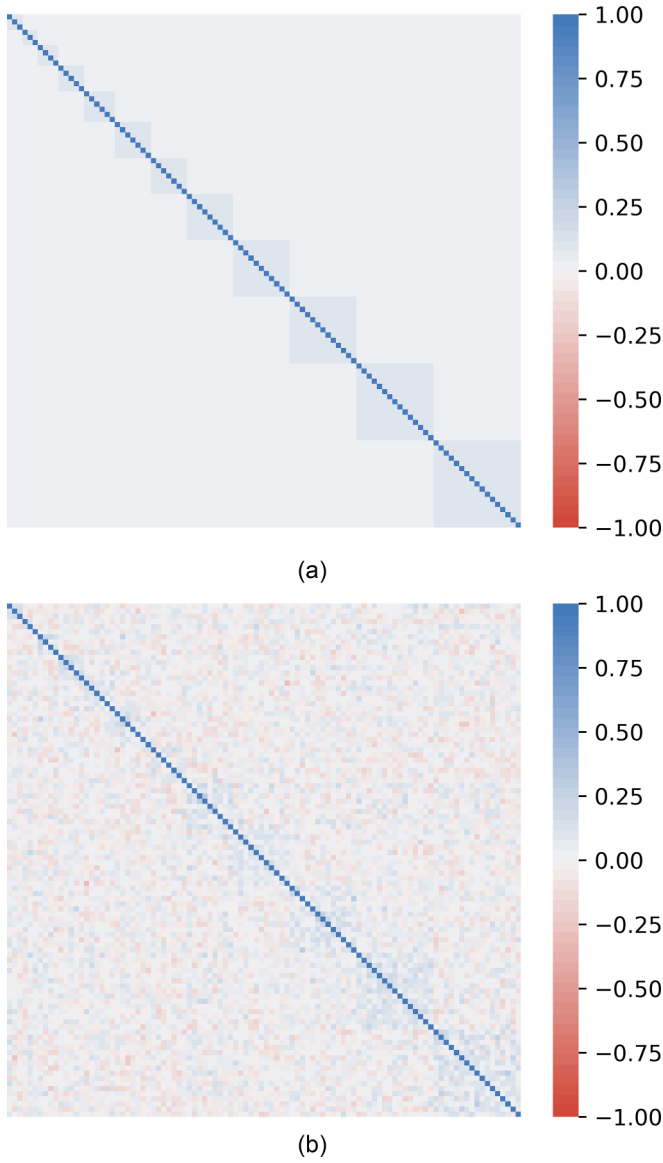


FIG. 2. Block diagonal model. (a) Population correlation matrix. (b) A single realization of such processes with autocorrelation matrix  $\mathbf{A} = \mathbf{I}$  and dimensions  $p = 100, n = 200$ .

Figure 4 shows the behavior of the average loss functions over  $m = 1000$  realizations and dimensions  $p = 100, n = 200$  for case 1 (blue), case 2 (green), and case 3 (brown). (a)–(f) show  $\langle \mathcal{L}(\mathbf{C}, \mathfrak{E}_i) \rangle$  vs  $\langle \mathcal{L}(\mathfrak{E}_i, \mathfrak{E}_j) \rangle$ . We denote by  $\mathcal{L}$  each of the loss functions (KL divergence, Frobenius, etc.),  $\langle \cdot \rangle$  represents the average and  $\mathfrak{E}$  represents the filtered correlation matrix under each of the filtering strategies described in Sec. II. The  $\mathfrak{E}^{\text{mwcV}}$  estimator is set with  $T_{\text{total}} = 10T$  and  $T_{\text{out}} = T = n$ . Moreover, the  $\mathfrak{E}^{\text{BJ}}$  estimator is set with  $\tau = 3$  [or, equivalently,  $\eta = \text{coth}(1/3) \approx 3.11$ ; see Eqs. (11) and (15)]. Under this setting, the  $\mathfrak{E}^{\text{BJ}}$  filter is expected to obtain optimal results for case 3, while the filter would be misspecified to deal with cases 1 and 2. Thus, we have omitted the results under the  $\mathfrak{E}^{\text{BJ}}$  and the related two-step (II) estimator for the latter cases. In these curves, we are comparing the value that minimizes the loss function (y axis) given the stability level of the estimators (x axis). The dotted

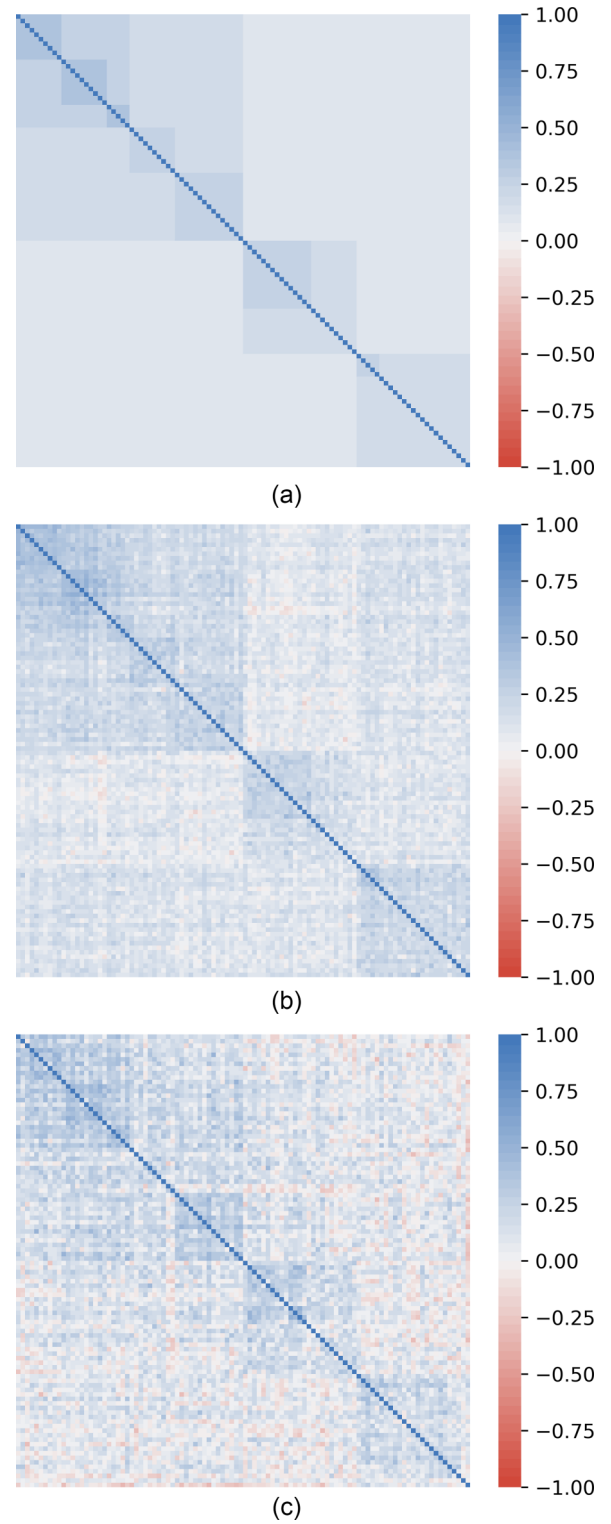


FIG. 3. Hierarchical nested model. (a) Population correlation matrix. (b) A single realization of such processes with autocorrelation matrix  $\mathbf{A} = \mathbf{I}$ . (c) A single realization of such processes with autocorrelation elements  $\mathbf{A}_{ij} = e^{-|i-j|/3}$ . The dimensions of the samples are  $p = 100, n = 200$ .

line represents the average RMT estimator when the number of eigenvalues  $\lambda$  that are kept in the filtering procedure varies from 1 to  $p$ , where the shadow band represents the standard deviation. The lower left corner corresponds to the

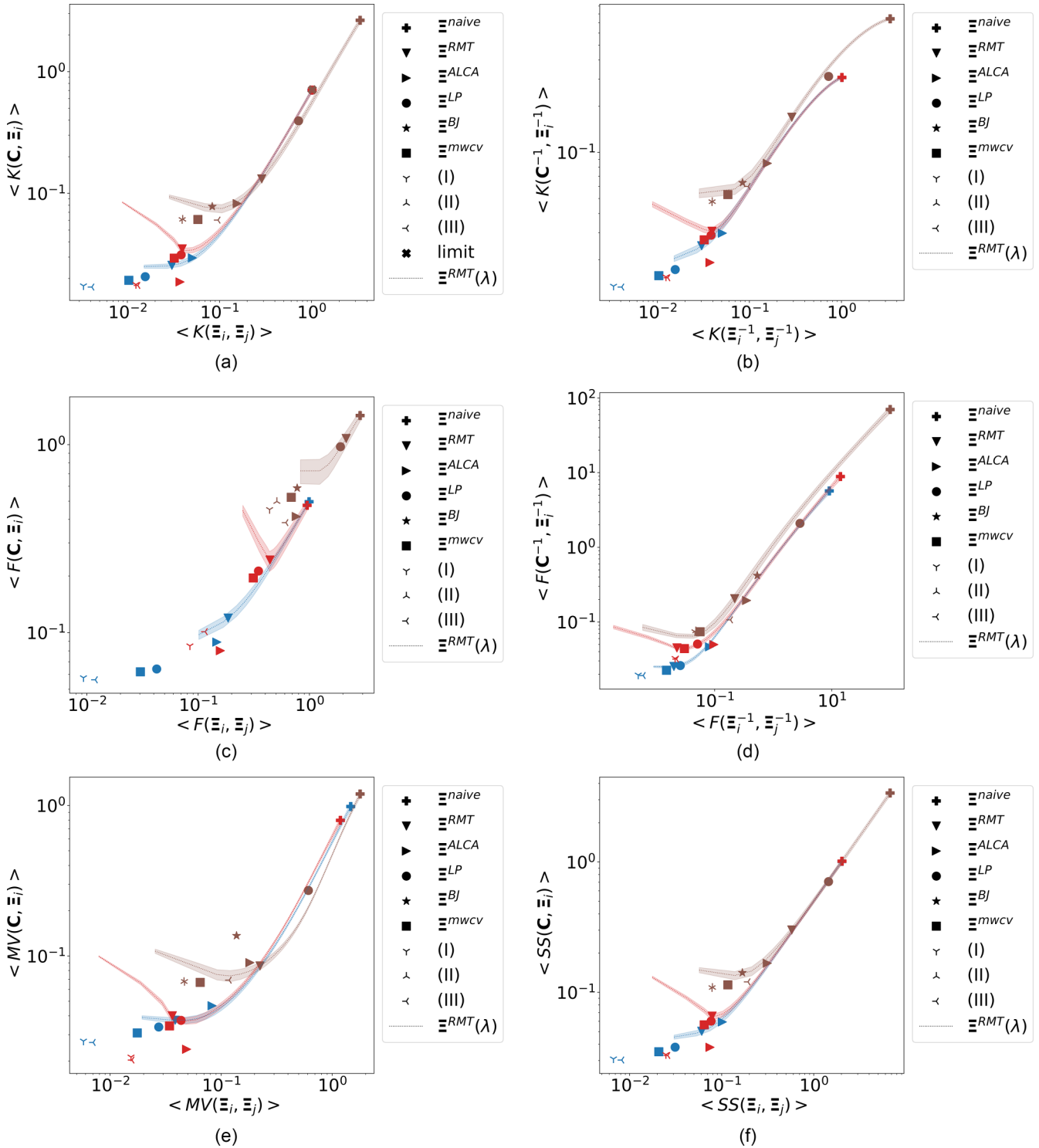


FIG. 4. Average loss functions over  $m = 1000$  realizations of the multiplicative noise model [see Eq. (27)] for dimensions  $p = 100, n = 200$ . The block diagonal model without memory (case 1) is represented by blue, the hierarchical nested model without memory (case 2) is represented by red, and the hierarchical nested model with memory (case 3) is represented by brown. (a)  $\langle K(\mathbf{C}, \Xi_i) \rangle$  vs  $\langle K(\Xi_i, \Xi_j) \rangle$ , where the cross-marker represents the theoretical limits given by Eq. (22). (b)  $\langle K(\mathbf{C}^{-1}, \Xi_i^{-1}) \rangle$  vs  $\langle K(\Xi_i^{-1}, \Xi_j^{-1}) \rangle$ . (c)  $\langle F(\mathbf{C}, \Xi_i) \rangle$  vs  $\langle F(\Xi_i, \Xi_j) \rangle$ . (d)  $\langle F(\mathbf{C}^{-1}, \Xi_i^{-1}) \rangle$  vs  $\langle F(\Xi_i^{-1}, \Xi_j^{-1}) \rangle$ . (e)  $\langle MV(\mathbf{C}, \Xi_i) \rangle$  vs  $\langle MV(\Xi_i, \Xi_j) \rangle$ . (f)  $\langle SS(\mathbf{C}, \Xi_i) \rangle$  vs  $\langle SS(\Xi_i, \Xi_j) \rangle$ . Both axes are on a logarithmic scale.

case where we have kept only the signal associated with the largest eigenvalue. The upper right corner corresponds to the extreme case where we have kept all the signals or eigenvalues, so the estimator is identical to the empirical cor-

relation matrix  $\mathbf{E}$ . Our simulations confirm that  $\langle K(\mathbf{C}, \Xi_i^{\text{naive}}) \rangle$  is in agreement with the theoretical limits of the KL divergence given by Eqs. (22) [represented by the cross-marker in Fig. 4(a)].

TABLE I. Block diagonal model without memory (case 1). Performance of estimators in terms of  $\langle \mathcal{L}(\mathbf{C}, \Xi_i) \rangle$ , where  $\mathcal{L}$  denotes the loss function and  $\langle \cdot \rangle$  represents the average over  $m = 1000$  realizations and considering dimensions  $p = 100, n = 200$ . In boldface we highlight the lowest value observed for each loss function.

	$\langle K(\mathbf{C}, \Xi_i) \rangle$	$\langle K(\mathbf{C}^{-1}, \Xi_i^{-1}) \rangle$	$\langle F(\mathbf{C}, \Xi_i) \rangle$	$\langle F(\mathbf{C}^{-1}, \Xi_i^{-1}) \rangle$	$\langle MV(\mathbf{C}, \Xi_i) \rangle$	$\langle SS(\mathbf{C}, \Xi_i) \rangle$
$\Xi^{\text{naive}}$	0.702978	0.307302	0.496612	5.682767	0.985036	1.010279
$\Xi^{\text{RMT}}$	0.025633	0.024646	0.119511	0.025346	0.037386	0.050279
$\Xi^{\text{ALCA}}$	0.029513	0.029777	0.089017	0.046473	0.046671	0.059290
$\Xi^{\text{LP}}$	0.020697	0.017244	0.064047	0.026045	0.033724	0.037942
$\Xi^{\text{mwcv}}$	0.019300	0.015717	0.061817	0.022541	0.030759	0.035017
Two-step (I)	0.017404	0.013441	0.057510	0.019467	0.027199	0.030845
Two-step (III)	<b>0.017012</b>	<b>0.013252</b>	<b>0.056114</b>	<b>0.019154</b>	<b>0.026599</b>	<b>0.030264</b>

The curves' behavior of the block diagonal model (blue color) is monotonically increasing almost for every value, except very near the origin. In contrast, the curves of the hierarchical nested model (red and brown colors) are monotonically increasing only relatively far from the origin. Interestingly, the RMT filter roughly coincides with the numerical minimum of  $\langle \mathcal{L}(\mathbf{C}, \Xi^{\text{RMT}}(\lambda)) \rangle$  (dotted lines) for all the metrics  $\mathcal{L}$  when no autocorrelations are considered (case 2). Thus, the Marchenko-Pastur bound effectively gives us the number of optimal signals to preserve in the hierarchical nested model without autocorrelation but fails to recover the true number of signals if the model violates the i.i.d. assumption (case 3). In general, we can see that the two-step estimators are the ones that obtain the optimal and most stable values within each case and for all the considered loss functions.

Tables I–III summarize the performance of estimators in terms of  $\langle \mathcal{L}(\mathbf{C}, \Xi_i) \rangle$  for the three studies (see Appendix A). The filter stability  $\langle \mathcal{L}(\Xi_i, \Xi_j) \rangle$  of each estimator  $\Xi$  in relation to the loss function  $\mathcal{L}$  is shown in Appendix A (see Tables IV–VI).

Table I shows the two-step (III) estimator minimizes all the loss functions for case 1. In other words, the best strategy is to apply the  $\Xi^{\text{LP}}$  estimator followed by the ALCA filter. The second best option is the two-step (I) estimator, which implies applying the estimator  $\Xi^{\text{mwcv}}$  followed again by the ALCA filter. Notably, in third place, and very close to the minimum values of the two-step estimators mentioned above, are the results of simply applying the strategy  $\Xi^{\text{mwcv}}$ .

The results for case 2 are similar for the two best performances. The exception concerns the Frobenius metric, where

now the ALCA filter beats the two-step estimator (III) performance and moves it to third place, with the two-step (I) estimator having the second-best performance against this metric. Actually, the ALCA estimator turn shifts  $\Xi^{\text{mwcv}}$  to obtain the third-best performance about the KL, inverse KL, MV, and SS loss functions. Only with the inverse Frobenius metric does the estimator  $\Xi^{\text{mwcv}}$  obtain third place.

For case 3, we have included the  $\Xi^{\text{BJ}}$  filter, which systematically beats the  $\Xi^{\text{LP}}$  filter as it should because it is calibrated with the same parameter  $\tau = 3$  of the generating process, although the best performance is disputed between the  $\Xi^{\text{mwcv}}$ , two-step (I), and two-step (III) filters depending on the loss function. The top three also include the two-step (II) and  $\Xi^{\text{ALCA}}$  filters under some metrics.

Figure 5 shows the average shrinkage eigenvalues ( $\xi$ ) as a function of the average empirical eigenvalues ( $\lambda$ ) for case 1 (a), case 2 (b), and case 3 (c) under each of the considered filters. We can see the single-step filters do not deal well with the extreme eigenvalues, and the two-step estimators somehow regularize the estimations. Notably, the misspecification of the  $\Xi^{\text{LP}}$  filter in case 3 presents a huge bias on almost the entire spectrum. In general, the bias of the smallest eigenvalue has severe consequences on the metrics that require inverting the correlation matrix because a near singular matrix could be obtained. Hence, the importance of correctly estimating these eigenvalues.

On the other hand, the behavior of the eigenvectors can be characterized by the inverse participation ratio (IPR) [36]. The IPR of the eigenvector  $v_i$  is defined as [7]  $\text{IPR}(v_i) = \sum_{j=1}^p [v_i^{(j)}]^4$ ; where  $v_i^{(j)}$  is the  $j$ th element of the eigenvector  $v_i$ . An eigenvector  $v_i$  located in only one component has the

TABLE II. Hierarchical nested model without memory (case 2). Performance of estimators in terms of  $\langle \mathcal{L}(\mathbf{C}, \Xi_i) \rangle$ , where  $\mathcal{L}$  denotes the loss function and  $\langle \cdot \rangle$  represents the average over  $m = 1000$  realizations and considering dimensions  $p = 100, n = 200$ . In boldface we highlight the lowest value observed for each loss function.

	$\langle K(\mathbf{C}, \Xi_i) \rangle$	$\langle K(\mathbf{C}^{-1}, \Xi_i^{-1}) \rangle$	$\langle F(\mathbf{C}, \Xi_i) \rangle$	$\langle F(\mathbf{C}^{-1}, \Xi_i^{-1}) \rangle$	$\langle MV(\mathbf{C}, \Xi_i) \rangle$	$\langle SS(\mathbf{C}, \Xi_i) \rangle$
$\Xi^{\text{naive}}$	0.704715	0.308027	0.475366	8.871255	0.796652	1.012742
$\Xi^{\text{RMT}}$	0.035000	0.030668	0.243333	0.044913	0.040020	0.065668
$\Xi^{\text{ALCA}}$	0.018728	0.019151	<b>0.080116</b>	0.049903	0.023976	0.037879
$\Xi^{\text{LP}}$	0.031284	0.028860	0.212684	0.050627	0.037316	0.060144
$\Xi^{\text{mwcv}}$	0.029365	0.026917	0.195309	0.043923	0.034257	0.056282
Two-step (I)	0.017844	0.015434	0.084999	0.032445	0.021296	0.033278
Two-step (III)	<b>0.017522</b>	<b>0.015251</b>	0.101058	<b>0.031783</b>	<b>0.020347</b>	<b>0.032772</b>



TABLE III. Hierarchical nested model with memory (case 3). Performance of estimators in terms of  $\langle \mathcal{L}(\mathbf{C}, \Xi_i) \rangle$ , where  $\mathcal{L}$  denotes the loss function and  $\langle \cdot \rangle$  represents the average over  $m = 1000$  realizations and considering dimensions  $p = 100, n = 200$ . In boldface we highlight the lowest value observed for each loss function.

	$\langle K(\mathbf{C}, \Xi_i) \rangle$	$\langle K(\mathbf{C}^{-1}, \Xi_i^{-1}) \rangle$	$\langle F(\mathbf{C}, \Xi_i) \rangle$	$\langle F(\mathbf{C}^{-1}, \Xi_i^{-1}) \rangle$	$\langle MV(\mathbf{C}, \Xi_i) \rangle$	$\langle SS(\mathbf{C}, \Xi_i) \rangle$
$\Xi^{\text{naive}}$	2.636596	0.741668	1.428074	70.128771	1.192546	3.378264
$\Xi^{\text{RMT}}$	0.131541	0.169630	1.078062	0.205236	0.085844	0.301171
$\Xi^{\text{ALCA}}$	0.082404	0.084835	0.414592	0.194150	0.090226	0.167239
$\Xi^{\text{LP}}$	0.394159	0.312629	0.974296	2.093030	0.272487	0.706789
$\Xi^{\text{BJ}}$	0.077930	0.063499	0.587735	0.419335	0.136441	0.141428
$\Xi^{\text{mwcv}}$	0.060966	0.053122	0.524506	<b>0.073864</b>	<b>0.066712</b>	0.114087
Two-step (I)	0.060880	<b>0.047672</b>	0.451754	0.074127	0.067626	<b>0.108552</b>
Two-step (II)	0.061383	0.047843	0.498786	0.076085	0.067689	0.109226
Two-step (III)	<b>0.060302</b>	0.059961	<b>0.384679</b>	0.106745	0.069054	0.120263

upper bound  $\text{IPR}(v_i) = 1$ , while an eigenvector uniformly distributed over the  $p$  components has the lower bound  $\text{IPR}(v_i) = 1/p$ . Figure 6 shows the average IPR as a function of the  $i$ th eigenvector of the filtered correlation matrix for case 1 (a), case 2 (b), and case 3 (c) under each of the considered filters. It can be observed that the IPR of the eigenvectors related to the filters that fall into the RIE family present a stable behavior with a uniform distribution of its elements, which is natural since these filters assume that the eigenvectors do not change. The small fluctuations are due to the normalization effect to obtain orthonormal eigenvectors after reconstructing the correlation matrix. On the contrary, the behavior of the eigenvectors that involve the ALCA filter is more localized and is closer to the eigenvectors of the population model (black line).

## VI. DISCUSSION

In principle, one would expect that asymptotic estimators based on random matrices and free probability perform better as the dimension of the correlation matrix increases. However, we have found at least one case where the hierarchical estimators perform better than the RIE estimators, even at  $p = 500$  [37]. This behavior can be explained due to the assumptions of the semianalytical solution of the nonlinear shrinkage function proposed by Ledoit and Wolf. Their expression was also used in the Burda and Jarosz approach and our analysis. The central assumption of the solution is the existence of a compact interval that contains all the eigenvalues as the matrix dimensions tend to infinity. In other words, the eigenvalues should not grow with the dimension to converge to a well-defined density. However, diagonal block and hierarchical nested models have one eigenvalue that grows with the dimension of each of their blocks (see Appendix B). That is, a model of  $k$  blocks has  $k$  unbounded eigenvalues, which violates the assumptions of the asymptotic results of the RIE solutions. More precisely, the expressions in Eqs. (7) and (11) are correct and valid as long as  $p, n \rightarrow \infty$ . What is problematic is the kernel approximation of the density  $\rho_E$  and the Hilbert transform  $h_E$  since they do not converge for our block structure models. Hence, our models violate this principle. Thereupon, the hierarchical clustering and two-step estimators can give better estimates regarding optimality and stability.

Analyzing eigenvalues and eigenvectors reveals why the selected two-step filters perform well against most metrics. If we change the order in constructing these estimators, we lose the regularizing effect by filtering the high-dimensional noise before applying the clustering methods. We have verified that estimating the smallest eigenvalues get worse by inverting the order of the single estimators composing the two-step filters. This behavior is particularly noticeable for the inverse KL and inverse Frobenius metrics since they can be written as a function of the inverse of the eigenvalues. Therefore, if these are very close to zero, we obtain metrics that tend to infinity or indeterminate. On the part of the eigenvectors, the models that involve the hierarchical estimators modify the distribution of their elements and bring them closer to the population behavior qualitatively. Nevertheless, block models present eigenvalues' multiplicity, and the set of eigenvectors is not unique. There may be slightly different solutions depending on the algorithm to compute them. Thus, a future question to explore is to what extent it is possible to filter the correlation matrix by modifying only the eigenvalues.

On the other hand, the excellent performance of the data-driven estimator  $\Xi^{\text{mwcv}}$  is notable. We have seen that the  $\Xi^{\text{mwcv}}$  filter outperforms the two-step filters in case 3 under some metrics. A preliminary explanation is that the  $\Xi^{\text{mwcv}}$  filter can capture autocorrelations due to its construction as a time-varying estimator. Then, the correlation matrix of a nonstationary financial time series might be better estimated by the  $\Xi^{\text{mwcv}}$  and two-step (I) filters. Furthermore, the authors of Ref. [31] proved that it is possible to approximate the optimal RIE estimator  $\xi(\lambda) = l$  (the true eigenvalues) by overlapping the eigenvectors of two different realizations of the same population covariance matrix  $\Sigma$ —even valid if the test sample covariance matrix can be rank deficient, i.e.,  $n = T_{\text{out}} < p$ . Intuitively, the superposition of the training and testing eigenvectors helps estimate the empirical eigenvalues, as if rotating them into the test direction unveils their true value. This evidence opens the door to considering other types of nonlinear shrinkage  $\xi(\lambda)$  under the RIE approach.

An interesting future work would consider statistical learning models to shape the function  $\xi(\lambda)$  and consider the stylized fact of heterogeneous structures in financial correlation matrices under more general distributional assumptions. Moreover, the nonlinear shrinkage functions  $\Xi^{\text{LP}}$  and  $\Xi^{\text{BJ}}$  are optimal concerning the Frobenius loss function. Then,

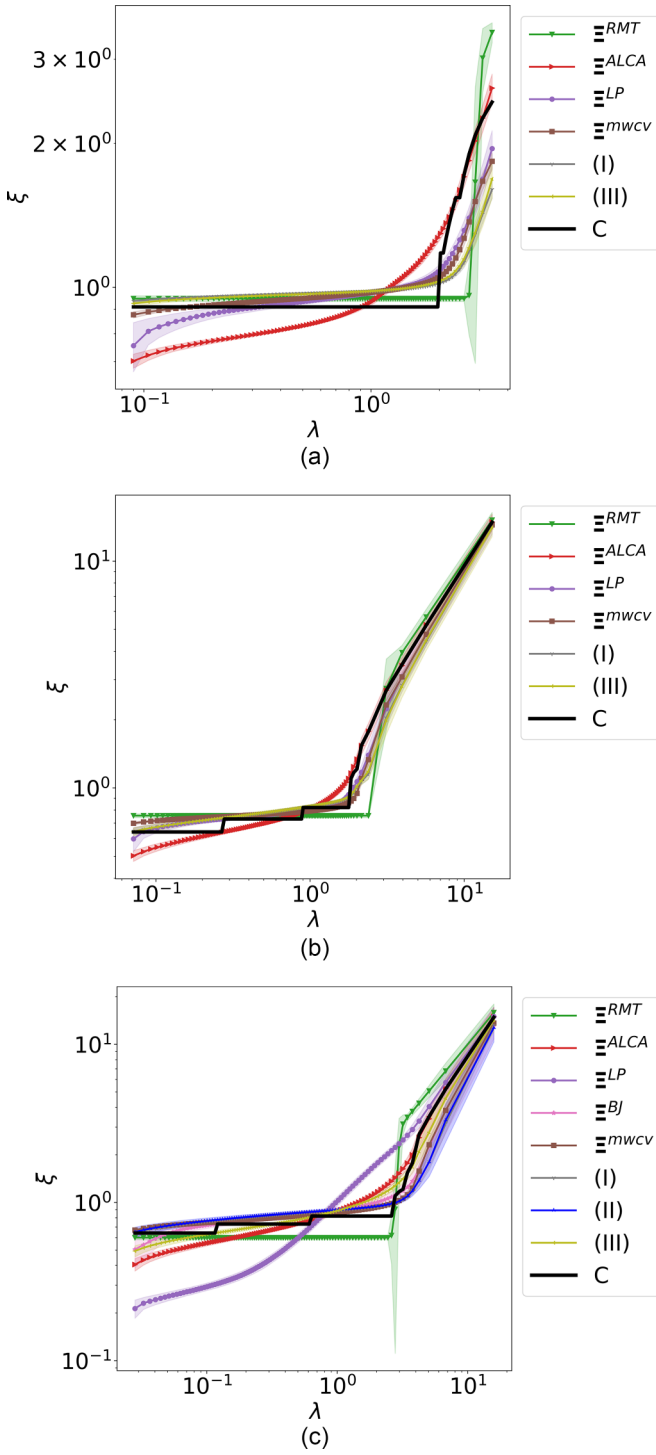


FIG. 5. Average shrinkage eigenvalues ( $\xi$ ) vs average empirical eigenvalues ( $\lambda$ ). (a) Case 1. (b) Case 2. (c) Case 3. The black line represents the corresponding population model  $C$ . The shadow band represents one standard deviation. Both axes are on logarithmic scales.

further future work could also go in the direction of analyzing the performance of the block diagonal and hierarchical nested models under a nonlinear shrinkage formula optimized having as a target the loss function used to evaluate their performance and in the spirit of the proposed expressions in Refs. [24,34,38].

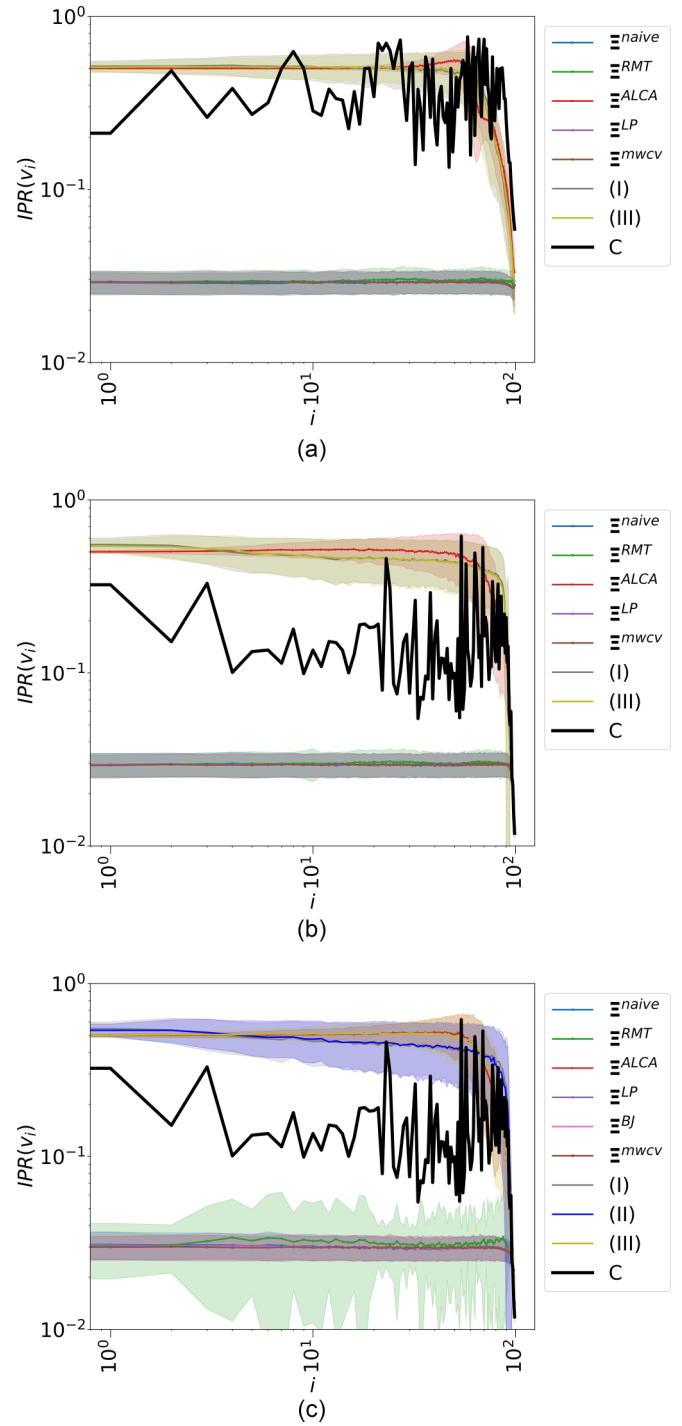


FIG. 6. Average IPR as a function of the rank of the  $i$ th eigenvector of the filtered correlation matrix. (a) Case 1. (b) Case 2. (c) Case 3. The black line represents the corresponding population model  $C$ . The shadow band represents one standard deviation. Both axes are on logarithmic scales. The rank of eigenvectors runs from 1 (smallest eigenvalue) to 100 (largest eigenvalue).

**ACKNOWLEDGMENTS**

S.M. and R.N.M. acknowledge financial support by the MIUR PRIN Project No. 2017WZFTZP, Stochastic forecasting in complex systems. A.G.M. acknowledge financial

TABLE IV. Block diagonal model without memory (case 1). Stability of estimators in terms of  $\langle \mathcal{L}(\Xi_i, \Xi_j) \rangle$ , where  $\mathcal{L}$  denotes the loss function and  $\langle \cdot \rangle$  represents the average over the realizations of  $m = 1000$  and considering dimensions  $p = 100, n = 200$ . In boldface we highlight the lowest value observed for each loss function.

	$\langle K(\Xi_i, \Xi_j) \rangle$	$\langle K(\Xi_i^{-1}, \Xi_j^{-1}) \rangle$	$\langle F(\Xi_i, \Xi_j) \rangle$	$\langle F(\Xi_i^{-1}, \Xi_j^{-1}) \rangle$	$\langle MV(\Xi_i, \Xi_j) \rangle$	$\langle SS(\Xi_i, \Xi_j) \rangle$
$\Xi$ naive	1.011480	1.011512	0.993376	9.102604	1.469460	2.022991
$\Xi$ RMT	0.030454	0.030248	0.185746	0.020132	0.038568	0.060702
$\Xi$ ALCA	0.050803	0.050802	0.147216	0.081805	0.082820	0.101605
$\Xi$ LP	0.015637	0.015614	0.042401	0.025937	0.027382	0.031251
$\Xi$ mwcv	0.010394	0.010385	0.030306	0.015103	0.017561	0.020779
Two-step (I)	<b>0.003341</b>	<b>0.003339</b>	<b>0.009333</b>	<b>0.004962</b>	<b>0.005754</b>	<b>0.006680</b>
Two-step (III)	0.004123	0.004108	0.011924	0.006024	0.007032	0.008231

support by Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) Project No. A1-S-43514.

APPENDIX A

Tables IV–VI show the average stability  $\langle \mathcal{L}(\Xi_i, \Xi_j) \rangle$  of each estimator  $\Xi$  in relation to the loss function  $\mathcal{L}$  (see Tables I–III in the main text.)

APPENDIX B

1. Top eigenvalues of diagonal block and hierarchical nested models

a. Diagonal block model

Consider a block diagonal matrix  $\mathbf{A}$  of dimension  $p \times p$  with  $b$  blocks  $\mathbf{A}_l (l = 1, \dots, b)$ , each of dimensions  $p_l \times p_l$  satisfying  $\sum_l p_l = p$ :

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & 0 & \dots & 0 \\ 0 & \mathbf{A}_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \mathbf{A}_b \end{pmatrix}. \tag{B1}$$

Let each block  $\mathbf{A}_l$  be of the form

$$\mathbf{A}_l = \begin{pmatrix} 1 & a^{(l)} & \dots & a^{(l)} \\ a^{(l)} & 1 & \dots & a^{(l)} \\ \dots & \dots & \dots & \dots \\ a^{(l)} & a^{(l)} & a^{(l)} & 1 \end{pmatrix}, \tag{B2}$$

where  $a^{(l)} \in [0, 1]$ . The characteristic polynomial of  $\mathbf{A}_l$  is found to be  $\det(\mathbf{A}_l - \lambda \mathbf{I}) = (1 - a^{(l)} - \lambda)^{p_l - 1} (1 + (p_l - 1)a^{(l)} - \lambda) = 0$ . Thus, the eigenvalues of block  $\mathbf{A}_l$  are given by

$$\lambda = \begin{cases} 1 + a^{(l)}(p_l - 1); & \text{with multiplicity } 1 \\ \lambda = 1 - a^{(l)}; & \text{with multiplicity } p_l - 1. \end{cases} \tag{B3}$$

The eigenvalues of  $\mathbf{A}$  are the combined eigenvalues of each block due to the property

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \det(\mathbf{A}_1 - \lambda \mathbf{I}) \dots \det(\mathbf{A}_b - \lambda \mathbf{I}). \tag{B4}$$

Therefore, there are  $b$  eigenvalues of  $\mathbf{A}$  that grow with the dimension of their blocks at the rate  $p_l$ . Consequently,  $b$  eigenvalues are not bounded when  $p \rightarrow \infty$ .

In addition, we can notice that  $\mathbf{A}$  is reducible because there does not exist a directed path between the blocks in the associated directed graph  $G(\mathbf{A})$ , that is,  $G(\mathbf{A})$  is not strongly connected [39]. Nevertheless, each directed subgraph  $G(\mathbf{A}_l)$  is strongly connected given that  $\mathbf{A}_l > \mathbf{O}$ . Then, each block matrix  $\mathbf{A}_l$  is irreducible, and either

$$\sum_{j=1}^{p_l} [a_{ij}]_l = \rho(\mathbf{A}_l) \quad \text{for all } 1 \leq i \leq p_l \tag{B5}$$

or

$$\min_{1 \leq i \leq p_l} \left( \sum_{j=1}^{p_l} [a_{ij}]_l \right) < \rho(\mathbf{A}_l) < \max_{1 \leq i \leq p_l} \left( \sum_{j=1}^{p_l} [a_{ij}]_l \right), \tag{B6}$$

where  $[a_{ij}]_l$  are the  $(i, j)$ th elements of  $\mathbf{A}_l$ , and  $\rho(\mathbf{A}_l)$  is its spectral radius. Further, the sum of each row of  $\mathbf{A}_l$  is the same,

TABLE V. Hierarchical nested model without memory (case 2). Stability of estimators in terms of  $\langle \mathcal{L}(\Xi_i, \Xi_j) \rangle$ , where  $\mathcal{L}$  denotes the loss function and  $\langle \cdot \rangle$  represents the average over the realizations of  $m = 1000$  and considering dimensions  $p = 100, n = 200$ . In boldface we highlight the lowest value observed for each loss function.

	$\langle K(\Xi_i, \Xi_j) \rangle$	$\langle K(\Xi_i^{-1}, \Xi_j^{-1}) \rangle$	$\langle F(\Xi_i, \Xi_j) \rangle$	$\langle F(\Xi_i^{-1}, \Xi_j^{-1}) \rangle$	$\langle MV(\Xi_i, \Xi_j) \rangle$	$\langle SS(\Xi_i, \Xi_j) \rangle$
$\Xi$ naive	1.015298	1.013145	0.952442	14.119574	1.188050	2.028444
$\Xi$ RMT	0.039437	0.039263	0.441274	0.022705	0.036337	0.078700
$\Xi$ ALCA	0.037348	0.037287	0.158198	0.096944	0.048786	0.074635
$\Xi$ LP	0.038438	0.038397	0.348333	0.051018	0.043579	0.076836
$\Xi$ mwcv	0.032339	0.032347	0.313248	0.030515	0.034258	0.064685
Two-step (I)	<b>0.012385</b>	<b>0.012390</b>	<b>0.084258</b>	<b>0.021132</b>	<b>0.015412</b>	<b>0.024775</b>
Two-step (III)	0.013008	0.012994	0.115886	0.022353	0.015728	0.026002

TABLE VI. Hierarchical nested model with memory (case 3). Stability of estimators in terms of  $\langle \mathcal{L}(\Xi_i, \Xi_j) \rangle$ , where  $\mathcal{L}$  denotes the loss function and  $\langle \cdot \rangle$  represents the average over the realizations of  $m = 1000$  and considering dimensions  $p = 100, n = 200$ . In boldface we highlight the lowest value observed for each loss function.

	$\langle K(\Xi_i, \Xi_j) \rangle$	$\langle K(\Xi_i^{-1}, \Xi_j^{-1}) \rangle$	$\langle F(\Xi_i, \Xi_j) \rangle$	$\langle F(\Xi_i^{-1}, \Xi_j^{-1}) \rangle$	$\langle MV(\Xi_i, \Xi_j) \rangle$	$\langle SS(\Xi_i, \Xi_j) \rangle$
$\Xi$ naive	3.377219	3.384450	2.854999	99.721840	1.788246	6.761669
$\Xi$ RMT	0.288290	0.288862	2.137602	0.220317	0.223823	0.577152
$\Xi$ ALCA	0.156065	0.156478	0.760417	0.351341	0.181564	0.312543
$\Xi$ LP	0.723069	0.724795	1.899020	2.867135	0.610633	1.447864
$\Xi$ BJ	0.083587	0.084194	0.775484	0.534801	0.137637	0.167781
$\Xi$ mwcv	0.058353	0.058390	0.686160	0.056490	0.064800	0.116743
Two-step (I)	<b>0.039335</b>	<b>0.039237</b>	<b>0.438580</b>	<b>0.046152</b>	0.046753	<b>0.078572</b>
Two-step (II)	0.039901	0.039489	0.509946	0.047698	<b>0.046319</b>	0.079389
Two-step (III)	0.096686	0.096735	0.615370	0.186985	0.118327	0.193421

then the minimum and maximum is equal. Hence, we have

$$\rho(\mathbf{A}_l) = \sum_{j=1}^{p_l} [a_{ij}]_l = 1 + (p_l - 1)a^{(l)}. \quad (\text{B7})$$

Moreover, the generalization of the Perron-Frobenius theorem assures that  $\mathbf{A}$  has a nonnegative real eigenvalue equal to its spectral radius. Therefore, one of the spectral radii  $\rho(\mathbf{A}_l)$ ,  $l = 1, \dots, b$ , is the spectral radius of  $\mathbf{A}$ . It can be corroborated that Eq. (B7) coincides with the first part of Eq. (B3).

### b. Hierarchical nested model

We have something similar for the hierarchical nested model. In this case, the number of independent blocks is reduced. However, each of them is irreducible by the same argument given above. Let  $\mathbf{A}_k$  be an independent hierarchical nested block, where  $\sum_{k=1}^c p_k = p$  ( $k = 1, \dots, c$ ), such that  $c < b$ . In other words, each independent hierarchical block is composed of several overlapping blocks. We have by construction

$$\min_{1 \leq i \leq n} \left( \sum_{j=1}^n [a_{ij}]_k \right) = 1 + (p_k - 1)a^{(k)}, \quad (\text{B8})$$

$$\max_{1 \leq i \leq n} \left( \sum_{j=1}^n [a_{ij}]_k \right) = 1 + p_k(p_k - 1)a^{(k)}. \quad (\text{B9})$$

The minimum is reached when no overlapping exists, and the model is reduced to the diagonal block model. The maximum is reached when each internal block overlaps with each other, and as we can have at most  $p_k$  blocks, a factor  $p_k$  appears. Then, Eqs. (B5) and (B6) apply, and the bounds for the spectral radius of each independent block are

$$1 + (p_k - 1)a^{(k)} < \rho(\mathbf{A}_k) < 1 + p_k(p_k - 1)a^{(k)}. \quad (\text{B10})$$

Again, the generalization of the Perron-Frobenius theorem assures that  $\mathbf{A}$  has a nonnegative real eigenvalue equal to its spectral radius. Thus, one of the spectral radius  $\rho(\mathbf{A}_k)$ ,  $k = 1, \dots, c$ , is the spectral radius of  $\mathbf{A}$ . Therefore,  $c$  eigenvalues of  $\mathbf{A}$  grow with the dimension of their blocks at the rate  $p_k$  (at least). Consequently,  $c$  eigenvalues are not bounded when  $p \rightarrow \infty$ .

### c. Observations

We observe that the number of independent blocks in the hierarchical nested model is less than the number in the diagonal block model, i.e.,  $c < b$ . Consequently, the block's size of the former should be bigger to satisfy  $\sum_{k=1}^c p_k = \sum_{l=1}^b p_l = p$ . Therefore,  $p_k > p_l$ , and the top eigenvalue of the hierarchical nested model grows faster than the top eigenvalue of the diagonal block model.

In our models  $a^{(k)} = a^{(l)} = \gamma^2 = (0.3)^2 = 0.09$ , the diagonal block model has  $b = 12$  diagonal blocks, while the hierarchical nested model has  $c = 3$  independent (nonoverlapping) blocks. Then it is clear that the top eigenvalue of the latter grows faster to infinity than the former as  $p \rightarrow \infty$ .

- [1] V. A. Marčenko and L. A. Pastur, Distribution of eigenvalues for some sets of random matrices, *Math. USSR-Sbornik* **1**, 457 (1967).
- [2] J. Bun, J.-P. Bouchaud, and M. Potters, Cleaning large correlation matrices: Tools from random matrix theory, *Phys. Rep.* **666**, 1 (2017).
- [3] 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).
- [4] M. Tumminello, F. Lillo, and R. N. Mantegna, Shrinkage and spectral filtering of correlation matrices: A comparison via the Kullback-Leibler distance, *Acta Polonica B* **38**, 4079 (2007).
- [5] M. Tumminello, F. Lillo, and R. N. Mantegna, Hierarchically nested factor model from multivariate data, *Europhys. Lett.* **78**, 30006 (2007).
- [6] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Noise dressing of financial correlation matrices, *Phys. Rev. Lett.* **83**, 1467 (1999).
- [7] V. Plerou, P. Gopikrishnan, B. Rosenow, Luis A. Nunes Amaral, T. Guhr, and H. E. Stanley, Random matrix approach to cross correlations in financial data, *Phys. Rev. E* **65**, 066126 (2002).
- [8] I. M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, *Ann. Statist.* **29**, 295 (2001).

- [9] C. Stein, Rietz lecture, Estimation of a covariance matrix, in *39th Annual Meeting IMS, Atlanta, GA, 1975* (1975).
- [10] O. Ledoit and M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, *J. Multivariate Anal.* **88**, 365 (2004).
- [11] O. Ledoit and S. Péché, Eigenvectors of some large sample covariance matrix ensembles, *Probab. Theory Relat. Fields* **151**, 233 (2011).
- [12] O. Ledoit and M. Wolf, Numerical implementation of the QuEST function, *Comput. Stat. Data Anal.* **115**, 199 (2017).
- [13] O. Ledoit and M. Wolf, Analytical nonlinear shrinkage of large-dimensional covariance matrices, *Ann. Stat.* **48**, 3043 (2020).
- [14] Z. Burda and A. Jarosz, Cleaning large-dimensional covariance matrices for correlated samples, *Phys. Rev. E* **105**, 034136 (2022).
- [15] D. L. Donoho, M. Gavish, and I. M. Johnstone, Optimal shrinkage of eigenvalues in the spiked covariance model, *Ann. Statist.* **46**, 1742 (2018).
- [16] F. Lillo and R. N. Mantegna, Spectral density of the correlation matrix of factor models: A random matrix theory approach, *Phys. Rev. E* **72**, 016219 (2005).
- [17] M. Tumminello, F. Lillo, and R. N. Mantegna, Spectral properties of correlation matrices for some hierarchically nested factor models, in *AIP Conference Proceedings*, Vol. 965 (American Institute of Physics, Melville, New York, 2007), pp. 300–307.
- [18] M. Tumminello, F. Lillo, and R. N. Mantegna, Correlation, hierarchies, and networks in financial markets, *J. Econ. Behav. Org.* **75**, 40 (2010).
- [19] V. Tola, F. Lillo, M. Gallegati, and R. N. Mantegna, Cluster analysis for portfolio optimization, *J. Econ. Dyn. Control* **32**, 235 (2008).
- [20] E. Pantaleo, M. Tumminello, F. Lillo, and R. N. Mantegna, When do improved covariance matrix estimators enhance portfolio optimization? An empirical comparative study of nine estimators, *Quant. Financ.* **11**, 1067 (2011).
- [21] C. Bongiorno, D. Challet, and G. Loeper, Cleaning the covariance matrix of strongly nonstationary systems with time-independent eigenvalues, [arXiv:2111.13109](https://arxiv.org/abs/2111.13109).
- [22] C. Bongiorno and D. Challet, Covariance matrix filtering with bootstrapped hierarchies, *PloS one* **16**, e0245092 (2021).
- [23] C. Bongiorno and D. Challet, Reactive global minimum variance portfolios with  $k$ -BAHC covariance cleaning, *Eur. J. Finance* **28**, 1344 (2022).
- [24] O. Ledoit and M. Wolf, Shrinkage estimation of large covariance matrices: Keep it simple, statistician? *J. Multivariate Anal.* **186**, 104796 (2021).
- [25] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis* (Springer-Verlag, Berlin, Heidelberg, 2005).
- [26] J. A. Mingo and R. Speicher, *Free Probability and Random Matrices*, Vol. 35 (Springer, New York, NY, 2017).
- [27] M. Potters and J.-P. Bouchaud, *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists* (Cambridge University Press, Cambridge, UK, 2020).
- [28] Z. Burda, J. Jurkiewicz, and B. Waclaw, Spectral moments of correlated Wishart matrices, *Phys. Rev. E* **71**, 026111 (2005).
- [29] Z. Burda, A. Jarosz, M. A. Nowak, J. Jurkiewicz, G. Papp, and I. Zahed, Applying free random variables to random matrix analysis of financial data. Part I: The Gaussian case, *Quant. Financ.* **11**, 1103 (2011).
- [30] D. Bartz, Cross-validation based nonlinear shrinkage, [arXiv:1611.00798](https://arxiv.org/abs/1611.00798).
- [31] J. Bun, J.-P. Bouchaud, and M. Potters, Overlaps between eigenvectors of correlated random matrices, *Phys. Rev. E* **98**, 052145 (2018).
- [32] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis* (Pearson Prentice Hall, Upper Saddle River, NJ, 2002).
- [33] M. R. Anderberg, *Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks* (Academic Press, New York, 2014).
- [34] O. Ledoit and M. Wolf, Optimal estimation of a large-dimensional covariance matrix under Stein’s loss, *Bernoulli* **24**, 3791 (2018).
- [35] R. F. Engle, O. Ledoit, and M. Wolf, Large dynamic covariance matrices, *J. Bus. Econ. Stat.* **37**, 363 (2019).
- [36] W. Visscher, Localization of electron wave functions in disordered systems, *J. Non-Cryst. Solids* **8-10**, 477 (1972).
- [37] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevE.108.044137> for the results with  $p = 500$ ,  $n = 1000$ , and  $m = 100$  realizations of case 3, where we have additionally included the linear shrinkage estimator and the single linkage clustering analysis (SLCA) estimator.
- [38] O. Ledoit and M. Wolf, The power of (non)linear shrinking: A review and guide to covariance matrix estimation, *J. Fin. Econ.* **20**, 187 (2022).
- [39] R. S. Varga, *Matrix Iterative Analysis*, 2nd ed. (Springer-Verlag, Berlin, 2000).