Noise cleaning the precision matrix of short time series

Miguel Ibáñez-Berganza [*]
Networks Unit, IMT School for Advanced Studies Lucca, Piazza San Francesco 19, 50100 Lucca, Italy
and Istituto Italiano di Tecnologia. Largo Barsanti e Matteucci, 53, 80125 Napoli, Italy

Carlo Lucibello
AI Lab, Institute for Data Science and Analytics, Bocconi University, 20136 Milano, Italy

Francesca Santucci and Tommaso Gili
Networks Unit, IMT School for Advanced Studies Lucca, Piazza San Francesco 19, 50100 Lucca, Italy

Andrea Gabrielli
Dipartimento di Ingegneria Civile, Informatica e delle Tecnologie Aeronautiche, Universitá degli Studi Roma Tre,
Via Vito Volterra 62, 00146 Rome, Italy
and Centro Ricerche Enrico Fermi, Via Panisperna 89a, 00184 Rome, Italy

(Received 19 April 2023; accepted 2 August 2023; published 28 August 2023)

We present a comparison between various algorithms of inference of covariance and precision matrices in small data sets of real vectors of the typical length and dimension of human brain activity time series retrieved by functional magnetic resonance imaging (fMRI). Assuming a Gaussian model underlying the neural activity, the problem consists of denoising the empirically observed matrices to obtain a better estimator of the (unknown) true precision and covariance matrices. We consider several standard noise-cleaning algorithms and compare them on two types of data sets. The first type consists of synthetic time series sampled from a generative Gaussian model of which we can vary the fraction of dimensions per sample q and the strength of off-diagonal correlations. The second type consists of time series of fMRI brain activity of human subjects at rest. The reliability of each algorithm is assessed in terms of test-set likelihood and, in the case of synthetic data, of the distance from the true precision matrix. We observe that the so-called optimal rotationally invariant estimator, based on random matrix theory, leads to a significantly lower distance from the true precision matrix in synthetic data and higher test likelihood in natural fMRI data. We propose a variant of the optimal rotationally invariant estimator in which one of its parameters is optimzed by cross-validation. In the severe undersampling regime (large q) typical of fMRI series, it outperforms all the other estimators. We furthermore propose a simple algorithm based on an iterative likelihood gradient ascent, leading to very accurate estimations in weakly correlated synthetic data sets.

DOI: 10.1103/PhysRevE.108.024313

I. INTRODUCTION

Multiple, complex, rapidly changing brain activity patterns are constrained by an underlying structural connectivity (SC) network of neuronal fiber bundles that evolve on distinctly larger timescales [1–4]. It is precisely this degeneration (reminiscent of collective phases in physics) of emerging complex and segregated dynamical states, subtended by a relatively static and sparse SC network, that makes possible context-sensitive conscious cognition, perception, and action [5–7]. For this reason, the study of the relation between brain structure and the associated emergent cognitive functions in healthy and diseased subjects is, arguably, one of the most important challenges in neuroscience. Since the advent of high-quality functional magnetic resonance imaging (fMRI) and electro- and magnetoencephalography (EEG and MEG) data sets, a wide range of linear [2,8–11] and nonlinear [12–22] models explaining the emergent function in terms of a latent, subtending model of *effective connectivity* [4] have been considered.

In this context, the role played by *correlation* and *precision matrices* of brain activity patterns is receiving increasing interest [11,23–32]. Functional connectivity (FC) is usually estimated through the correlation matrix (C) among pairs of functional time series of activity (usually blood oxygen level dependent (BOLD) fMRI signals *at rest*) corresponding to different brain areas. A related quantity, considered as an alternative estimator of FC (see, for example, Ref. [11]), is the *precision matrix*, or the inverse of the correlation matrix $J = C^{-1}$. Assuming that the vector of brain activity patterns obeys Gaussian statistics [2,8,9,11] (or that the time series follow an Ornstein-Uhlenbeck process [3,33]), the precision matrix J represents harmonic coupling constants that constrain the emergent correlations C. For this reason, J may be understood as an (inferred) model of more direct [11] anatomical connections between brain areas. Indeed, differently from C, and within a Gaussian approximation, J accounts for direct causal relations only. As a matter of fact, in the context of the SC-FC relationship, the inferred

[*] Corresponding author: miguel.ibanezberganza@imtlucca.it

precision matrix J has been demonstrated to be a more accurate statistical estimator of SC (as retrieved, e.g., by diffusion tensor imaging techniques) than C and than the interarea couplings resulting from Granger causality and autoregressive inference ([11,24] and references therein). Furthermore, J has also been shown to provide better prediction scores than correlation-based FC for some diseases and nonimaging phenotypic measures [25–28] (see also Ref. [29]), and to better capture intrasubject FC differences [30]. Finally, recent results suggest [11] that the relation between the empirical SC matrix and the precision matrix from temporal BOLD series can be exploited, beyond the Ornstein-Uhlenbeck hypothesis, to infer the relative timescales of temporal correlation of the BOLD activity in different brain subnetworks of grey matter, that in turn reflect the relative complexity of cognitive functions involved in the cortical hierarchy.

The studies of precision matrices are, however, severely limited by the limited accuracy of their statistical estimation due to the short length of time series [11,27,29,31,32,34]. To overcome this issue, several techniques of statistical inference of the correlation (and, hence, of the precision) matrix have been proposed in the context of network neuroscience. These are Ledoit-Wolf and Tikhonov regularized precision matrix [11,25,32,35,36], L_1-regularization with or without population priors [31,37], addition of regularized aggregation to construct the group precision matrix [29], and (nonisotropic) population-shrinkage covariance estimators [30] (see further references in Refs. [25,29,30]). Accurate strategies for correlation and precision matrix regularization, or noise cleaning, are also crucial to improve the efficiency of autoregressive models of causal inference [11] since they rely on the inversion of sample covariance estimators constructed from a few data vectors (please note that we do not use *clean* in the sense of denoising of the fMRI primary data as in Ref. [38] but, rather, in the sense of noise cleaning the correlation and precision matrices, as in Ref. [39]).

The previous discussion highlights the need for accurate benchmarking of estimators of FC. In the present contribution, we compare standard noise-cleaning methods [such as linear shrinkage and principal component analysis (PCA)] on correlation matrices derived from short time series of the typical size and length of fMRI data. Importantly, we add to the comparison recent methods grounded on random matrix theory [40]. Despite such methods having been known in statistical physics and in theoretical finance for a few years, they have not been, to the best of our knowledge, benchmarked nor just employed in the study of the human brain connectome. The algorithms are compared in two types of data sets: *synthetic* ones, sampled from a known generative Gaussian model, and two *natural* data sets of human resting-state brain activity by fMRI. We evaluate the efficiency of each algorithm on each data set in terms of the out-of-sample (test) likelihood, that takes into account both variance and bias errors, and of two related criteria. In the case of synthetic data sets, the efficiency of the methods is further evaluated in terms of the elementwise distance from the *true* or *population* precision matrix. Correlation and precision matrices are here independently inferred for each multivariate time series (each subject). We do not leverage the groupwide information across subjects in our algorithms, which is a relevant direction for further investigations and comparisons.

The article is structured as follows. In Sec. II, we define the problem of noise-cleaning correlation matrices and set the notation. Then we describe the benchmarked algorithms (Sec. II A), the quality criteria (Sec. II B), the cross-validation (CV) strategies (Sec. II C), and the characteristics of the synthetic (Sec. II D) and natural fMRI (Sec. II E) data sets. We finally present the results and draw the conclusions in Secs. III and IV, respectively.

II. MATERIALS AND METHODS

a. Data sets. Let data set X be an $N \times T$ real matrix consisting of T N-dimensional vectors, and let $\mathbf{x}(t) := (X_{1t}, \dots, X_{Nt}) \in \mathbb{R}^N$. In the context of network neuroscience, X represents the single subject data, N is the number of anatomic areas or regions of interest, and T is the length of the time signal. We will assume that the observations $\mathbf{x}(t)$ are identically and independently distributed. Furthermore, we assume the signal distribution to be a multivariate Gaussian. Without loss of generality, we will assume that the data is normalized in such a way that it exhibits null temporal averages and unit standard deviation: $\sum_{t=1}^{T} x_i(t) = 0$, $\sum_{t=1}^{T} x_i^2(t) = T$. In this article, we will focus on the case $T \geqslant N$, or $q \leqslant 1$ being $q := N/T$.

b. Sample and population covariance matrices. We will call $E = XX^{\dagger}/T$ the *sample* correlation matrix, where † indicates matrix transpose. Whenever T is finite or $q = N/T$ is nonnegligible, E and its inverse, in particular, are not good estimators of the (unknown) *population* or true (*verus*) correlation and precision matrices, C^{v} and $J^{\text{v}} = C^{\text{v}-1}$, that would have been obtained in the limit of infinitely many data, $T \to \infty$ *and* $q \to 0$. From the Marchenko-Pastur equation (see, for example, Refs. [40,41]) one knows, in particular, that the differences in the spectral densities of E and C^{v} are governed by q and vanish only for $q = 0$.

c. Noise-cleaned estimators of the covariance matrix. The problem of noise cleaning a covariance matrix amounts to proposing a noise-cleaned matrix C, given X, that aims to be as similar as possible to C^{v} (and more than what E is) according to some criterion. Equivalently, the cleaned matrix C aims to correct the overfitting (for small T) and the curse of dimensionality (for large q) that affect the unbiased estimator E. In other words, C should present lower bias+variance error [42] at the expense of a higher bias error.

In Bayesian terms, E is the maximum likelihood (ML) estimator of the covariance matrix given the data set X [assuming a Gaussian likelihood, $\mathcal{N}(X|E)$], while C is a *beyond-ML estimator* [43] in the sense that it aims at achieving a higher test likelihood at the expense of a lower train likelihood. Indeed, the design of noise-cleaning algorithms can be cast in terms of Bayesian random matrix theory [40]. Most of the algorithms that we consider here depend on a hyperparameter, generically γ, such that, for $\gamma = 0$ the resulting cleaned matrix is $C_{\gamma=0} = E$ (minimum bias, maximum variance) while, for its maximum value γ_+, it is $C_{\gamma_+} = 1_N$ (maximum bias, minimum variance). The optimal value γ^* of the hyperparameter can be set by cross-validated maximization of a validation-set likelihood (or by maximization of the training-set Bayesian evidence, as in PCA-Minka, see before). For infinitely many data-set vectors $T \gg N$, there is no overfitting and no curse of dimensionality, hence $\gamma^* = 0$ and $C = E$.

d. Training- and test sets, and hyperparameters. Each subject data set X is decomposed by columns into training and test data sets, $X = (X^{(\mathrm{tr})}, X^{(\mathrm{te})})$, of dimensions $N \times T_{\mathrm{tr}}$, and $N \times T_{\mathrm{te}}$, respectively. Given $X^{(\mathrm{tr})}$, we will obtain noise-cleaned covariance matrices C from several algorithms, also called *methods* in this article.

Some of the considered methods lead to a cleaned matrix C_π depending on a hyper-parameter π. For these methods, the training set is in turn decomposed into *inversion* (or pure training) and validation sets, $X^{(\mathrm{tr})} = (X^{(\mathrm{in})}, X^{(\mathrm{va})})$, of dimensions $T_{\mathrm{in}}, T_{\mathrm{va}}$, respectively. The optimal value of the hyperparameter is chosen by maximization of a *criterion* Q, $\pi^* = \arg\max_\pi Q(X^{(\mathrm{va})}|C_\pi)$ evaluated on the validation set (while C_π is computed from the inversion set), and the inversion-validation split is given by K-fold CV. Finally, the *quality* of each method according to the criterion Q (cross validated across random $(X^{(\mathrm{in})}, X^{(\mathrm{va})})$ partitions when needed) is given by $Q(X^{(\mathrm{te})}|C_{\pi^*})$. We evaluate the average and errors of this quantity, across all the *subjects* X belonging to a given *collection* of data sets, $(X^{(s)})_s$.

We describe the considered methods, criteria, and data sets in Secs. II A, II B, II D and II E, respectively.

A. Algorithms of noise-cleaning correlation matrices

It is out of the scope of this article to present a complete review of the immense amount of results on the general problem of overfitting and curse of dimensionality mitigation of covariance matrices. We limit ourselves to mention and compare a list of the better known and most popular algorithms for cleaning correlation matrices according to Ref. [39]. Let the spectral decomposition of the sample correlation matrix $E = X^{(\mathrm{tr})} X^{(\mathrm{tr})\dagger}/T_{\mathrm{tr}}$ be $E = U^\dagger \Lambda U$, where U is orthogonal and Λ is a diagonal, real matrix. The noise-cleaned or regularized matrix will be called C and its spectral decomposition $C = W^\dagger \hat{\Lambda} W$ where, again, W is orthogonal and $\hat{\Lambda}$ diagonal. We will assume their eigenvalues $\lambda_i = \Lambda_{ii} \geqslant 0$ and $\hat{\lambda}_i = \hat{\Lambda}_{ii}$ to be in decreasing order. As we will see, most of the standard algorithms modify only the sample spectra, so $W = U$.

a. Eigenvalue clipping, or PCA, according to which only p eigenvalues of E are considered to be significant, with $0 \leqslant p \leqslant N$.

(1) The cleaned spectrum is set equal to the sample spectrum $\hat{\lambda}_i = \lambda_i$ whenever $i \leqslant p$, otherwise it is set to a common noise value: $\hat{\lambda}_{i>p} = \bar{\lambda}_p = \sum_{j>p} \lambda_j/(N-p)$, equal to the average of the $N - p$ neglected eigenvalues.

(2) The resulting cleaned matrix is $C = U^\dagger \hat{\Lambda}\ U$. Given p, Λ, U, the choice of $\bar{\lambda}$ corresponds to a ML prescription (see, for example, Ref. [44]).

In PCA, the γ hyperparameter is the number of nonfitted principal components $\gamma = N - p$.

We have implemented two variants of this method: For the first one [PCA (CV)], the value of p is set by CV (see Sec. II C). In the second one [PCA (Minka)], p is chosen with the Minka criterion [44], consisting of a maximization of the training-set Bayesian evidence. This method does not require CV.

b. Linear shrinkage (shrinkage). The cleaned matrix here is a convex combination of the unbiased sample estimator E and a completely biased matrix, not depending on the data, that we will take as the identity matrix 1_N [45,46]. Depending on $\alpha \in [0, 1]$, the cleaned estimator is $C_\alpha = (1-\alpha)1_N + \alpha E$ or $\hat{\lambda}_i = (1-\alpha) + \alpha\lambda_i$. In this case, it is $\gamma = 1 - \alpha$. As explained in Ref. [39], the shrinkage method corresponds, in Bayesian random matrix theory, to the posterior average of the covariance matrix when the prior distribution is the inverse-Wishart distribution whose mean is the identity matrix in N dimensions, 1_N.

c. Optimal rotationally invariant estimator (RIE). The cleaned spectrum is [39,40,47,48]

$$\hat{\lambda}_i = \frac{\lambda_i}{|1 - q + q z_i s(z_i)|^2}, \tag{1}$$

where $s(z) := \mathrm{tr}[(z 1_N - E)^{-1}]/N$ is the Cauchy transform of the E spectral density, being $z_i := \lambda_i - \imath\eta$, $\imath$ the imaginary unit, and η a small parameter, coming from the limit $\eta \to 0$ in the derivation of the RIE estimator for large N, T (through the Sokhotski–Plemelj identity, see Ref. [40], Sec. 4). In Refs. [39,40] it is explained that, for finite N, a convenient choice is $\eta = N^{-1/2}$.

Roughly speaking, the RIE estimator is derived by imposing that matrix C is, among those sharing the eigenvectors with E, $C = U^\dagger \hat{\Lambda}\ U$, the one that exhibits a minimum Hilbert-Schmidt distance $d_{\mathrm{HS}}(C, C^{\mathrm{v}}) = \mathrm{tr}[(C - C^{\mathrm{v}})^2]$ from the population matrix $C^{\mathrm{v}} = V^\dagger \Lambda^{\mathrm{v}} V$. Albeit, the population matrix is not known; for the minimization of $d_{\mathrm{HS}}(C, C^{\mathrm{v}})$, it is sufficient to know, for large enough T, its spectral density $\rho_{C^{\mathrm{v}}}$, which is in turn related to the sample spectral density ρ_E through the Marcenko-Pastur equation (see Appendix A and Refs. [39,40,47,48] for details).

The RIE estimator is expected to provide a better estimation, in terms of d_{HS}, than any other algorithm modifying only the sample spectrum Λ, at least for sufficiently large values of T. In particular, we expect the RIE estimator to be more efficient than the PCA, shrinkage, caut-PCA, and q-corrected raw estimators.

We have implemented two variants of the RIE algorithm, one in which the parameter η is chosen according to the value prescribed in the literature $\eta = N^{-1/2}$ [39,40] (simply called RIE), the other one [called RIE (CV)]. in which η is cross validated on a grid of values. The RIE estimator does not require being cross validated (hence, for it, $X^{(\mathrm{tr})} = X^{(\mathrm{in})}$ and $T_{\mathrm{va}} = 0$). The cross-validated hyperparameter η of RIE (CV) does not balance bias and variance errors, hence it does not play the role of the parameter γ mentioned above.

d. Factor analysis (FA). This method proposes a cleaned matrix of the form (lower-rank matrix) + (heteroschedastic noise diagonal matrix). The generic form of the cleaned matrix is given by $C = M^\dagger M + Q$. Here M is a $r \times N$ real matrix, $r \in \{0, \ldots, N\}$, and Q is a real diagonal square matrix of size N. Given r, the values of M and Q are found numerically by maximization of the inversion likelihood $\mathcal{N}(X^{(\mathrm{in})}|C)$ (see, for example, Refs. [49,50]). The value of the hyperparameter r (the rank of the $M^\dagger M$ matrix) is chosen by CV.

e. Graphical lasso (lasso). Given the positive regularization (L_1-norm) hyperparameter α_{L}, the cleaned precision matrix $J = C^{-1}$ is given by maximization of the training likelihood minus a regularization term [51] (see also references in Ref. [52]):

$$J^* = \arg\max_J \left\{ \ln \mathcal{N}(X^{(\mathrm{tr})}|J^{-1}) - \alpha_{\mathrm{L}} \sum_{i<j} |J_{ij}| \right\}. \qquad (2)$$

In this case, it is $\gamma = \alpha_{\mathrm{L}}$. For $\alpha_{\mathrm{L}} = 0$, the estimated $C = E$, while for $\alpha_{\mathrm{L}} = \infty$, it is $C = 1_N$.

As a comparative reference for the efficiency of the above algorithms, we propose two further, simple methods.

f. Cautious-PCA (caut-PCA). We propose the following simple variant of PCA. In this case, $\gamma = N - p$ is again the number of neglected principal components but the spectrum of the cleaned matrix is modified differently.

(1) First, one provisionally sets $\hat{\lambda}_{i\leqslant p} = \lambda_i$ and $\hat{\lambda}_{i>p} = \bar{\lambda}_p$ as in PCA, but with $\bar{\lambda}_p = \lambda_p$ (the value of the lowest fitted sample eigenvalue) instead of $\bar{\lambda}_p = \sum_{i>p} \lambda_i/(N-p)$ as in normal PCA.

(2) Second, the whole cleaned spectrum is hence rescaled in such a way that C exhibits the same total variance as E: $\mathrm{tr}(\hat{\Lambda}) = \mathrm{tr}(\Lambda)$ or $\hat{\lambda}_j := \kappa_p \hat{\lambda}_j$, with $\kappa_p = N/(Nv_p + (N-p)\lambda_p)$ and $v_p = (\sum_{j\leqslant p} \lambda_j)/p$.

(3) Finally, the cleaned matrix is $C = U^\dagger \hat{\Lambda} U$.

Alternatively, the rescaling of the spectrum in step II A 0 f may be substituted by a standardisation of C: $C_{ij} := C_{ij}/\sqrt{C_{ii}C_{jj}}$. In our numerical analysis, both strategies lead to almost identical results.

Both PCA and caut-PCA methods fit the p largest eigenvalues and corresponding eigenvectors of E and neglect the lowest $N - p$ sample eigenvalues and corresponding eigenvectors. In the $N - p$-dimensional subspace of $\mathbb{R}^N$ generated by the neglected eigenvectors, the associated cleaned matrices are degenerated with a single noise variance $\bar{\lambda}_p$. The difference is that in PCA the noise variance $\bar{\lambda}_p$ is substituted by its ML value $\bar{\lambda}_p^{(\mathrm{ml})} = N(1 - v_p)/(N - p)$ [44], while in caut-PCA the noise variance $\bar{\lambda}_p^{(\mathrm{cau})} = \kappa_p \lambda_p$ consistently equals the lowest fitted eigenvalue λ_p that is considered to be significant, apart from the normalizing constant κ_p, which is close to 1 in the relevant regime $\lambda_p \ll N v_p/(N - p)$. In this regime, and whenever $\lambda_p \geqslant v_p \bar{\lambda}_p^{(\mathrm{ml})}$, the noise variance of caut-PCA *is larger than its ML value* $\bar{\lambda}_p^{(\mathrm{cau})} \geqslant \bar{\lambda}_p^{(\mathrm{ml})}$. In this sense, caut-PCA is more cautious. See more details in Appendix B.

g. Early-stopping gradient ascent (GA) algorithms. consist of an iterative updating of the correlation matrix C. It is initially set to 1_N, hence updated following a GA search of the training likelihood $\ln \mathcal{N}(X^{(\mathrm{in})}|C)$ or $C := C + \eta_{\mathrm{GA}} (\partial \ln \mathcal{N}(X^{(\mathrm{in})}|C')/\partial C')|_C$, where η_{GA} is a constant learning rate. While the likelihood is maximized by $C = E$, we stop the optimization earlier. The stopping criterion is given by the time of the first decrease of one of the validation-set criteria in Sec. II B).

We have implemented and compared mainly two variants of this algorithm, differing as follows. DGAW (deterministic gradient ascent-Wishart): the iteration is not on the matrix elements of C, but on those of an $N \times N$ matrix Y such that $J = C^{-1} = YY^\dagger$ (in such a Wishart form, the symmetry and positive definiteness of C is guaranteed in the iterative dynamics); SGA (stochastic gradient ascent) and its Wishart form SGAW: the iterative update algorithm for C or Y is not deterministic but stochastic: the gradient in each iteration τ is not that of $\ln \mathcal{N}(X^{(\mathrm{in})}|C)$ but, similarly to minibatch learning in machine learning [42], that of a random bootstrapping $X^{(\mathrm{in})}(\tau)$ of the training data, different from iteration to iteration. We have as well implemented further optional variants, as the coupled dynamics of a Lagrange multiplier guaranteeing the condition $\mathrm{tr}(C) = \mathrm{tr}(E)$. Please see further details of the GA algorithms in Appendix D.

B. Quality of the cleaned matrices according to different criteria

We evaluate the quality of the cleaned matrix C in the test set according to different criteria $Q(X^{(\mathrm{te})}|C)$:

a. Test likelihood (referred to as ℓ). The criterion $\ln \mathcal{N}(X^{(\mathrm{te})}|C)/T_{\mathrm{te}}$ is the average of the logarithm of the Gaussian likelihood over the test-set vectors $\mathbf{x}(t)$: $\ln \mathcal{N}(X^{(\mathrm{te})}|C)/T_{\mathrm{te}} = -(1/2)[\ln(2\pi) + \ln\det C + \mathrm{tr}(C^{-1}E_{\mathrm{te}})]$.

b. Test pseudolikelihood. We compute the average over the test-set vectors $\mathbf{x}(t)$ of the pseudolikelihood $\ln \mathcal{L}(\mathbf{x}) = \sum_{i=1}^N \ln p_i(x_i|\mathbf{x}_{/i}, C)$, where $\mathbf{x}_{/i}$ is the vector $\mathbf{x}$ with missing ith coordinate, and where $p_i(x_i|\mathbf{x}_{/i}, C)$ is the marginal of $\mathcal{N}(\mathbf{x}|C)$ given all the coordinates but x_i; it is a univariate normal distribution $p_i(x_i|\mathbf{x}_{/i}, C) = \mathcal{N}(x_i - \mu_i|\sigma_i^2)$ with C- and $\mathbf{x}_{/i}$-dependent average μ_i and variance σ_i^2:

$$\ln \mathcal{L}(X^{(\mathrm{te})}|C)$$

$$= \frac{1}{N T_{\mathrm{te}}} \sum_{t=1}^{T_{\mathrm{te}}} \sum_{i=1}^{N} \ln \mathcal{N}\big(x_i(t) - \mu_i(\mathbf{x}(t), C)|\sigma_i^2(C)\big), \qquad (3)$$

$$\mu_i(\mathbf{x}, C) := -\frac{\sum_{m\neq i} J_{im} x_m}{J_{ii}}, \qquad \sigma_i^2(C) := {J_{ii}}^{-1}, \qquad (4)$$

and where $J = C^{-1}$.

c. Test-completion error (referred to as $\bar{c}$). We define the *completion error* of the ith coordinate of vector $\mathbf{x}$, c_i, as the absolute value of the difference between x_i and its expected value according to the marginal distribution $p_i(x_i|\mathbf{x}_{/i}, C)$ or $c_i := |x_i - \mu_i(\mathbf{x}, C)|$. The *completion error of the data set* $X^{(\mathrm{te})}$, $\bar{c}(X^{(\mathrm{te})}|C)$ is defined as the average of the single coordinate completion error c_i over all coordinates i and all vectors $\mathbf{x}$ in $X^{(\mathrm{te})}$. It is a variant of the pseudolikelihood in which σ_i is not taken into account:

$$\bar{c}(X^{(\mathrm{te})}|C) = \frac{1}{N T_{\mathrm{te}}} \sum_{t=1}^{T_{\mathrm{te}}} \sum_{i=1}^{N} |x_i(t) - \mu_i(\mathbf{x}(t), C))|. \qquad (5)$$

The completion error is, hence, interpretable: it is the error, in units of the coordinates' variance ($= 1$), of the expected value of the missing coordinates x_i, Eq. (4), according to the Gaussian model induced by the inferred C.

d. Distance to the true precision and correlation matrices (referred to as d). . If the generative model of the data defined by the probability density P^{v} is known, it is possible to evaluate the quality of the cleaned matrix by computing the similarity between C and C^{v} (being ${C^{\mathrm{v}}}_{ij} = \langle x_i x_j \rangle_{P^{\mathrm{v}}}$) and between $J = C^{-1}$ and $J^{\mathrm{v}} = {C^{\mathrm{v}}}^{-1}$, according to a given criterion.

We will consider the matrix-element-wise metric,

$$d(J^{\mathrm{v}}, J) = \frac{\sum_{i\leqslant j} \left|J^{\mathrm{v}}_{ij} - J_{ij}\right|}{\sum_{i\leqslant j} \left|J^{\mathrm{v}}_{ij}\right|}, \tag{6}$$

and equivalently for $d(C^{\mathrm{v}}, C)$. This metric is essentially equivalent to the Hilbert-Schmidt distance $d_{\mathrm{HS}}(C^{\mathrm{v}}, C) = \mathrm{tr}[(C^{\mathrm{v}} - C)^2]$. Indeed, *given the generative model*, the relative efficiency of various algorithms as presented in Sec. III is qualitatively equal using d or d_{HS}. The only essential difference is that in Eq. (6) the mean error between matrix elements is expressed in units of the mean absolute value of the population matrix elements. The results are again essentially unchanged using a variant of $d(\cdot,\cdot)$ in which one discards the diagonal [i.e., $i < j$ in (6)]. Notice that the metric is interpretable: $d(J^{\mathrm{v}}, J)$ is the average distance between the matrix elements of the cleaned and true precision matrices, in units of the average value of the true precision matrix elements.

C. Hyperparameter tuning

Hyperparameter tuning for the algorithms in Sec. II A is performed by K-fold CV with $K = 6$, using the four quality criteria Q defined in Sec. II B. In other words, the optimal hyperparameter π^* is chosen by maximization of $\langle Q(X^{(\mathrm{va})}|C_\pi)\rangle$, where C_π is computed from $X^{(\mathrm{in})}$ and the average is over the K-fold partitions of $X^{(\mathrm{tr})} = (X^{(\mathrm{in})}, X^{(\mathrm{va})})$.

D. Synthetic data generation

For the generation of the synthetic data, we employ a generative multivariate Gaussian model $\mathcal{N}(\cdot|C^{\mathrm{v}})$ whose population covariance matrix is drawn from a probability distribution over correlation matrices that we dub the Dirichlet-Haar model. In the Dirichlet-Haar model, C^{v} is defined through the spectral decomposition $C^{\mathrm{v}} = W^\dagger \tilde{\Lambda} W$, where W is drawn from the Haar distribution (the uniform distribution over orthogonal matrices) and $\tilde{\Lambda}$ is a diagonal matrix whose eigenvalues $\tilde{\boldsymbol{\lambda}}$ are drawn from the Dirichlet distribution with parameter α_{D}: $\tilde{\boldsymbol{\lambda}}/N \sim \mathrm{Dir}(\cdot|\alpha_{\mathrm{D}})$ so $\mathrm{tr}(C^{\mathrm{v}}) = \mathrm{tr}(E) = N$. Therefore, α_{D} is the parameter that determines the degree of homogeneity or sparsity of the spectrum of C^{v}: large values of α_{D} lead to homogeneous eigenvalues $\tilde{\lambda}_i$, hence to correlation matrices with off-diagonal elements much lower than the diagonal (for $\alpha_{\mathrm{D}} \to \infty$, the Dirichlet-Haar model converges to the delta distribution around $C^{\mathrm{v}} = 1_N$). Vice versa, lower values of α_{D} lead to larger (and W-dependent) off-diagonal correlation (and precision) matrix elements, with a single large eigenvalue close to N. For this reason, α_{D} (or, more precisely, α_{D}^{-1}) may be seen as a measure of the degree of interaction strength of the population precision matrices J^{v}.

Summarizing, given N, $q = N/T$ and α_{D}, we generate N_{s} synthetic data sets, representing the subjects, according to the following procedure: We first sample an orthogonal matrix W from the Haar ensemble and a set of eigenvalues from the Dirichlet distribution $\mathbf{y} \sim \mathrm{Dir}(\cdot|\alpha_{\mathrm{D}})$, $\tilde{\boldsymbol{\lambda}} = N\mathbf{y}$; we construct $C^{\mathrm{v}} = W^\dagger \tilde{\Lambda} W$ and sample T vectors from the resulting Gaussian distribution $\mathbf{x}(t) \sim \mathcal{N}(\cdot|C^{\mathrm{v}})$. Such vectors $[\mathbf{x}(1), \dots, \mathbf{x}(T)]$ constitute the synthetic sample X. We repeat the procedure N_{s} times, getting a $N_{\mathrm{s}} \times N \times T$ synthetic collection of data sets $(X^{(s)})_{s=1}^{N_{\mathrm{s}}}$ in such a way that to every subject corresponds a different correlation matrix, with different (random) eigenvectors and different (random, but with common interaction strength) eigenvalues.

E. fMRI data

We analyze two fMRI data-set collections of BOLD activity time series of human subjects at rest, called A and B. Collection A is the one analyzed and described in Refs. [53,54]. Collection B is the large population-derived CamCAN Data Repository from the Cambridge Center for Aging and Neuroscience [55–57]. In collection A, we have measurements of $N_{\mathrm{s}} = 40$ subjects, with $T = 180$ observations of $N = 116$ features, that we randomly split in $T_{\mathrm{tr}} = 144 = T_{\mathrm{in}} + T_{\mathrm{va}}$ with $T_{\mathrm{in}} = 120$, $T_{\mathrm{va}} = 24$, and $T_{\mathrm{te}} = 36$ for those algorithms necessitating a validation set, otherwise $T_{\mathrm{tr}} = T_{\mathrm{in}} = 144$. For collection B, we have $N_{\mathrm{s}} = 652$, $N = 114$, and $T = 260$ observations split in $T_{\mathrm{in}} = 174$, $T_{\mathrm{va}} = 34$, and $T_{\mathrm{te}} = 52$.

III. RESULTS

A. Noise cleaning the synthetic data set

We applied the methods from Sec. II A to the synthetic data sets described in Sec. II D with varying dimensions per sample $q = N/T$ and degree of interaction α_{D}^{-1}. We assess the quality of the cleaned correlation and precision matrices according to the criteria described in Sec. II B, evaluated on the test set of each subject.

We consider a grid of values of q, α_{D} constructed as follows: We take $N = 116$ fixed, that coincides with the dimension of the standard fMRI parcelization of the human brain in 116 regions of interest, and coinciding with value of N for fMRI collection A, see Sec. II E. The value of T_{tr} takes the values 144, 200, 300, 1000, 2000. The lower value $T_{\mathrm{tr}} = 144$ coincides with T_{tr} of fMRI collection A. α_{D} takes the values $\alpha_{\mathrm{D}} = 0.5, 1, 1.5, 2, 2.5, 3, 4$. A fraction 1/6 of the T_{tr} observations is used for validation, as we said in Sec. II C. The test sets are composed by $T_{\mathrm{te}} = 0.25\, T_{\mathrm{tr}}$ vectors. For each value of the couple q, α_{D} the results are averaged over $N_{\mathrm{s}} = 10^2$ realizations of the data set (different "subjects").

In this section and in Appendix C, we present results for the methods: PCA (CV-l), PCA (Minka), shrinkage (CV-l), Lasso (CV-l), FA (CV-l), RIE, RIE (CV-l), RIE (CV-e), DGAW (CV-l), SGAW (CV-l), SGAW (CV-e), caut-PCA (CV-l), and caut-PCA (CV-e). The first part of the methods' name is as explained in Sec. II, while the notations CV-l and CV-e refer to the cross-validation strategy by maximization of the test-likelihood or test-completion error, respectively (see Sec. II C). For the sake of clearness, in the figures of this section we omit the combinations of methods and CV strategies that are less distinguishable or efficient. Although the CV strategy may induce some statistically significant differences for some values of the parameters q, α_{D} (see Fig. 1, $\alpha_{\mathrm{D}} = 1$, $T_{\mathrm{tr}} = 144$), such differences are negligible in most cases. Importantly, in the analysis below we add the baselines oracle and raw, in which no cleaning procedure is applied but, instead, the quality estimators $Q(X^{(\mathrm{te})}|C^{\mathrm{v}})$ and $Q(X^{(\mathrm{te})}|E)$ are directly evaluated using as cleaned matrices the true and

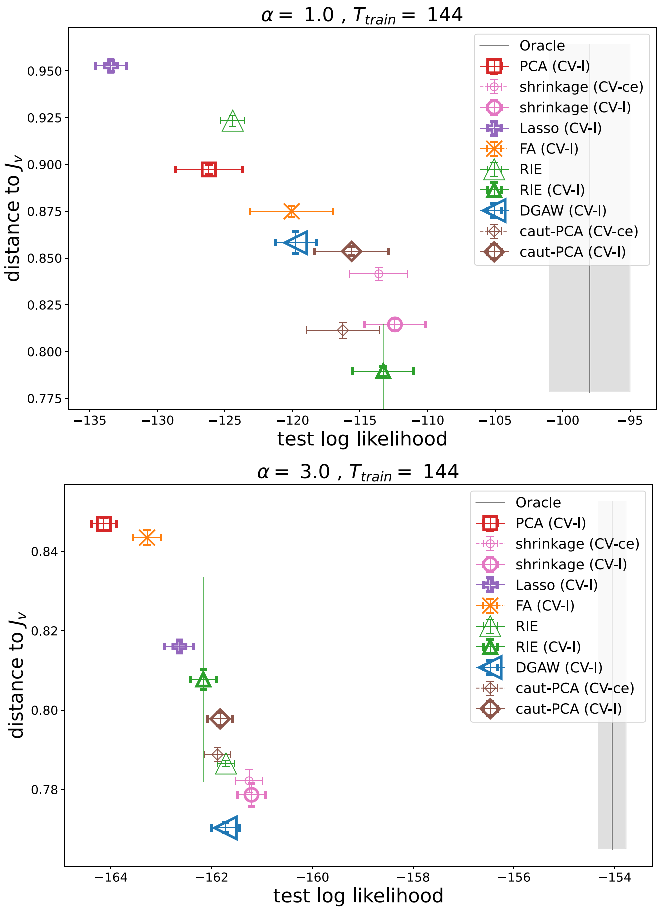

FIG. 1. Scatter plot of $d(J^{\text{v}}, J)$ (lower is better) versus ℓ (higher is better) for synthetic data in the severe undersampling regime $T_{\text{tr}} = 144$. Points and error bars are averages and standard errors of the mean across subjects, and each point corresponds to a different method. The thin vertical error bars with no cap over RIE are the standard deviation across subjects. The vertical line indicates the likelihood of the oracle method (whose $d(J^{\text{v}}, J)$ vanishes). Higher panel: $\alpha_{\text{D}} = 1$ (strongly correlated matrices, highly discontinuous spectrum). Lower panel: $\alpha_{\text{D}} = 3$ (weakly correlated matrices).

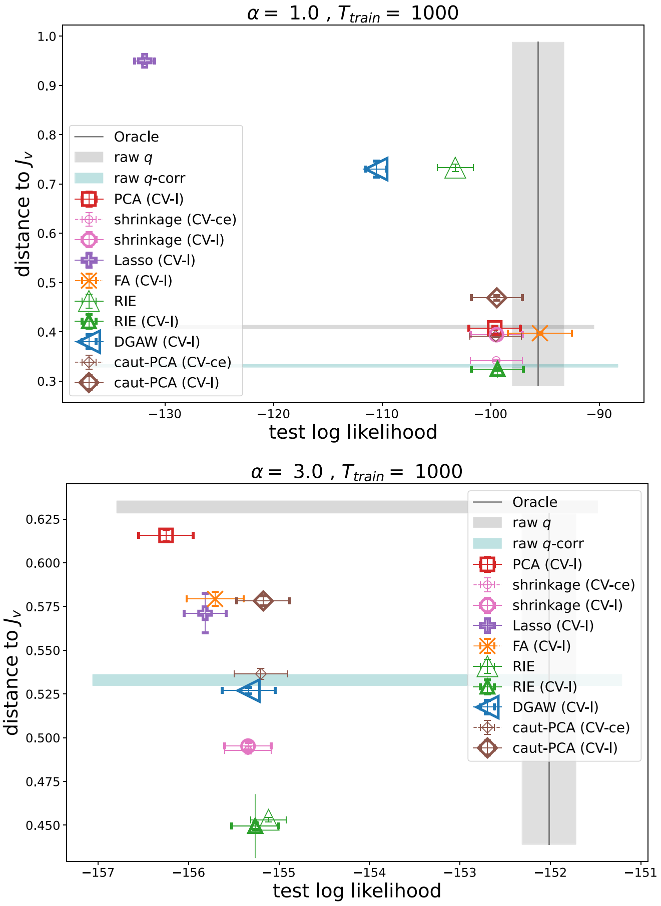

FIG. 2. As in Fig. 1 but for $T_{\text{tr}} = 1000$ (moderately undersampled regime). The horizontal bars indicate the mean±standard deviation of $d(J^{\text{v}}, J)$ corresponding to the raw (upper), and raw q-corr (lower) methods.

the sample matrices, respectively. An efficient estimator is expected to exhibit a value of Q larger than that of the raw estimator, and as close as possible to that of the oracle estimator.

We mainly focus on the analysis of the distance from the true precision matrix $d(J^{\text{v}}, J)$. This is the quality criterion presenting by far higher variability across methods in terms of its subject-to-subject errors (see Fig. 1). Information regarding the performance in terms of other quality criteria [test-likelihood, $d(C^{\text{v}}, C)$, test-completion error, test-pseudolikelihood] and for variants of these cleaning methods (cross validated with respect to the completion error or the pseudolikelihood) may be found inAppendix C and in Ref. [58], where a complete table of the performance of all algorithms according to all the criteria is made available [59].

In Figs. 1 and 2, we show the subject-averaged values of $d(J^{\text{v}}, J)$ ($J = C^{-1}$ being the best precision matrix for a given method) versus the test likelihood. Each point corresponds to a different algorithm, and each figure to a different combination of the data-set parameters q, α_{D}. The error bars represent the standard error of the mean (SEM) across subjects. For reference, we also include the standard deviation across subjects ($N_{\text{s}}{}^{1/2} = 10$ times larger than the SEM) *only for the RIE (CV-l) method* (yaxis thin error bars without cap). Please notice that the SEM, indicated by capped error bars are, instead, of the same order than the symbol size. The grey vertical strip (oracle method) indicates the value of the test-likelihood error corresponding to the true correlation matrix C^{v}. Mind that the oracle estimator exhibits null $d(J^{\text{v}}, J = J^{\text{v}}) = 0$.

Figure 1 corresponds to $T_{\text{tr}} = 144$ ($q \simeq 0.85$, severe undersampling), while Fig. 2 to $T_{\text{tr}} = 1000$ ($q = 0.116$, moderate undersampling). In the severe undersampling regime, all the considered methods lead to a cleaned precision matrix whose error is lower than one, while the distance $d(J^{\text{v}}, J = E^{-1})$ corresponding to the raw estimator lies far outside the figure at $(\ell, \bar{c}, d) \simeq (-327, 0.79, 11.7)$ for $\alpha_{\text{D}} = 1$ and $(\ell, \bar{c}, d) \simeq (-381, 1.5, 17.6)$ for $\alpha_{\text{D}} = 3$. In such a severe undersampling situation, the raw unbiased estimator of the precision matrix exhibits matrix elementwise errors which are more than ten times larger than the average of the matrix elements of J^{v}.

This is precisely what we expect from random matrix theory: whatever distribution the C^{v} has been sampled from, the empirical matrix E, follows (assuming a Gaussian data likelihood) the Wishart distribution $P_{\rm W}(E|C^{\rm v})$ whose average is $C^{\rm v}$ (see, for example, Ref. [40]), while the sample precision matrix E^{-1} follows the inverse-Wishart distribution $P_{\rm iW}(E^{-1}|C^{{\rm v}-1})$ whose average is $(1-q)^{-1}C^{{\rm v}-1}$. Hence, we expect that the precision matrix in a situation of $q \simeq 0.9$ is about ten times larger than the true precision matrix. This argument suggests we compare our results with an additional, reference cleaning method simply consisting of multiplying the raw precision by $1-q$:

$$E \to J = (1-q)E^{-1}. \tag{7}$$

We refer to this method to infer J as raw (q-corr.) in the figure legend. The raw q-corrected method systematically reduces the distance to the true precision matrix with respect to E^{-1} ($d \simeq 2.0$ for $\alpha_{\rm D}=1$ and $d \simeq 3.3$ for $\alpha_{\rm D}=3$) for the lowest $T_{\rm tr}$ but, in this case, it still leads to a much larger distance d than the rest of the considered cleaning methods. The situation is the opposite for the largest $T_{\rm tr}=1000$, see Fig. 2: For $\alpha_{\rm D}=1$, there is no cleaning method leading to a significantly lower d than the raw q-corrected method: cleaning is likely to be counterproductive [only RIE (CV) leads to a nonlarger d than raw (q-corr.)]. The intermediate situation for $T_{\rm tr}=300$ is shown in Fig. 11.

We now draw some conclusions on the synthetic data analysis from the results in Figs. 1, 2 and Appendix C.

(1) The across-method differences in terms of distance from the true precision matrix are more significant than in terms of test likelihood and completion error: significant differences between two methods in ℓ and $\bar{c}$ also imply significant differences in d, while the opposite does not hold (see Figs. 1 and 2). In any case, the method ranking resulting from d, $\bar{c}$, and ℓ are consistent, these quantities being strongly correlated across methods (Figs. 10 and 11).

The across-method differences in d are, in some cases, significant not only in terms of SEM but even in terms of standard deviation across subjects (Fig. 1). In these cases, the difference between the best and worst algorithms' average d amounts to two or more standard deviations of d across subjects and, consequently, the best methods present lower distance than the worst methods *for most of the subjects*.

(2) The ranking of methods providing a lower distance to the population precision matrix depends much on the dataset characteristics $q, \alpha_{\rm D}$. While for high values of q all the considered noise-cleaning algorithms reduce the distance to the population precision matrix, beyond the raw and raw (q-corr.) algorithms, for low values of q the noise-cleaning may be counter-productive (Fig. 6).

(3) For sufficiently large values of $T_{\rm tr}$, the optimal RIE is the best performing in terms of all the criteria, *when complemented with the CV for the parameter* η suggested in this article [algorithm RIE (CV), see Figs. 2 and 6].

(4) Considering the whole grid of data-set parameters $q, \alpha_{\rm D}$, the best performing algorithms (according to d) are RIE (CV), shrinkage (CV), DGAW (CV-l) (see Figs. 1 and 6). The first two, however, have the advantage of being principled, faster (not requiring an optimisation at the level of $X^{(\rm in)}$), and robust (performing well in all regimes of the synthetic benchmark and on the fMRI data as well—see below), while GAW works well for low degrees of interaction only, and it is less robust. The high-$\alpha_{\rm D}$, high q regime in which the GAW algorithm performs well is, incidentally, the most difficult regime, presenting the highest values of d, $\bar{c}$, and the lowest values of ℓ (Figs. 6, 10, and 11). The results of this section (limited to the specific Haar-Dirichlet generative model that we use to generate the synthetic data) suggest using the RIE (CV) algorithm, since it is the one providing a distance d lower or statistically compatible with the (q-corrected) raw estimator even for large values of $T_{\rm tr}$.

(5) The proposed algorithm caut-PCA significantly improves PCA for almost all considered values of $q, \alpha_{\rm D}$.

We note that, for some of the probed values of $q, \alpha_{\rm D}$, RIE (CV) performs worse than RIE (Fig. 6) despite the fact that the parameter η is cross validated from a list that actually contains the value $\eta = N^{-1/2}$ used by the plain RIE method. This is possible since RIE does not have any hyperparameter and consequently does not need a validation set: the spectrum λ in Eq. (1) is computed from the $T_{\rm tr}$ vectors in $X^{(\rm tr)}$, while in RIE (CV) it is computed from the $T_{\rm va}=(5/6)T_{\rm tr}$ vectors in $X^{(\rm va)}$. The same happens with PCA (CV) and PCA (Minka) (Fig. 6).

B. Noise cleaning the fMRI brain activity datasets

We have applied the noise-cleaning algorithms to the two fMRI collections A and B described in Sec. II E. In this case, we do not have a population precision matrix $J^{\rm v}$ to compute the distance d from the inferred J. We assess the quality in terms of the criteria $\ell, \bar{c}$.

We present, in Fig. 3, the average of the criteria $\ell, \bar{c}$ across the subjects of collections A and B for various noise-cleaning methods. The short, capped error bars indicate the SEM across data-set subjects, while the thin error bars without cap over the PCA (CV-l) method indicate the standard deviation. The lasso (CV-l,e) algorithms are absent in the figure since they did not achieve convergence (see the details in Appendix E). The GA algorithms not in Wishart form may present problems with the positive-definiteness of the optimizing matrix C, depending on the choice of the learning rate—this is why they are absent in collection B. Conversely, in collection A, the DGAW and SGAW algorithms have been excluded since they present lower performance (generally speaking, the performance of the GA-based algorithm is rather sensible to the choice of the learning rate and bootstrapping fraction, see Appendix D).

We draw the following conclusions for this section:

(1) The differences across algorithms are definitely statistically significant in both data sets, in some cases even at the subject level. Assuming that the results on synthetic data sets hold in this context, we expect that the across-algorithm differences in the (inaccessible) d are even more significant.

(2) The results are qualitatively consistent in the two data collections. The best performing algorithms in terms of both ℓ and $\bar{c}$ are RIE (CV) and shrinkage (CV), followed by caut-PCA, consistently with the synthetic data-set results.

(3) On the natural data sets, only RIE (CV), not RIE, performs well: the CV of the parameter η becomes particularly useful. We make note that the RIE algorithm Eq. (1) is a function of q and that, in natural data, the choice $q = N/T_{\rm tr}$ neglects the temporal correlation between data-set vectors. A more convenient choice would be $q_{\rm eff} = N\tau/T$, where τ is

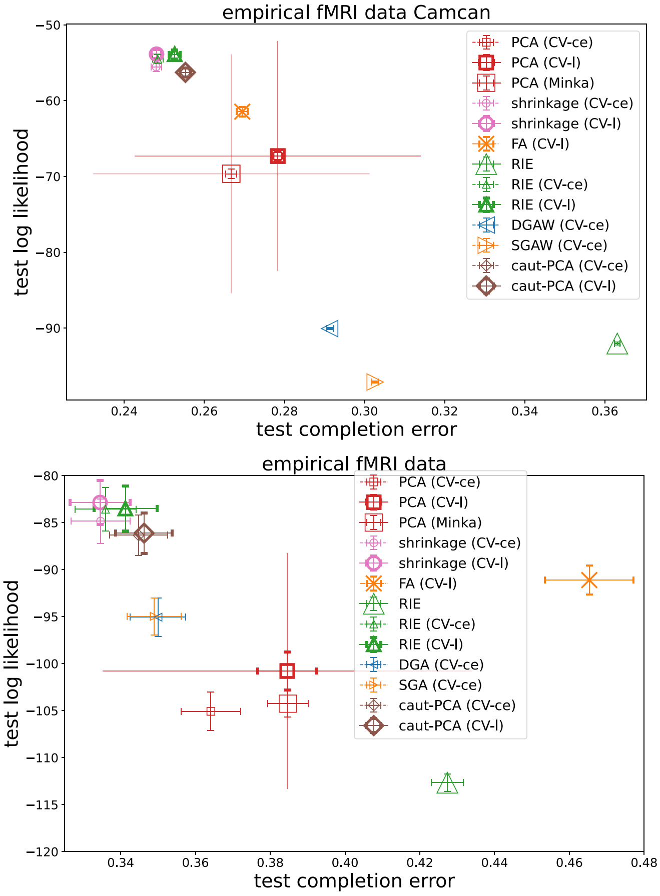

FIG. 3. Scatter plot of test-set likelihood ℓ (higher is better) versus completion error $\bar{c}$ (lower is better) in empirical fMRI data. Points and error bars are averages and standard errors of the mean across subjects, and each point corresponds to a different method. The thin vertical error bars with no cap over PCA are the standard deviation across subjects. Higher and lower panels: fMRI databases A and B, respectively.

the correlation time of the series (so the effective number of uncorrelated vectors is taken to be $T_{\rm eff} = T/\tau$). It is possible that the CV of η in RIE (CV) compensates for the misleading choice $q = N/T_{\rm tr}$. This is a possible origin of the inadequacy of plain RIE for natural data.

IV. CONCLUSIONS AND DISCUSSION

The precision matrix between different brain regions, inferred from fMRI of MEG, is a fundamental quantity in the context of network neuroscience. It is widely studied as a model of SC between brain areas in the harmonic approximation and to capture significant intersubject and intergroup differences, beyond those encoded in the correlation matrix [11,24,25,27,27–32].

This motivates the interest in an assessment of the absolute and relative utility of various noise-cleaning strategies for an accurate inference of the precision matrix in the context of network neuroscience. In particular, we are interested in assessing, and comparing with known methods, the efficiency of an overfitting mitigation strategy based on random matrix theory, the Ledoit-Péché, or optimal RIE [39,47] (see as well Ref. [60]), whose efficiency and potential utility has not yet been, to the best of our knowledge, addressed in the context of neuroscience.

In this article, we have performed a numerical analysis of the relative efficiency of several well-known strategies of regularization of the covariance (and hence precision) matrix of data sets in the $T \gtrsim N$ regime, being N, T of the order of typical fMRI and MEG neural data. For such a comparison, we have used both synthetic data sets of Gaussian vectors, of varying inverse sample ratio $q = N/T$ and degree of off-diagonal correlation, and two data sets of human brain activity at rest, measured by fMRI. We have performed such a comparison in terms of both the distance d between the noised-cleaned (inferred) and population (true) precision matrices, in the case of the synthetic data sets, and of the out-of-sample likelihood ℓ.

We have observed that:

(1) At least in our synthetic data set, the distance d, or *the average error in the inferred precision matrix elements, may significantly depend on the chosen cleaning strategy* (that may induce typical differences of the 20% in d or larger). Such interalgorithm differences in d are larger than those in ℓ. This suggests that, in a context in which the precision matrix should be inferred accurately (e.g., for classification purposes), the choice of the noise-cleaning method may be crucial.

(2) The analysis of both fMRI data sets consistently suggests that *the algorithms shrinkage (CV) and RIE (CV) are those providing* a higher ℓ and, consequently, *a more faithful precision matrix. Notably, the RIE method is accurate only in its RIE (CV) variant, proposed here, in which one of the plain RIE parameters is cross validated*, hence compensating for the presence of temporal correlations in the data (Fig. 3). The method RIE (CV) has the further advantage of being, by construction, optimal with respect to other rotationally invariant methods (as shrinkage) for large enough values of T, as confirmed by the synthetic data-set analysis. Indeed,

(3) *In synthetic data, RIE (CV) exhibits*, as expected, *significantly lower d for large T's*, being the only method that improves the raw estimator E^{-1} [with the Marcenko-Pastur q-correction Eq. (7)] in all the simulated regimes (see Fig. 6). Again, and specially for strongly correlated synthetic data, only our cross-validated variant RIE (CV) method performs well (Figs. 1 and 2).

(4) *The simple GA algorithms*, consisting in a (train-data-set likelihood) GA iterative updating of the covariance matrix, combined with an early stopping criterion to prevent overfitting, *outperforms the most efficient algorithms in low-T, weakly correlated synthetic data* (Figs. 1 and 6). It is not our aim to present a systematic nor rigorous study of the efficiency of such algorithms that could be optimized in several ways (bootstrapping strategy and fraction, learning rate, initial condition, stopping criterion). We rather show numerically that, as a proof of principle, such a simple early stopping GA technique is enough to accurately infer weak correlations of strongly undersampled data, at least in the synthetic data set at hand.

(5) *The cautious PCA algorithm*, simply consisting of raising the value $\bar{\lambda}$ of the noise eigenvalues in the PCA method,

systematically improves the inferred precision matrix with respect to PCA in the natural and synthetic datasets (see Figs. 3, 6, and Appendix B).

Summarizing, the present analysis results suggest that, whenever accurate statistical estimators of the precision matrices are needed in brain connectivity studies, the optimal RIE, *if completed with the simple CV strategy for the parameter η proposed in this article*, is the best one in terms of robustness, accuracy, and computational cost.

In this article, we have cast the inference of brain structure from single-subject temporal fMRI of MEG series as a problem of covariance matrix noise cleaning, hence deliberately restricting the analysis to (1) linear inference: the data nonlinearities are neglected; (2) noncausal inference: we neglect the data temporal correlations; (3) inference from single-subject data only: we do not exploit group information. In this precise context, we have performed a systematic comparison between well-known noise-cleaning algorithms, together with a further method (RIE) based on random matrix theory.

All such algorithms stand on a Gaussian likelihood $\mathcal{N}(X|C)$, and some of them on the Marchenko-Pastur relation or on the inverse Wishart distribution for C: i.e., on a statistical theory of finite-T correction of the sample spectrum, again under Gaussian hypotheses. The quality of the noise cleaning will consequently depend on the extent to which the data meet the above mentioned assumptions [1, 2].

The working hypotheses [1, 2] are, in principle, not satisfied in fMRI data. Nevertheless, in the presence of moderate temporal correlations and relatively small nonlinear cumulants, as those exhibited by BOLD resting-state fMRI data [61], the use of (linear) noise-cleaning algorithms may still be advantaged in front of causal or nonlinear inferring models. First, it is not obvious that, in severe undersampled situations, cumulants of higher order can be significantly inferred. Second, in the linear setup we can count, as said before, on a statistical theory of finite-T corrections of the population matrix spectrum.

Starting from linear inference, and sequentially accounting for nonlinearities, temporal correlations [12–21], and group information [11,25,29,30,32,35–37] would allow addressing the relative importance of these elements in the inference of functional data. Particularly interesting could be the comparison with one of the standard tools for inferring brain structure, dynamic causal modelling [62–66], accounting for both nonlinearities and temporal correlations. We suggest as well a comparison with the recent promising algorithm [60], rooted as well in random matrix theory.

We publicly release the algorithm's implementations and the code for reproducing the experiments [58].

ACKNOWLEDGMENTS

We acknowledge the advice of M. X. Cohen, F. Lillo, and A. Pagnani. Thanks to Z. Drogosz, M. Nowak, and P. Oświecimka for further discussions and for making us aware of Ref. [60]. This work is supported by the European Union's Horizon 2020 Program under the scheme INFRAIA-01-2018-2019—Integrating Activities for Advanced Communities, Grant Agreement No. 871042, SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics [67]. This article is also supported by the European Union's NextGenerationEU—National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR), project SoBigData.it—Strengthening the Italian RI for Social Mining and Big Data Analytics—Grant No. IR0000013 (No. 3264, 28/12/2021). M.I.-B. has been supported by the grant EU FESR-FSE PON *Ricerca e Innovazione* 2014-2020 BraVi, awarded to Stefano Panzeri; the Spanish Ministry and Agencia Estatal de Investigación (AEI) through Project I+D+i Ref. No. PID2020-113681GB-I00, financed by MICIN/AEI/10.13039/501100011033 and FEDER, "A way to make Europe".

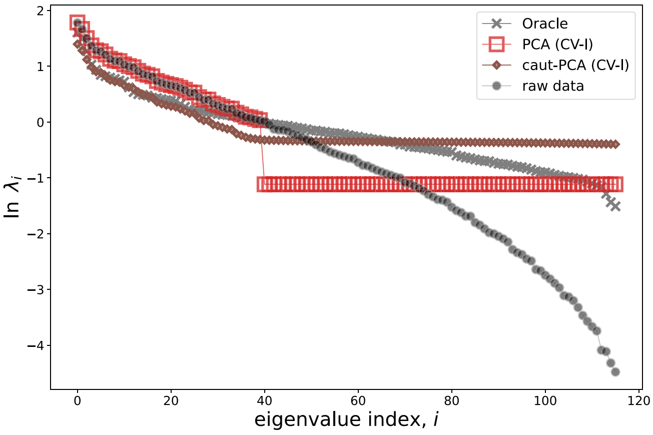

FIG. 4. Logarithm of the spectra of true (oracle), sample (raw data) and cleaned matrices (according to PCA and caut-PCA with fixed $p = 40$) versus the eigenvalue index. See the details of the database in the main text.

APPENDIX A: OPTIMAL RIE DERIVATION SKETCH

We present an informal sketch of the derivation of the Optimal RIE algorithm; please see Refs. [39,40,47,48] for details. Minimizing $d_{\mathrm{HS}}(C, C^{\mathrm{v}}) = \mathrm{tr}[(C - C^{\mathrm{v}})^2]$ with the constraint $C = U^\dagger \hat{\Lambda}\, U$ leads to $\hat{\lambda}_i = \sum_k (\hat{\mathbf{v}}_k^\dagger \hat{\mathbf{u}}_i)^2 \lambda_k^{(\mathrm{v})}$. The optimal

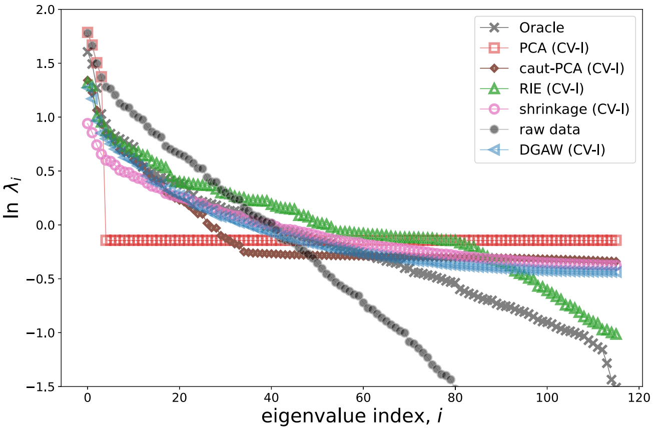

FIG. 5. As in Fig. 4, but cross validating the value of p and comparing (in the same database X) PCA (CV) and caut-PCA (CV) with other methods.

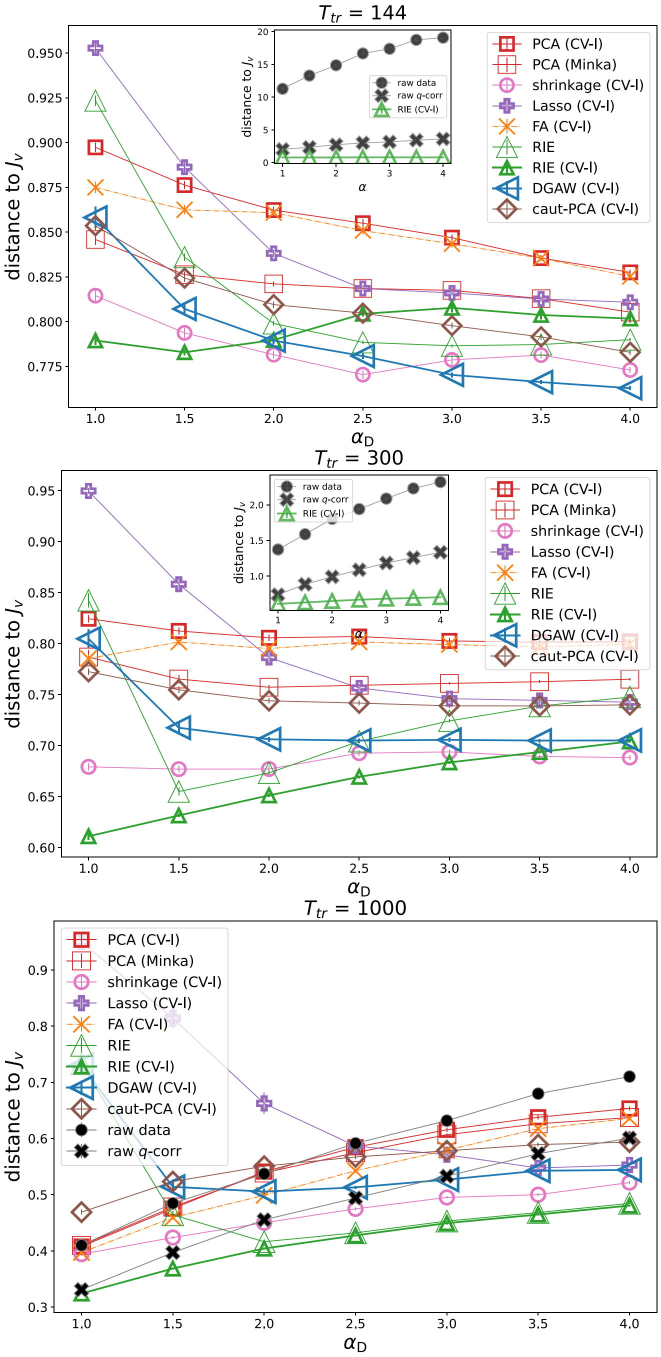

FIG. 6. $d(J^{v}, J)$ (lower is better) versus α_D in synthetic data. Points and error bars are averages and standard errors of the mean across subjects, and each curve corresponds to a different method. Higher, middle and lower panel correspond to $T_{tr} = 144, 300, 1000$, respectively. In the two highest panels, the inset compare the raw, raw (q-corr.) and RIE estimators, while in the lower panel all estimators are in the main figure.

$\hat{\Lambda}$ is, in other words, a function of the true spectrum and of the overlap between the true and empirical eigenvectors. Fortunately, and roughly speaking, there exist RMT relations allowing us to express the (average) overlap between true and sample eigenvectors in terms of the true spectrum and the (Cauchy transform of the) sample spectrum (see Ref. [40]). In this way, one can write $\hat{\Lambda}$ in terms of the true spectrum Λ^{v}

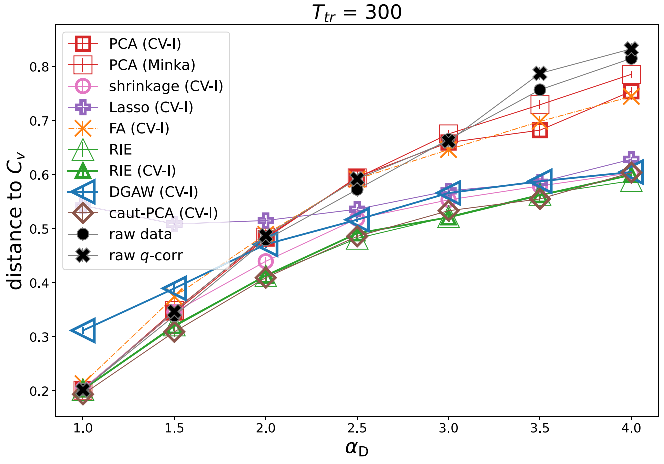

FIG. 7. $d(C^{v}, C)$ (lower is better) versus α_D in synthetic data, for $T_{tr} = 300$. Points and error bars are averages and standard errors of the mean across subjects, and each curve corresponds to a different method.

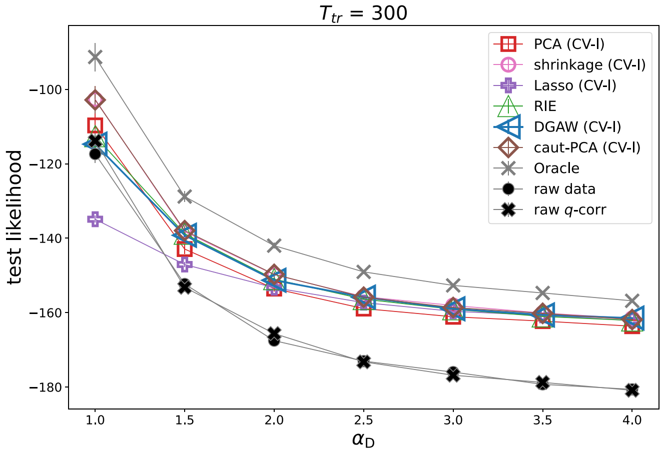

FIG. 8. Test-likelihood ℓ (higher is better) versus α_D in synthetic data, for $T_{tr} = 300$. Points and error bars are averages and standard errors of the mean across subjects, and each curve corresponds to a different method.

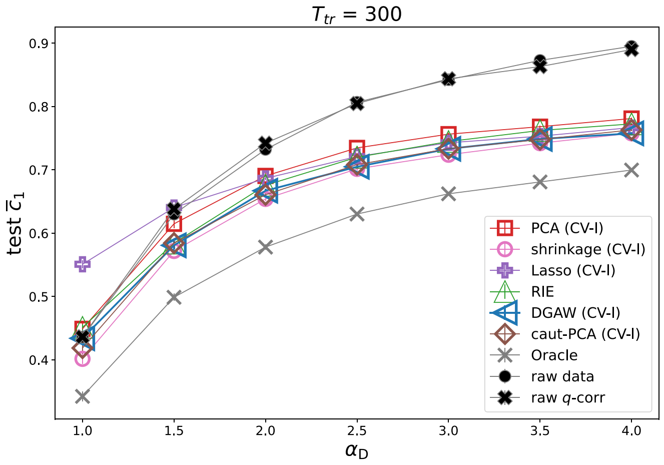

FIG. 9. Completion error $\bar{c}$ (lower is better) versus α_D in synthetic data, for $T_{tr} = 300$. Points and error bars are averages and standard errors of the mean across subjects, and each curve corresponds to a different method.

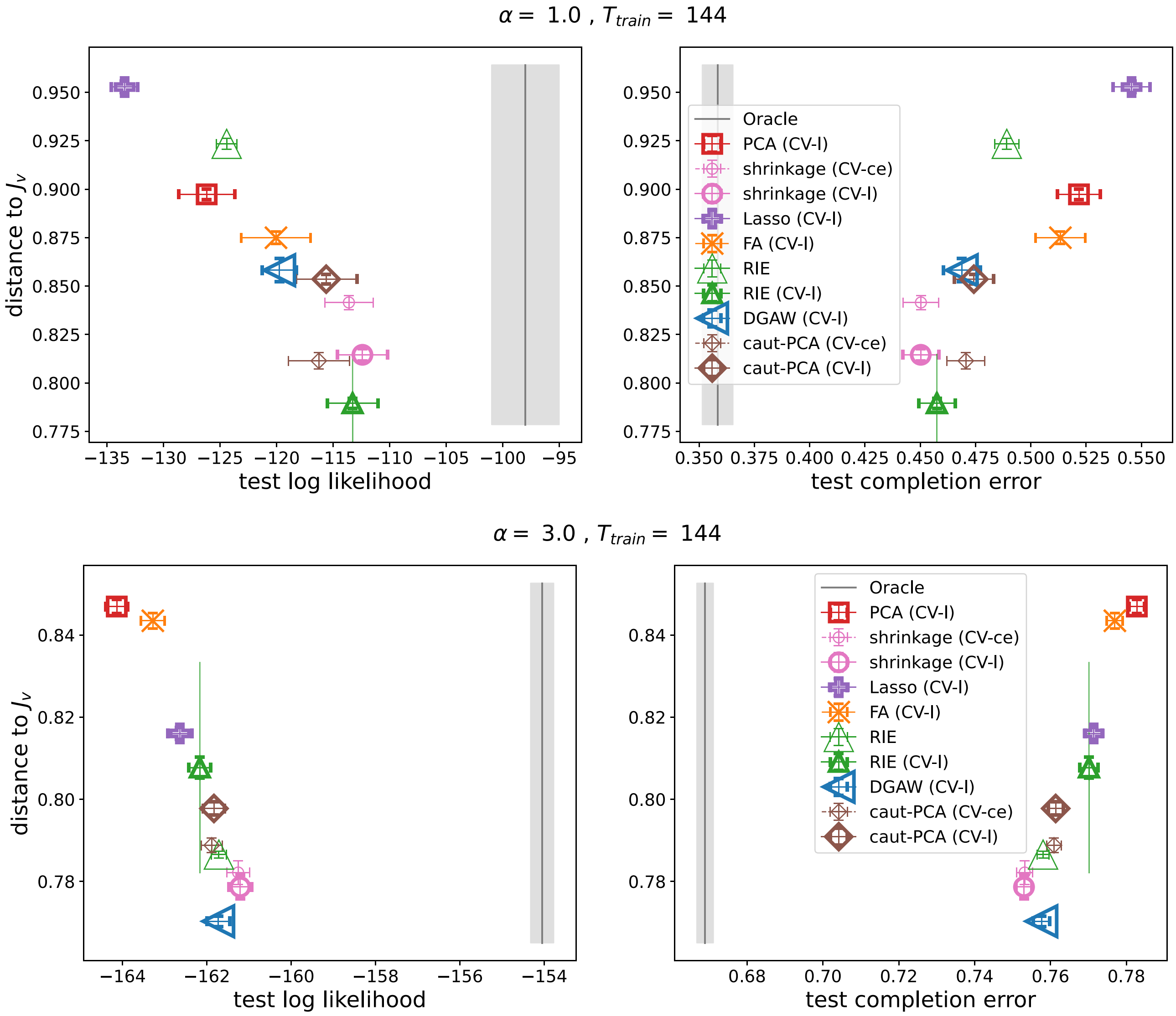

FIG. 10. Top left: scatter plot of $d(J^{\text{v}}, J)$ (lower is better) versus test-likelihood ℓ (higher is better), for $T_{\text{tr}} = 144$, $\alpha_{\text{D}} = 1$. Top right: idem but $d(J^{\text{v}}, J)$ versus $\bar{c}$ (lower is better). Bottom, left and right: idem, but for $\alpha_{\text{D}} = 3$.

only. Finally, the true spectrum may be related to the empirical spectrum through the Marchenko-Pastur equation, which relates the spectral density of C^{v} to (the Cauchy transform of) that of E:

$$g_E(z) = \int d\lambda \frac{\rho_{C^{\text{v}}}(\lambda)}{z - (1 - q + qzg_E(z))\lambda},$$

$$g_{\text{M}}(z) := \int d\lambda \frac{\rho_{\text{M}}(\lambda)}{z - \lambda} \quad \text{Cauchy transform.}$$

In this way, it is possible to write an expression for $\hat{\Lambda}$ in terms of Λ only:

$$\hat{\lambda}_i \simeq \frac{\lambda_i}{|1 - q + q\lambda_i \lim_{\eta \searrow 0} g_E(\lambda_i - \imath\eta)|^2},$$

of which Eq. (1) is a further simplification.

We make note that in our repository [58], we actually implement the further *debiased* RIE heuristic correction for low values of N. Please see the details in Ref. [39].

APPENDIX B: CLEANED SPECTRA

In Fig. 4, we illustrate the effect of the algorithm caut-PCA. We generate a single synthetic database $X \sim \mathcal{N}(\cdot|C^{\text{v}})$, where C^{v} is sampled from the Haar-Dirichlet model with $T_{\text{tr}} = 144$, $N = 116$, $\alpha_{\text{D}} = 3$. Afterward, we plot the spectra of the cleaned matrix C according to PCA and caut-PCA for a fixed value of $p = 40$. As explained in Sec. II A, the noise eigenvalue of caut-PCA ($\bar{\lambda}_p^{(\text{cau})}$) is larger than that of PCA ($\bar{\lambda}_p^{(\text{ml})}$) for a fixed p, whenever $\bar{\lambda}v_p \leqslant \lambda_p$.

The reader may notice that the $\lambda_{j>p}$ eigenvalues of caut-PCA in Fig. 4 are not constant. This is because we are using the standardization in step 2 of the description of cautious-PCA in Sec. II A. Using the rescaling instead, one obtains a similar spectrum with a constant noise eigenvalue.

As a consequence of $\bar{\lambda}_p^{(\text{cau})} > \bar{\lambda}_p^{(\text{ml})}$, the cross-validated value of p^* tends to be lower in PCA (CV) than in caut-PCA (CV), as illustrated in Fig. 5. The reason is that the low value of $\bar{\lambda}_p^{(\text{ml})}$ penalises large values of p. In the presence of overfitting, for low T, the validation-set energy term in

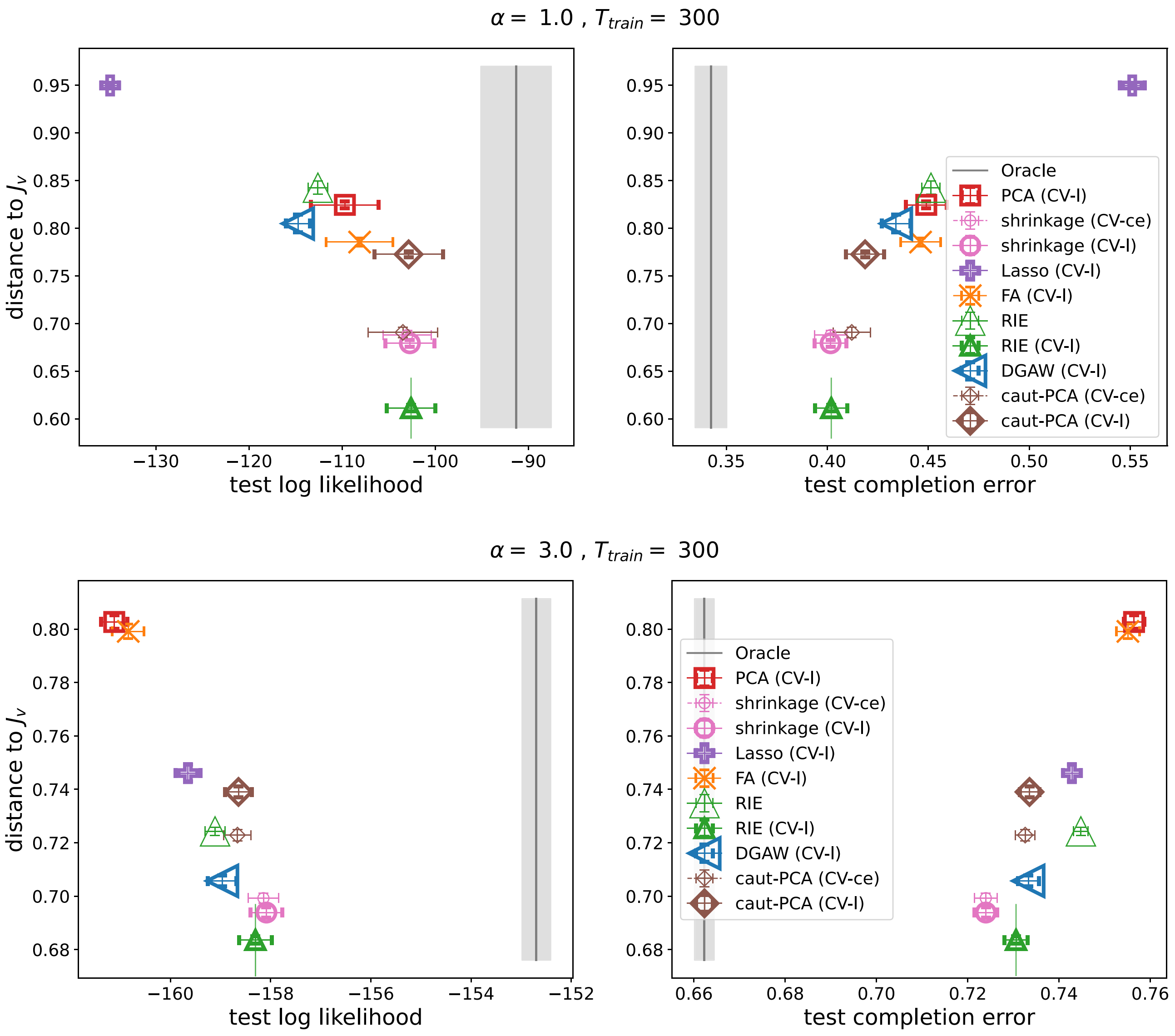

FIG. 11. As in figure 10 but for $T_{\mathrm{tr}} = 300$.

the likelihood, $-(1/2)\sum_t \sum_{j>p}(x'_j(t))^2/\bar{\lambda}_p$ (being $\mathbf{x}' = U\mathbf{x}$) decreases fast with p, since the average of $(x'_{j>p})^2$ over the validation set tends to be larger than in the inversion set for low T, i.e., larger than its ML value $\bar{\lambda}_p^{(\mathrm{ml})}$ (as predicted by the Marchenko-Pastur equation). Raising the value of $\bar{\lambda}_p > \bar{\lambda}_p^{(\mathrm{ml})}$, one takes into account this fact. Therefore, the resulting value of the cross-validated p^* tends to be larger in caut-PCA. As a consequence (see Fig. 5), a larger number p^* of eigenvalues is more similar to the sample (and, more importantly, to the oracle) spectrum.

For reference, in Fig. 5 we also compare the methods PCA (CV) and caut-PCA (CV) with shrinkage (CV), RIE (CV), GAW, raw, and oracle.

APPENDIX C: SYSTEMATIC RESULTS FOR THE SYNTHETIC DATA SET

We here present some complementary results of the synthetic simulations. In Fig. 6 we show d versus α_{D} for various algorithms, and different values of T_{tr} (in different panels). This is a different perspective of the same data of Figs. 1 and 2, but for more values of α_{D}. We show $d(C^{\mathrm{v}}, C)$ versus α_{D} for a single value of $T_{\mathrm{tr}} = 300$ in Fig. 7. Intermethod differences are less significant, with respect to their statistical errors, than for $d(J^{\mathrm{v}}, J)$. The same occurs with the test-likelihood (Fig. 8) and the test-completion error (Fig. 9). Figures 10, 11 present the likelihood versus the completion error for $T_{\mathrm{tr}} = 300$, $\alpha_{\mathrm{D}} = 1, 3$, respectively.

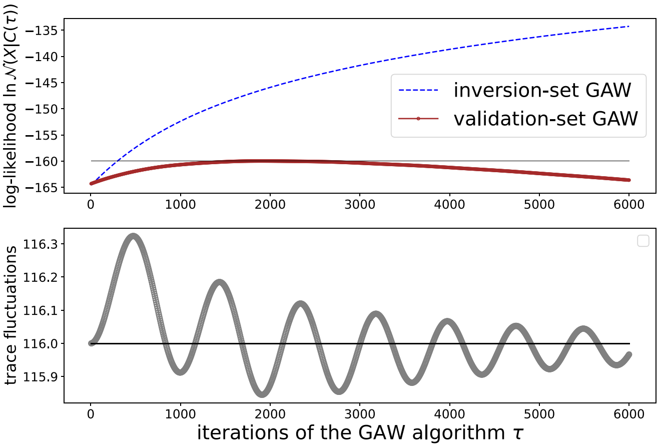

FIG. 12. Illustration of the GAW algorithm. Upper panel: Validation- and inversion-set likelihoods versus the number of iterations τ. Lower panel: The trace of the resulting covariance matrix $\mathrm{tr}(C(\tau))$ versus τ. In this example, the database X is a synthetic data set with $\alpha_{\rm D} = 3$, $T_{\rm tr} = 144$, $T_{\rm in} = (5/6)T_{\rm tr}$. The algorithm parameters are $\eta_{\rm GA} = 10^{-4}$, $\eta_\lambda = 5\,10^{-2}$.

APPENDIX D: EARLY-STOPPING GRADIENT ASCENT ALGORITHMS

We now describe the GAW algorithm. We perform a GA search of $\ln\mathcal{N}(X^{(\rm in)}|C)$ on the $N\times N$ real matrix Y, defined such that $C^{-1} = YY^\dagger$. Performing the GA search on Y guarantees the positive-definiteness of the precision matrix $YY^\dagger$ at each iteration. One first takes an initial condition in the first $\tau = 0$ step, $Y(0) = 1_N$; afterward we follow the simple gradient iteration for the rs element of matrix Y,

$$Y_{rs}(\tau+1) - Y_{rs}(\tau) = \eta_{\rm GA}\frac{\partial}{\partial Y'_{rs}}\bigg|_{Y(\tau)} \ln\mathcal{N}(X^{(\rm in)}|C), \tag{D1}$$

where $C = (Y'Y'^\dagger)^{-1}$ and where the *learning rate* $\eta_{\rm GA}$ is a small, positive parameter. The gradient in Eq. (D1) takes the form

$$\frac{\partial}{\partial Y'_{rs}}\ln\mathcal{N}(X^{(\rm in)}|C) \tag{D2}$$

$$= (CY' + (CY')^\dagger) - (EY' + (EY')^\dagger)_{rs}, \tag{D3}$$

where E is the unbiased estimator of the covariance matrix given $X^{(\rm in)}$. The term CY may be computed from Y using the singular value decomposition $Y = W\Lambda^{-1/2}Z$, where W, Z are unitary matrices and Λ is the diagonal eigenvalue matrix of C, afterward taking $CY = W\Lambda^{1/2}Z$. The iterations stop when the quality criterion $Q(X^{(\rm va)}|C(\tau))$ decreases from the τth to the $\tau + 1$th iteration, and the cleaned correlation matrix is taken as $C(\tau)$.

The SGA algorithm is based on the above described (deterministic) GA algorithm but, instead of the deterministic gradient ascent of Eq. (D1), we use a stochastic gradient ascent rule, inspired in artificial neural network learning,

$$Y(\tau+1) - Y(\tau) = \eta(M(\tau) + M^\dagger(\tau)), \tag{D4}$$

$$M(\tau) = (C(\tau) - E^{(\tau)})Y(\tau), \tag{D5}$$

where $E^{(\tau)}$ is a random bootstrapping of the sample correlation matrix, consisting of the covariance matrix of a subset of $B \leqslant T$ sample vectors composing the training set, with repeating indices. In other words, at each iteration of the GA algorithm, the sample term of the gradient in Eq. (D1) is not constant, but calculated with a random bootstrapping of the data, different at each epoch. In this case, the stopping criterion is consequently modified: the iterations stop when $Q(X^{(\rm va)}|Q(\tau))$ decreases for $\tau_{\rm d}$ consecutive iterations.

Adding the constant trace constraint. Suppose that Y^* is the solution satisfying $Y^* = \arg\max[\mathcal{N}(X^{(\rm in)}|(YY^\dagger)^{-1})]$ subject to the constraint $\mathrm{tr}(C) = N$ with $C = (YY^\dagger)^{-1}$. Then, the Y^* satisfies

$$\partial_Y[\ln\mathcal{N}(X^{(\rm in)}|C) - \mu\,[\mathrm{tr}(C) - N]]|_{Y^*,\mu^*} = 0, \tag{D6}$$

$$\partial_\mu[\ln\mathcal{N}(X^{(\rm in)}|C) - \mu\,[\mathrm{tr}(C) - N]]|_{Y^*,\mu^*} = 0, \tag{D7}$$

where μ is the Lagrange multiplier associated to the constraint. We have, hence, a further scalar variable and a further equation in the satisfaction problem (that we solve only approximately, since we apply the early stopping criterion). The two above coupled equations induce the following Euler iterative dynamics in the variables Y, μ:

$$\mu(\tau+1) - \mu(\tau) = \pm\eta_\mu\,[N - \mathrm{tr}(C(\tau))], \tag{D8}$$

$$Y(\tau+1) - Y(\tau) = \eta_{\rm GA}\,[M(\tau) + M^\dagger(\tau)], \tag{D9}$$

$$M(\tau) := C(\tau)Y(\tau) - EY(\tau) + 2\mu(\tau)C^2(\tau)Y(\tau), \tag{D10}$$

where η_μ is the learning rate associated to the updating of μ that we set constant and larger than $\eta_{\rm GA}$ (in the numerical calculations, we actually set $\eta_\mu = 10\eta_{\rm GA}$).

We show the validation and inversion likelihood as a function of the number of iterations in the GAW algorithm in Fig. 12. When the validation-set likelihood reaches its maximum value (horizontal line in Fig. 12), the iterations stop and the resulting C is taken as the regularized matrix. The lower panel of the figure shows the behavior of $\mathrm{tr}(C(\tau))$ and its oscillations around its required value $N = 116$. Increasing the value of η_λ reduces the amplitude of the oscillations in $\mathrm{tr}(C)$, but this does not have a statistically significant impact on the results for the subject-averaged values of d, ℓ.

APPENDIX E: DETAILS OF THE NUMERICAL SIMULATIONS

We present some details of the numerical algorithms (see Ref. [58]). The array of values of the cross-validated hyperparameters is $p = 1, \ldots, N-1$ for PCA and caut-PCA; for FA, the hyperparameter r takes the same values; for shrinkage, $\alpha \in \{1 - 10^x\}$ with x taking 30 equally spaced values between -2 and -0.1; for RIE, $\eta \in x N^{-1/2}$ with x taking the values 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100. For the Lasso algorithm, we have employed the scikit-learn implementation [68], called GraphicalLassoCV, with an initial four-length grid with four refinements and 1000 maximum number of iterations. Also, for FA and shrinkage, we use the scikit-learn versions. For the GA and GAW algorithms we employ $\eta_{\rm GA} = 10^{-4}$ and $\eta_\mu = 10^{-2}$. For the stochastic version, SGAW, we employ a batch size $B = T_{\rm in}/4$.

[1] J. Cabral, M. L. Kringelbach, and G. Deco, Functional connectivity dynamically evolves on multiple time-scales over a static structural connectome: Models and mechanisms, NeuroImage **160**, 84 (2017), functional Architecture of the Brain.

[2] C. J. Honey, O. Sporns, L. Cammoun, X. Gigandet, J. P. Thiran, R. Meuli, and P. Hagmann, Predicting human resting-state functional connectivity from structural connectivity, Proc. Natl. Acad. Sci. USA **106**, 2035 (2009).

[3] M. P. van den Heuvel, R. C. Mandl, R. S. Kahn, and H. E. Hulshoff Pol, Functionally linked resting-state networks reflect the underlying structural connectivity architecture of the human brain, Hum. Brain Mapp. **30**, 3127 (2009).

[4] K. J. Friston, Functional and effective connectivity: A review, Brain Connect. **1**, 13 (2011).

[5] H.-J. Park and K. Friston, Structural and functional brain networks: From connections to cognition, Science **342**, 1238411 (2013).

[6] A. Demertzi, E. Tagliazucchi, S. Dehaene, G. Deco, P. Barttfeld, F. Raimondo, C. Martial, D. Fernández-Espejo, B. Rohaut, H. U. Voss, N. D. Schiff, A. M. Owen, S. Laureys, L. Naccache, and J. D. Sitt, Human consciousness is supported by dynamic complex patterns of brain signal coordination, Sci. Adv. **5**, eaat7603 (2019).

[7] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon, Functional connectivity in the resting brain: A network analysis of the default mode hypothesis, Proc. Natl. Acad. Sci. USA **100**, 253 (2003).

[8] S. Gu, R. F. Betzel, M. G. Mattar, M. Cieslak, P. R. Delio, S. T. Grafton, F. Pasqualetti, and D. S. Bassett, Optimal trajectories of brain state transitions, NeuroImage **148**, 305 (2017).

[9] M. Gilson, R. Moreno-Bote, A. Ponce-Alvarez, P. Ritter, and G. Deco, Estimation of directed effective connectivity from fMRI functional connectivity hints at asymmetries of cortical connectome, PLoS Comput. Biol. **12**, e1004762 (2016).

[10] F. Abdelnour, M. Dayan, O. Devinsky, T. Thesen, and A. Raj, Functional brain connectivity is predictable from anatomic network's Laplacian eigen-structure, NeuroImage **172**, 728 (2018).

[11] R. Liégeois, A. Santos, V. Matta, D. Van De Ville, and A. H. Sayed, Revisiting correlation-based functional connectivity and its relationship with structural connectivity, Network Neurosci. **4**, 1235 (2020).

[12] F. Morone, K. Roth, B. Min, H. E. Stanley, and H. A. Makse, Model of brain activation predicts the neural collective influence map of the brain, Proc. Natl. Acad. Sci. USA **114**, 3849 (2017).

[13] I. Fortel, M. Butler, L. E. Korthauer, L. Zhan, O. Ajilore, I. Driscoll, A. Sidiropoulos, Y. Zhang, L. Guo, H. Huang, D. Schonfeld, and A. Leow, Brain dynamics through the lens of statistical mechanics by unifying structure and function, in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*, edited by D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan (Springer International Publishing, Cham, 2019), pp. 503–511

[14] T. Watanabe, S. Hirose, H. Wada, Y. Imai, T. Machida, I. Shirouzu, S. Konishi, Y. Miyashita, and N. Masuda, A pairwise maximum entropy model accurately describes resting-state human brain networks, Nat. Commun. **4**, 1370 (2013).

[15] I. Fortel, M. Butler, L. E. Korthauer, L. Zhan, O. Ajilore, A. Sidiropoulos, Y. Wu, I. Driscoll, D. Schonfeld, and A. Leow, Inferring excitation-inhibition dynamics using a maximum entropy model unifying brain structure and function, Network Neurosci. **6**, 420 (2022).

[16] W. Niu, X. Huang, K. Xu, T. Jiang, and S. Yu, Pairwise interactions among brain regions organize large-scale functional connectivity during execution of various tasks, Neuroscience **412**, 190 (2019).

[17] P. M. Abeyasinghe, Structure-function relationship of the brain: A comparison between the 2D classical Ising model and the generalized Ising model, Master's thesis, The University of Western Ontario, 2015.

[18] B. Kadirvelu, Y. Hayashi, and S. J. Nasuto, Inferring structural connectivity using ising couplings in models of neuronal networks, Sci. Rep. **7**, 8156 (2017).

[19] G. Hahn, M. A. Skeide, D. Mantini, M. Ganzetti, A. Destexhe, A. D. Friederici, and G. Deco, A new computational approach to estimate whole-brain effective connectivity from functional and structural MRI, applied to language development, Sci. Rep. **9**, 1 (2019).

[20] T. K. Das, P. M. Abeyasinghe, J. S. Crone, A. Sosnowski, S. Laureys, A. M. Owen, and A. Soddu, Highlighting the structure-function relationship of the brain with the Ising model and graph theory, BioMed Res. Int. **2014**, 237898 (2014).

[21] G. Deco, M. Senden, and V. Jirsa, How anatomy shapes dynamics: a semi-analytical study of the brain at rest by a simple spin model, Front. Comput. Neurosci. **6**, 1 (2012).

[22] P. Wang, R. Kong, X. Kong, R. Liëgeois, C. Orban, G. Deco, M. P. van den Heuvel, and B. T. Yeo, Inversion of a large-scale circuit model reveals a cortical hierarchy in the dynamic resting human brain, Sci. Adv. **5**, eaat7854 (2019).

[23] J. F. Strain, M. R. Brier, A. Tanenbaum, B. A. Gordon, J. E. McCarthy, A. Dincer, D. S. Marcus, J. P. Chhatwal, N. R. Graff-Radford, G. S. Day, C. la Fougére, R. J. Perrin, S. Salloway, P. R. Schofield, I. Yakushev, T. Ikeuchi, J. Vöglein, J. C. Morris, T. L. Benzinger, R. J. Bateman *et al.*, Covariance-based vs. correlation-based functional connectivity dissociates healthy aging from Alzheimer disease, NeuroImage **261**, 119511 (2022).

[24] F. Deligianni, G. Varoquaux, B. Thirion, E. Robinson, D. J. Sharp, A. D. Edwards, and D. Rueckert, A probabilistic framework to infer brain functional connectivity from anatomical connections, in *Information Processing in Medical Imaging*, edited by G. Székely and H. K. Hahn (Springer, Berlin, 2011), pp. 296–307.

[25] U. Pervaiz, D. Vidaurre, M. W. Woolrich, and S. M. Smith, Optimising network modelling methods for fMRI, NeuroImage **211**, 116604 (2020).

[26] H. Liu, H. Hu, H. Wang, J. Han, Y. Li, H. Qi, M. Wang, S. Zhang, H. He, and X. Zhao, A brain network constructed on an l1-norm regression model is more sensitive in detecting small world network changes in early AD, Neural Plast. **2020**, 9436406 (2020).

[27] S. M. Smith, D. Vidaurre, C. F. Beckmann, M. F. Glasser, M. Jenkinson, K. L. Miller, T. E. Nichols, E. C. Robinson, G. Salimi-Khorshidi, M. W. Woolrich *et al.*, Functional connectomics from resting-state fMRI, Trends Cognit. Sci. **17**, 666 (2013).

[28] K. Dadi, M. Rahim, A. Abraham, D. Chyzhyk, M. Milham, B. Thirion, and G. Varoquaux, Benchmarking functional

connectome-based predictive models for resting-state fMRI, NeuroImage **192**, 115 (2019).

[29] J. Chung, B. S. Jackson, J. E. McDowell, and C. Park, Joint estimation and regularized aggregation of brain network in fMRI data, J. Neurosci. Methods **364**, 109374 (2021).

[30] M. Rahim, B. Thirion, and G. Varoquaux, Population shrinkage of covariance (PoSCE) for better individual brain functional-connectivity estimation, Med. Image Anal. **54**, 138 (2019).

[31] S. Ryali, T. Chen, K. Supekar, and V. Menon, Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty, NeuroImage **59**, 3852 (2012).

[32] M. R. Brier, A. Mitra, J. E. McCarthy, B. M. Ances, and A. Z. Snyder, Partial covariance based functional connectivity computation using Ledoit-Wolf covariance regularization, NeuroImage **121**, 29 (2015).

[33] J. Lefort-Besnard, D. S. Bassett, J. Smallwood, D. S. Margulies, B. Derntl, O. Gruber, A. Aleman, R. Jardri, G. Varoquaux, B. Thirion, S. B. Eickhoff, and D. Bzdok, Different shades of default mode disturbance in schizophrenia: Subnodal covariance estimation in structure and function, Human Brain Mapping **39**, 644 (2018).

[34] D. Cordes and R. R. Nandy, Estimation of the intrinsic dimensionality of fMRI data, NeuroImage **29**, 145 (2006).

[35] A. F. Mejia, M. B. Nebel, A. D. Barber, A. S. Choe, and M. A. Lindquist, Effects of scan length and shrinkage on reliability of resting-state functional connectivity in the human connectome project, arXiv:1606.06284 [stat.AP] (2016).

[36] F. Deligianni, M. Centeno, D. W. Carmichael, and J. D. Clayden, Relating resting-state fMRI and EEG whole-brain connectomes across frequency bands, Front. Neurosci. **8**, 258 (2014).

[37] G. Varoquaux, A. Gramfort, J.-B. Poline, and B. Thirion, Brain covariance selection: better individual functional connectivity models using population prior, *Advances in Neural Information Processing Systems*, edited by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Vol. 23 (Curran Associates, Inc., 2010).

[38] R. Ciric, A. F. Rosen, G. Erus, M. Cieslak, A. Adebimpe, P. A. Cook, D. S. Bassett, C. Davatzikos, D. H. Wolf, and T. D. Satterthwaite, Mitigating head motion artifact in functional connectivityMRI, Nat. Protocols **13**, 2801 (2018).

[39] J. Bun, J.-P. Bouchaud, and M. Potters, My beautiful laundrette: Cleaning correlation matrices for portfolio optimization, 2016, doi:10.13140/RG.2.1.2782.2961.

[40] J. Bun, J.-P. Bouchaud, and M. Potters, Cleaning large correlation matrices: Tools from random matrix theory, Phys. Rep. **666**, 1 (2017).

[41] G. Livan, M. Novaes, and P. Vivo, *Introduction to Random Matrices*, 1st ed. SpringerBriefs in Mathematical Physics (BRIEFSMAPHY) Vol. 26 (Springer, Cham, 2018), p. 124.

[42] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to machine learning for physicists, Phys. Rep. **810**, 1 (2019), a high-bias, low-variance introduction to Machine Learning for physicists.

[43] D. J. C. MacKay, *Information Theory, Inference and Learning Agorithms* (Cambridge University Press, Cambridge, 2003).

[44] T. P. Minka, Automatic choice of dimensionality for PCA, in *Proceedings of the 13th International Conference on Neural Information Processing Systems* (MIT Press, United States, 2000), pp. 577–583.

[45] L. Haff, Empirical bayes estimation of the multivariate normal covariance matrix, Ann. Stat. **8**, 586 (1980).

[46] J.-P. Bouchaud and M. Potters, *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management* (Cambridge University Press, Cambridge, 2003).

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

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

[49] D. Barber, *Bayesian Reasoning and Machine Learning* (Cambridge University Press, New York, USA, 2012).

[50] 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).

[51] J. Friedman, T. Hastie, and R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics **9**, 432 (2008).

[52] L. Zhou, L. Wang, and P. Ogunbona, Discriminative sparse inverse covariance matrix: Application in brain functional network classification, in *2014 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2014), pp. 3097–3104.

[53] R. Mastrandrea, A. Gabrielli, F. Piras, G. Spalletta, G. Caldarelli, and T. Gili, Organization and hierarchy of the human functional brain network lead to a chain-like core, Sci. Rep. **7**, 4888 (2017).

[54] R. Mastrandrea, F. Piras, A. Gabrielli, N. Banaj, G. Caldarelli, G. Spalletta, and T. Gili, The unbalanced reorganization of weaker functional connections induces the altered brain network topology in schizophrenia, Sci. Rep. **11**, 15400 (2021).

[55] CamCAN Data Repository, http://www.mrc-cbu.cam.ac.uk/datasets/camcan/.

[56] M. A. Shafto, L. K. Tyler, M. Dixon, J. R. Taylor, J. B. Rowe, R. Cusack, A. J. Calder, W. D. Marslen-Wilson, J. Duncan, T. Dalgleish *et al.*, The cambridge centre for ageing and neuroscience (cam-can) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing, BMC Neurology **14**, 204 (2014).

[57] J. R. Taylor, N. Williams, R. Cusack, T. Auer, M. A. Shafto, M. Dixon, L. K. Tyler, Cam-CAN, and R. N. Henson, The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample, NeuroImage **144**, 262 (2017), data Sharing Part II.

[58] C. Lucibello and M. Ibanez-Berganza, Covariance estimators (2022), https://github.com/CarloLucibello/covariance-estimators.

[59] See the Python-Jupyter notebook plot_estimators.ipynb in Ref. [58].

[60] Z. Drogosz, J. Jurkiewicz, G. Lukaszewski, and M. A. Nowak, Comparison of eigeninference based on one- and two-point Green's functions, Phys. Rev. E **92**, 022111 (2015).

[61] T. O. Laumann, A. Z. Snyder, A. Mitra, E. M. Gordon, C. Gratton, B. Adeyemo, A. W. Gilmore, S. M. Nelson, J. J. Berg, D. J. Greene, J. E. McCarthy, E. Tagliazucchi, H. Laufs, B. L. Schlaggar, N. U. F. Dosenbach, and S. E. Petersen, On the

Stability of BOLD fMRI Correlations, Cereb. Cortex **27**, 4719 (2016).

[62] S. Frässle, Z. M. Manjaly, C. T. Do, L. Kasper, K. P. Pruessmann, and K. E. Stephan, Whole-brain estimates of directed connectivity for human connectomics, NeuroImage **225**, 117491 (2021).

[63] S. Frässle, Y. Yao, D. Schöbi, E. A. Aponte, J. Heinzle, and K. E. Stephan, Generative models for clinical applications in computational psychiatry, Wiley Interdiscip. Rev.: Cognit. Sci. **9**, e1460 (2018).

[64] L. Rigoux and J. Daunizeau, Dynamic causal modelling of brain-behaviour relationships, NeuroImage **117**, 202 (2015).

[65] K. H. Brodersen, T. M. Schofield, A. P. Leff, C. S. Ong, E. I. Lomakina, J. M. Buhmann, and K. E. Stephan, Generative embedding for model-based classification of fMRI data, PLoS Comput. Biol. **7**, e1002079 (2011).

[66] K. E. Stephan, L. Kasper, L. M. Harrison, J. Daunizeau, H. E. den Ouden, M. Breakspear, and K. J. Friston, Nonlinear dynamic causal models for fMRI, NeuroImage **42**, 649 (2008).

[67] http://www.sobigdata.eu.

[68] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).