Large-deviation theory for diluted Wishart random matrices

Isaac Pérez Castillo
Department of Quantum Physics and Photonics, Institute of Physics, UNAM, P.O. Box 20-364, 01000 Mexico City, Mexico and London Mathematical Laboratory, 14 Buckingham Street, London WC2N 6DF, United Kingdom

Fernando L. Metz
*Institute of Physics, Federal University of Rio Grande do Sul, 91501-970 Porto Alegre, Brazil;
Physics Department, Federal University of Santa Maria, 97105-900 Santa Maria, Brazil;
and London Mathematical Laboratory, 14 Buckingham Street, London WC2N 6DF, United Kingdom*

(Received 11 January 2018; published 19 March 2018)

Wishart random matrices with a sparse or diluted structure are ubiquitous in the processing of large datasets, with applications in physics, biology, and economy. In this work, we develop a theory for the eigenvalue fluctuations of diluted Wishart random matrices based on the replica approach of disordered systems. We derive an analytical expression for the cumulant generating function of the number of eigenvalues $\mathcal{I}_N(x)$ smaller than $x \in \mathbb{R}^+$, from which all cumulants of $\mathcal{I}_N(x)$ and the rate function $\Psi_x(k)$ controlling its large-deviation probability $\text{Prob}[\mathcal{I}_N(x) = kN] \asymp e^{-N\Psi_x(k)}$ follow. Explicit results for the mean value and the variance of $\mathcal{I}_N(x)$, its rate function, and its third cumulant are discussed and thoroughly compared to numerical diagonalization, showing very good agreement. The present work establishes the theoretical framework put forward in a recent letter [Phys. Rev. Lett. **117**, 104101 (2016)] as an exact and compelling approach to deal with eigenvalue fluctuations of sparse random matrices.

DOI: 10.1103/PhysRevE.97.032124

I. INTRODUCTION

In the past few decades, we have experienced an explosion of available information—the so-called Big Data era. Problems in modern data analysis usually involve a large number of variables and observations, posing new challenges in the processing of data. This high dimensionality of the dataset typically occurs in climate studies, genetics, biomedical imaging, and economics [1].

Suppose one performs P measurements of N variables characterizing a system. For instance, the variables could be assets in a stock market or a collection of climate observables, while the measurements of all variables could be simultaneously performed for P different times. The collected data can be organized in an $N \times P$ matrix $\boldsymbol{\xi}$, with the element ξ_{ij} providing the measurement j of the variable i. The $N \times N$ sample covariance matrix $\boldsymbol{M}$ of the dataset is built from the product $\boldsymbol{M} = \boldsymbol{\xi}\boldsymbol{\xi}^T$ and it encodes all possible correlations among the variables. The covariance matrix is at the core of multivariate statistical analysis, with applications in dimensional reduction methods and classifying procedures, such as principal component [2] and linear discriminant analyses [3], respectively.

Generally speaking, it is reasonable to expect that in many natural phenomena, each variable is significantly correlated with only a few others, giving rise to sparse covariance matrices, whose main feature is the presence of a large amount of entries that are very small or even zero. In this context, an important example is the problem of inferring, from the empirical covariance matrix, the causal influences among the individual components of a system. This is typically the case, for instance, in the experimental reconstruction of the interactions between the elements of biological systems, such as cellular signaling networks [4], gene regulatory networks [5], and ecological association networks [6,7]. Besides that, performing numerical tasks with large covariance matrices, where all entries are strictly nonzero, is computationally very expensive, and, in this case, one usually resorts to regularization techniques in order to bring the matrix into a sparse form [8].

Since the pioneering work of Wishart [9], random matrix theory has been playing a pivotal role in multivariate statistics [10]. Essentially, results derived from random matrix models serve as important benchmarks through which comparisons with real data can be made. The simplest null model for the covariance matrix $\boldsymbol{M}$ consists in assuming that the entries of $\boldsymbol{\xi}$ are independent Gaussian random variables. In this case, the joint distribution of the eigenvalues of $\boldsymbol{M}$ is analytically known, which forms the starting point to employ the Coulomb gas technique [11–13] and derive a wealth of quantitative information about the eigenvalue statistics of the covariance random matrix [14–17], including a detailed account of the typical and atypical fluctuations of its eigenvalues [16]. Unfortunately, apart from the averaged spectral density [18], much less is known about the eigenvalue statistics of sparse covariance matrices. The main reason lies in the absence of an analytical expression for the joint distribution of eigenvalues, which hampers the application of the Coulomb gas approach. This is a general problem in ensembles of diluted random matrices, and, although novel approaches have helped us to push forward the understanding of the eigenvalue statistical properties in these ensembles [18–27], from an analytical viewpoint we still have a long way to go when compared to classical ensembles of random matrices.

In this work, we develop an analytical approach to quantify the eigenvalue fluctuations of sparse covariance random matrices. By relying on a recent technique [28], based on the replica method of disordered systems [29], we derive an analytical expression for the large-N behavior of the cumulant generating function of the number of eigenvalues $\mathcal{I}_N(x)$ smaller than a certain threshold $x \in \mathbb{R}^+$. This function gives access to all cumulants of the random variable $\mathcal{I}_N(x)$ as well as to its large-deviation probability, providing a full picture of the eigenvalue fluctuations for this class of random matrices. From the numerical solution of our analytical equations, we present explicit results for the mean and the variance of $\mathcal{I}_N(x)$, its third cumulant, and its rate function governing the large-deviation probability. We show that, similarly to the atypical eigenvalue fluctuations of the adjacency matrix of random graphs [30], the rate function of $\mathcal{I}_N(x)$ is asymmetric around its minimum, which characterizes an unbalanced contribution of rare samples responsible for increasing or decreasing $\mathcal{I}_N(x)$ with respect to its typical value. The exactness of our theoretical findings is fully supported by direct diagonalization of finite random matrices.

In the next section, we define the ensemble of sparse Wishart matrices and the main quantity we consider in order to study eigenvalue fluctuations. In Sec. III we present explicit results derived from our theoretical approach, while in the final section we discuss some final remarks. All technical details are concentrated in two appendixes. Appendix A explains all steps involved in the derivation of the cumulant generating function of $\mathcal{I}_N(x)$. In that appendix, we also discuss some mathematical subtleties regarding the representation of $\mathcal{I}_N(x)$ in terms of the complex logarithm. In Appendix B, we describe the algorithmic approach employed to solve numerically the main analytical equations obtained from the theory.

II. RANDOM MATRIX MODEL AND THE GENERAL SETTING

As we are interested in the ensemble of sparse or diluted Wishart matrices, we need to decide a way to introduce dilution in the Gaussian or classical Wishart ensemble. We proceed as follows: we consider rectangular $N \times P$ matrices $\boldsymbol{\xi}$ whose entries are independent and identically distributed random variables drawn from the distribution

$$p(\xi_{i\mu}) = \frac{d}{N} P_\xi(\xi_{i\mu}) + \left(1 - \frac{d}{N}\right)\delta(\xi_{i\mu}), \tag{1}$$

where $P_\xi(\xi)$ is the probability density for the nonzero entries of $\boldsymbol{\xi}$. From a graph viewpoint [31], the random matrix $\boldsymbol{\xi}$ corresponds to the adjacency matrix of a weighted Poissonian bipartite random graph with two types of nodes [18]: i-nodes, associated with the row index of $\boldsymbol{\xi}$, and μ-nodes, associated with the column index of $\boldsymbol{\xi}$. The total number of nodes is $N + P$ and the average degree of the μ-nodes is d, while $c = \alpha d$ is the average degree of the i-nodes, with $\alpha = P/N$. Now we introduce the ensemble of $N \times N$ symmetric diluted Wishart random matrices, in which the entries M_{ij} of a given covariance random matrix $\boldsymbol{M}$ are obtained from

$$M_{ij} = \frac{1}{d}\sum_{\mu=1}^{P} \xi_{i\mu}\xi_{j\mu}. \tag{2}$$

Defining $\lambda_1, \ldots, \lambda_N$ as the N positive eigenvalues of $\boldsymbol{M}$, we are interested here in the statistics of the random variable,

$$\mathcal{I}_N(x) = \sum_{\alpha=1}^{N} \Theta(x - \lambda_\alpha), \tag{3}$$

which counts the number of eigenvalues smaller than a threshold $x \in \mathbb{R}^+$ (Θ is the Heaviside function). The main goal here is to study the cumulants of $\mathcal{I}_N(x)$ and the rate function controlling its large-deviation probability. For the random matrix model defined by Eqs. (1) and (2), the main difficulty in pursuing an analytical study of the fluctuations of $\mathcal{I}_N(x)$ lies in the absence of an invariance property of the ensemble and the corresponding unknown analytical form of the joint distribution of eigenvalues [32]. Therefore, the analytical method of [16] is simply inapplicable. In spite of that, one may still hope to derive analytical results for the statistics of $\mathcal{I}_N(x)$, provided one finds an explicit link between $\mathcal{I}_N(x)$ and the random matrix $\boldsymbol{M}$. Such a link is obtained by using the representation of the Heaviside function in terms of the discontinuity of the principal value of the complex logarithm,

$$\mathcal{I}_N(x) = \frac{1}{2\pi i} \lim_{\epsilon \to 0^+} \sum_{\alpha=1}^{N} [\mathrm{Ln}(\lambda_\alpha - x_\epsilon) - \mathrm{Ln}(\lambda_\alpha - \overline{x_\epsilon})], \tag{4}$$

where $x_\epsilon = x - i\epsilon$, and $\overline{(\cdots)}$ denotes the complex conjugate.

The above identity is the starting point to study the statistics of $\mathcal{I}_N(x)$. The cumulant generating function of $\mathcal{I}_N(x)$ is defined as

$$\mathcal{F}_x(y) = -\lim_{N\to\infty} \frac{1}{N} \mathrm{Ln}\langle e^{-y\mathcal{I}_N(x)}\rangle_{\boldsymbol{M}}, \tag{5}$$

in which $\langle \cdots \rangle_{\boldsymbol{M}}$ is the ensemble average with the distribution of the random matrix $\boldsymbol{M}$. The ℓ-cumulant of $\mathcal{I}_N(x)$, defined as

$$\kappa_\ell(x) = \frac{\langle \mathcal{I}_N^\ell(x)\rangle_{\boldsymbol{M}}^c}{N} \tag{6}$$

in terms of the connected correlation $\langle \cdots \rangle^c$, is obtained from $\mathcal{F}_x(y)$ according to

$$\kappa_\ell(x) = (-1)^{\ell+1} \frac{\partial^\ell \mathcal{F}_x(y)}{\partial y^\ell}\bigg|_{y=0}. \tag{7}$$

Invoking the Gärtner-Ellis theorem [33], the probability of having kN eigenvalues smaller than x behaves asymptotically for large N as

$$\mathrm{Prob}[\mathcal{I}_N(x) = kN] \asymp e^{-N\Psi_x(k)}, \tag{8}$$

where the rate function $\Psi_x(k)$ is obtained from $\mathcal{F}_x(y)$ by the Legendre transform:

$$-\Psi_x(k) = \inf_{y\in\mathbb{R}}[ky - \mathcal{F}_x(y)]. \tag{9}$$

Thus, all cumulants of $\mathcal{I}_N(x)$ and its large-deviation probability follow from the analytical expression for $\mathcal{F}_x(y)$, presented in the next section.

III. RESULTS

The exponential form of $\mathcal{F}_x(y)$ [see Eq. (5)], when combined with the identity (4), is suitable for the application of the replica approach in order to calculate the ensemble average and the limit $N \to \infty$ in Eq. (5). We thoroughly explain all steps of such a calculation in Appendix A, while here we only present the final outcome, namely the analytical expression for the cumulant generating function:

$$
\begin{aligned}
\mathcal{F}_x(y) = {} & \mathcal{A} \int d\Delta\, d\Gamma\, w_\rho(\Delta) w_k(\Gamma) \exp\left[-\frac{iy}{2\pi} \mathrm{Ln}\left(\frac{\Delta + \frac{1}{\Gamma}}{\overline{\Delta} + \frac{1}{\overline{\Gamma}}}\right)\right] \\
& - \alpha \mathrm{Ln}\left\{\int d\sigma\, w_\sigma(\sigma) \exp\left[-\frac{iy}{2\pi}\mathrm{Ln}\left(\frac{1+\sigma}{1+\overline{\sigma}}\right)\right]\right\} \\
& - \mathrm{Ln}\left\{\sum_{\ell=0}^{\infty} \frac{\mathcal{A}^\ell}{\ell!} \int \left(\prod_{s=1}^{\ell} d\Gamma_s w_k(\Gamma_s)\right)\right. \\
& \left. \times \exp\left[\frac{iy}{2\pi} W_\epsilon(\Gamma_1,\ldots,\Gamma_\ell)\right]\right\}. \qquad (10)
\end{aligned}
$$

The weight $W_\epsilon(\Gamma_1,\ldots,\Gamma_\ell)$ in the above equation is given by

$$
W_\epsilon(\Gamma_1,\ldots,\Gamma_\ell) = \sum_{s=1}^{\ell} \mathrm{Ln}\left(\frac{\Gamma_s}{\overline{\Gamma_s}}\right) - \mathrm{Ln}\left(\frac{\sum_{s=1}^{\ell} \Gamma_s - x_\epsilon}{\sum_{s=1}^{\ell} \overline{\Gamma_s} - \overline{x_\epsilon}}\right), \qquad (11)
$$

while $w_\rho(\Delta)$, $w_k(\Gamma)$, and $w_\sigma(\sigma)$ are the distributions of the complex variables Δ, Γ, and σ. These distributions solve the following set of self-consistency equations:

$$
\begin{aligned}
w_\rho(\Delta) = {} & \frac{1}{\mathcal{N}_1} \sum_{\ell=0}^{\infty} \frac{e^{-\mathcal{A}}\mathcal{A}^\ell}{\ell!} \int \left(\prod_{s=1}^{\ell} d\Gamma_s w_k(\Gamma_s)\right) \\
& \times \exp\left[\frac{iy}{2\pi} W_\epsilon(\Gamma_1,\ldots,\Gamma_\ell)\right] \delta\left(\Delta - \frac{1}{\sum_{s=1}^{\ell}\Gamma_s - x_\epsilon}\right), \qquad (12)
\end{aligned}
$$

$$
\begin{aligned}
w_\sigma(\sigma) = {} & \sum_{k=0}^{\infty} \frac{e^{-d} d^k}{k!} \int \left(\prod_{\ell=1}^{k} d\Delta_\ell w_\rho(\Delta_\ell) d\xi_\ell P_\xi(\xi_\ell)\right) \\
& \times \delta\left(\sigma - \frac{1}{d}\sum_{\ell=1}^{k} \xi_\ell^2 \Delta_\ell\right), \qquad (13)
\end{aligned}
$$

$$
w_k(\Gamma) = \int d\xi\, P_\xi(\xi) \int d\sigma\, w_\sigma(\sigma) \delta\left[\Gamma - \frac{\xi^2}{d(1+\sigma)}\right]. \qquad (14)
$$

The constant factors $\mathcal{N}_1$ and $\mathcal{A}$, appearing in Eqs. (10) and (12), are defined as follows:

$$
\begin{aligned}
\mathcal{N}_1 = {} & \sum_{\ell=0}^{\infty} \frac{e^{-\mathcal{A}}\mathcal{A}^\ell}{\ell!} \int \left(\prod_{s=1}^{\ell} d\Gamma_s w_k(\Gamma_s)\right) \\
& \times \exp\left[\frac{iy}{2\pi} W_\epsilon(\Gamma_1,\ldots,\Gamma_\ell)\right], \qquad (15)
\end{aligned}
$$

$$
\mathcal{A} = \frac{\alpha d}{\int d\sigma\, w_\sigma(\sigma) \exp\left[-\frac{iy}{2\pi}\mathrm{Ln}\left(\frac{1+\sigma}{1+\overline{\sigma}}\right)\right]}. \qquad (16)
$$

Equation (10) is the main analytical result of our work, since it provides the exact cumulant generating function of $\mathcal{I}_N(x)$ for the sparse Wishart ensemble in the limit $N \to \infty$, from which the rate function and the behavior of all cumulants readily follow. Notice that $\mathcal{F}_x(y)$ depends on the distributions $w_\rho(\Delta)$, $w_k(\Gamma)$, and $w_\sigma(\sigma)$ through the solution of Eqs. (12)–(14), for which there is no closed, analytical form in the general case. Hence we must solve Eqs. (12)–(14) numerically, which can be done very efficiently using the weighted population dynamics algorithm. All steps of this numerical method are carefully discussed in Appendix B. We point out that the limit $\epsilon \to 0^+$ is implicitly assumed in Eqs. (10)–(16).

Equation (10) can be interpreted as a combination of effective moments or cumulant generating functions of the following random variables:

$$
\mathcal{I}_1(\Delta,\Gamma) \equiv \frac{i}{2\pi}\mathrm{Ln}\left(\frac{\frac{1}{\overline{\Gamma}} + \overline{\Delta}}{\frac{1}{\Gamma} + \Delta}\right), \qquad (17)
$$

$$
\mathcal{I}_2(\sigma) \equiv \frac{i}{2\pi}\mathrm{Ln}\left(\frac{1+\overline{\sigma}}{1+\sigma}\right), \qquad (18)
$$

$$
\mathcal{I}_3(\{\Gamma_s\}_{s=1}^{\ell}) \equiv \frac{i}{2\pi} W_\epsilon(\Gamma_1,\ldots,\Gamma_\ell). \qquad (19)
$$

From Eq. (10), we note that the dependence of $\mathcal{F}_x(y)$ on y is explicit, as the generating parameter, and implicit, since the measures w_ρ, w_k, and w_σ also depend on y through the saddle-point Eqs. (12)–(14). However, it can be shown that, for the first two cumulants, the implicit derivative is not needed and one readily obtains

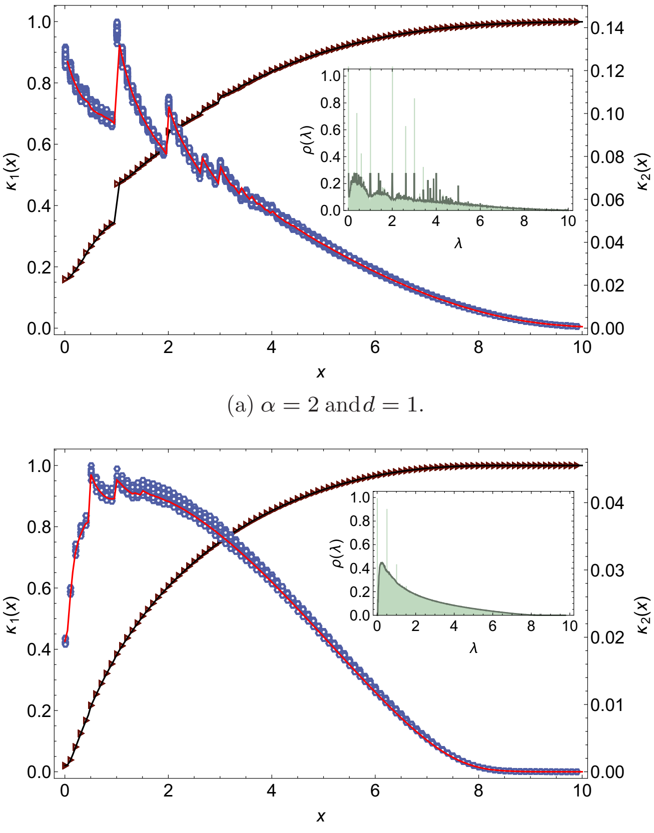
$$
\begin{aligned}
\kappa_1(x) = {} & -\alpha d \langle \mathcal{I}_1(\Delta,\Gamma)\rangle + \big\langle \mathcal{I}_3\big(\{\Gamma_s\}_{s=1}^{\ell}\big)\big\rangle \\
& + \alpha \langle \mathcal{I}_2(\sigma)\rangle, \qquad (20)
\end{aligned}
$$

$$
\begin{aligned}
\kappa_2(x) = {} & \alpha d \big\langle \mathcal{I}_1^2(\Delta,\Gamma)\big\rangle \\
& - \left[\big\langle \mathcal{I}_3^2\big(\{\Gamma_s\}_{s=1}^{\ell}\big)\big\rangle - \big\langle \mathcal{I}_3\big(\{\Gamma_s\}_{s=1}^{\ell}\big)\big\rangle^2\right] \\
& - \alpha\left[\big\langle \mathcal{I}_2^2(\sigma)\big\rangle - \langle \mathcal{I}_2(\sigma)\rangle^2\right], \qquad (21)
\end{aligned}
$$

in which the brackets $\langle \cdots \rangle$ denote the average over the complex random variables Δ, Γ, and σ with their corresponding distributions evaluated at $y = 0$. This average is calculated through the population dynamics method, as explained in Appendix B. Although we do not derive here the analytical equations for higher-order cumulants, in principle one can study their large-N behavior by computing numerically the derivatives of Eq. (10) with respect to y.

Let us now present explicit numerical results obtained from our theoretical approach. Although our equations are valid for arbitrary $\alpha = P/N$, here we limit ourselves to the regime $\alpha > 1$, in which the covariance random matrix is nonsingular and, consequently, more relevant for practical applications. For the sake of simplicity, we present results only for the case $P_\xi(\xi) = \delta(\xi - 1)$.

In Fig. 1 we illustrate the behavior of the first two cumulants of $\mathcal{I}_N(x)$ and compare the outcome of our theory with direct diagonalization of finite random matrices. Since the diluted Wishart ensemble is characterized by two parameters (α,d), we have chosen in Fig. 1 the values $(\alpha,d) = (2,1)$ and $(\alpha,d) = (2,2)$, as the spectral density displays rather distinct features for these two pairs of values: while in the first case the continuous part of the spectrum is bathed in multiple Dirac δ peaks, in the second case the contribution of the discrete eigenvalues is less important. The existence of δ peaks is



(a) $\alpha = 2$ and $d = 1$.

(b) $\alpha = 2$ and $d = 2$.

FIG. 1. The first and second cumulants of $\mathcal{I}_N(x)$ as functions of x for different values of α and d. The solid lines are obtained from the solution of Eqs. (12)–(14) using the weighted population dynamics method with $\mathcal{L} = 10^7$ and $\epsilon = 10^{-8}$ (see Appendix B). The black curve is the mean value of $\mathcal{I}_N(x)$, whereas the red curve gives the variance of $\mathcal{I}_N(x)$. The markers correspond to results coming from numerical diagonalization of an ensemble with 10^4 sparse Wishart matrices of dimension $N = 1000$. The insets display the average spectral density $\rho(\lambda)$ for each pair of model parameters.

a typical feature in the spectral properties of sparse random matrices [18,19,22]. The peaks are commonly located at the eigenvalues of isolated, disconnected trees that appear for small average degrees of the associated random graph model [34]. As clearly shown in Fig. 1(a), the presence of δ peaks in the spectrum manifests itself as discontinuities in the behavior of the first two cumulants, similarly to the analogous results derived for the adjacency matrix of random graphs [30,35].

To further inspect the validity of our theory, we have also compared the third cumulant of $\mathcal{I}_N(x)$, derived from Eq. (10), with direct diagonalization. In this case, we need to take into account the explicit and implicit dependences of the cumulant generating function with respect to y. Instead of trying to derive a set of self-consistency equations for the derivatives of the distributions w_ρ, w_σ, and w_k, we have opted to evaluate the third derivative of $\mathcal{F}_x(y)$ by finite differences, using the algorithm presented in [36]. In Fig. 2 we show the behavior of the third cumulant $\kappa_3(x)$ for $(\alpha,d) = (2,1)$. Apart from small fluctuations, the agreement between our theoretical approach and numerical diagonalization is very good.

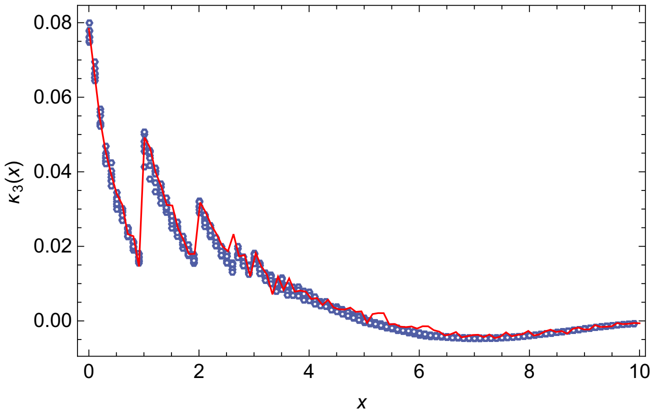
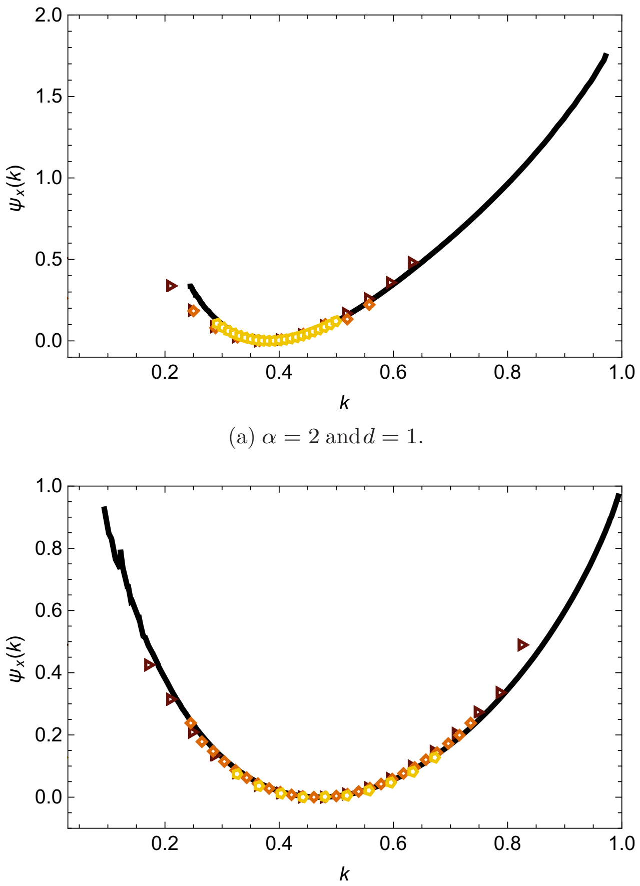


FIG. 2. The third cumulant of $\mathcal{I}_N(x)$ as a function of x for $(\alpha,d) = (2,1)$. The solid line has been obtained by evaluating the cumulant generating function $\mathcal{F}_x(y)$ from Eq. (10) at seven points of y, followed by a finite-differences calculation of the third-order derivative of $\mathcal{F}_x(y)$. These results have been computed through the weighted population dynamics algorithm with $\mathcal{L} = 10^6$ and $\epsilon = 10^{-8}$. Moreover, each estimate of $\mathcal{F}_x(y)$ has been averaged 100 times in order to reduce its fluctuations and improve the accuracy of the finite-differences calculation. The markers correspond to results coming from numerical diagonalization of an ensemble with 10^6 sparse Wishart matrices of dimension $N = 300$.

Let us now turn our attention to the behavior of the rate function $\Psi_x(k)$. Figure 3 illustrates the rate function for $x = 1.01$ and the same combinations of model parameters as discussed in Fig. 1: $(\alpha,d) = (2,1)$ and $(\alpha,d) = (2,2)$. The outcome of our theory is compared with direct diagonalization of finite random matrices, showing an excellent agreement in the range of values of k that can be probed through numerical diagonalization. Indeed, the probability of observing $\mathcal{I}_N(x) = kN$ decays exponentially with N, which promptly hinders any attempt to explore a sizable interval of k for large N, as one needs to diagonalize an unfeasible number of finite random matrices. This remarkable limitation of numerical diagonalization procedures further illustrates the importance of our theory, since we can virtually determine the rate function $\Psi_x(k)$ for any value of $0 \leqslant k \leqslant 1$.

As can be noted in Fig. 3, the function $\Psi_x(k)$ is asymmetric around its minimum, whose value of k coincides with the first cumulant of $\mathcal{I}_N(x)$, calculated from Eq. (20). Such an asymmetric feature is strikingly distinct from the symmetry of the analogous rate function in the classical Wishart ensemble [16]. This seems to be a general property of sparse random matrix ensembles [30], essentially due to the presence of Dirac δ peaks in their spectral density [34]. Concerning Fig. 3(a), we conclude that random matrix samples that increase $\mathcal{I}_N(x)$ are more likely, since the rate function grows slowly for k larger than its typical value. This is due to the presence of a Dirac δ peak with a large weight precisely at $x = 1$, as clearly shown by the discontinuous behavior of $\kappa_1(x)$ at $x = 1$ [see Fig. 1(a)]. One expects that the weights of the discrete contributions to the spectral density decrease exponentially with the average degree d [34], which results in a more symmetric rate function for larger values of d. This picture is consistent with the results in Figs. 1 and 3.



(a) $\alpha = 2$ and $d = 1$.

(b) $\alpha = 2$ and $d = 2$.

FIG. 3. The rate function of the fraction of eigenvalues k inside the interval $(-\infty,x]$ for $x = 1.01$ and two pairs of parameters characterizing the sparse Wishart ensemble. The solid black line is derived from the solution of Eqs. (12)–(14) using the weighted population dynamics method with $\mathcal{L} = 10^6$ and $\epsilon = 10^{-8}$ (see Appendix B). The markers are results obtained from numerical diagonalization of an ensemble with 10^6 sparse Wishart matrices of dimensions $N = 25$ (brown), $N = 50$ (orange), and $N = 100$ (yellow).

IV. CONCLUSION

We have discussed here a theoretical approach to study the eigenvalue fluctuations of sparse Wishart random matrices, in which the members of the random matrix ensemble are covariance matrices with a large amount of zero entries. Such a class of random matrices is of fundamental importance in various techniques to process large datasets in multivariate statistics. The main outcome of our work is an analytical expression [see Eq. (10)] for the $N \to \infty$ behavior of the cumulant generating function of the number of eigenvalues $\mathcal{I}_N(x)$ within the interval $(-\infty,x]$. This analytical formula is the main source of quantitative information about the fluctuations of $\mathcal{I}_N(x)$. In fact, we have presented explicit results for the mean and the variance of $\mathcal{I}_N(x)$, its rate function characterizing the probability of rare events, and the third-order cumulant of $\mathcal{I}_N(x)$.

Similarly to previous results for the adjacency matrix of random graphs [30], the rate function controlling the large-deviation probability of $\mathcal{I}_N(x)$ is asymmetric in the case of the sparse Wishart ensemble. This feature is particularly evident in the regime of high sparseness, where δ peaks due to isolated clusters in the associated random graph are abundant in the spectral density. It is reasonable to expect that the existence of Dirac δ peaks with large weights, at which a large number of eigenvalues accumulates, strongly influences the typical and rare fluctuations of $\mathcal{I}_N(x)$, being responsible for the asymmetry of the rate function. It would be interesting to further explore this issue by explicitly disentangling the contributions to the rate function coming from the discrete and continuous parts of the spectral density [37].

Finally, the present work further reinforces the exactness of the powerful method designed in [30]. To test the validity of our analytical results, we have carefully compared them with numerical diagonalization of large random matrices, and the agreement between these two independent approaches is very good. This further corroborates our theoretical method, in spite of its lack of a full mathematical rigor (see Appendix A).

ACKNOWLEDGMENTS

F.L.M. and I.P.C. thank London Mathematical Laboratory for financial support. F.L.M. acknowledges financial support from CNPq (Edital Universal 406116/2016-4). This work has been funded by the program UNAM-DGAPA-PAPIIT IA101815. The authors thank the support of DGTIC for the use of the HP Cluster Platform 3000SL, codename Miztli, under the Mega-project LANCAD-UNAM-DGTIC-333.

APPENDIX A: DERIVATION OF THE CUMULANT GENERATING FUNCTION

The number of eigenvalues $\mathcal{I}_N(x)$ within the unbounded interval $(-\infty,x]$, defined in Eq. (3), can be rewritten using an identity for the Heaviside function [30,35],

$$\mathcal{I}_N(x) = \lim_{\epsilon\to 0^+} \frac{1}{2\pi i} \sum_{\alpha=1}^{N} [\mathrm{Ln}(\lambda_\alpha + i\epsilon - x) - \mathrm{Ln}(\lambda_\alpha - i\epsilon - x)], \tag{A1}$$

where $\mathrm{Ln}(\cdots)$ corresponds to the principal branch of the complex logarithm. One of the central formulas in applying spin-glass techniques to random matrix theory is the following result for the multidimensional Fresnel integral:

$$\begin{aligned} Z(x_\epsilon) &= \int \left(\prod_{i=1}^{N} dy_i\right) \exp\left[-\frac{i}{2}\boldsymbol{y}^T \cdot (x_\epsilon \mathbb{I} - \boldsymbol{M})\boldsymbol{y}\right] \\ &= (2\pi)^{N/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{N} \mathrm{Ln}(\lambda_i - x_\epsilon) + i\frac{N\pi}{4}\right], \end{aligned} \tag{A2}$$

where $x_\epsilon = x - i\epsilon$, $\boldsymbol{y}^T = (y_1,\ldots,y_N)$, and $\mathbb{I}$ is the $N \times N$ identity matrix. In the complex plane, neither the exponential nor the logarithm are injective functions. While this is not a problem when studying the spectral density, some initial caution is warranted when considering the statistics of $\mathcal{I}_N(x)$. Being naive with these multivalued functions in the complex

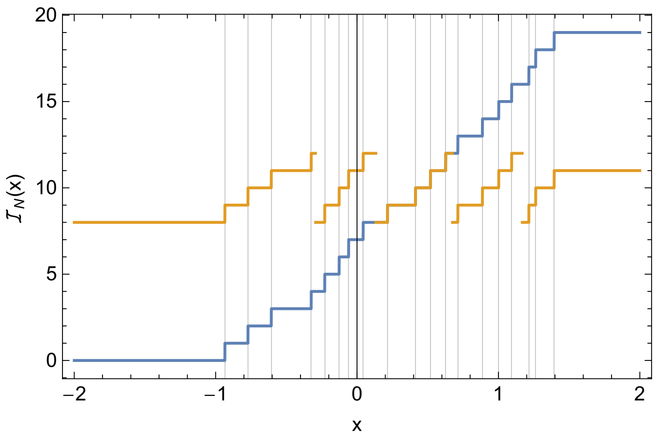


FIG. 4. The shifted index number $\mathcal{I}_N(x)$ for a single realization of an $N \times N$ random matrix. The function $\mathcal{I}_N(x)$ counts the number of eigenvalues smaller than x. Here we show a comparison between two representations of $\mathcal{I}_N(x)$: Eq. (A3) (solid orange line) and Eq. (A1) (solid blue line). The vertical gray lines denote the positions of each one of the eigenvalues $\{\lambda_\alpha\}_{\alpha=1}^{20}$.

plane, Eq. (A1) can be written as follows:

$$\mathcal{I}_N(x) = \lim_{\epsilon\to 0^+} \frac{1}{\pi i}\left[\mathrm{Ln}\overline{Z(x_\epsilon)} - \mathrm{Ln}Z(x_\epsilon) + \frac{N\pi i}{2}\right], \tag{A3}$$

with $\overline{(\cdots)}$ denoting complex conjugation.

Clearly, expression (A3) does not provide the correct result for $\mathcal{I}_N(x)$, given precisely by Eq. (A1). This situation is illustrated in Fig. 4, where we compare the outcomes of Eqs. (A3) and (A1) for a single random matrix of size $N = 20$. Thus, Eq. (A3) is unfit to count the number of eigenvalues for a single realization of $\boldsymbol{M}$. The reason is that we have naively folded in the sum appearing in Eq. (A1) as products inside the logarithms in Eq. (A3), without any regard of the multivalued character of the complex logarithm. In other words, we have assumed that the complex logarithm fulfills the same standard properties as the logarithm of a real variable.

Although Eq. (A3) is unsuitable to count the number of eigenvalues, it is the appropriate starting point to compute the ensemble average using the replica approach. It is thus a relevant question whether the above procedure leads to correct results for the statistics of the index. As we show in the paper, our method does yield correct results when comparing the final analytical expressions for the rate function and for the cumulants of $\mathcal{I}_N(x)$ with numerical diagonalization, which shows that Eq. (A3) correctly encodes the statistical properties of the index. We have some hint as to why this naive procedure works: we call it the *folding-unfolding mechanism*. After going from Eq. (A1) to Eq. (A3), one derives an effective theory using the replica method, which essentially consists of decoupling sites by coupling replicas [29]. This theory can then be unfolded by factorizing over the sites through the introduction of a suitable order parameter, restoring the correct scaling of the moments of $\mathcal{I}_N(x)$ with respect to N. The noninjective nature of the logarithm in the complex plane seems irrelevant for the statistical properties of $\mathcal{I}_N(x)$, which implies that we can be careless about the actual prescription used to represent the multidimensional Gaussian integral.

Let us now proceed to the analytical computation of the cumulant generating function. By substituting Eq. (A3) in Eq. (5), the cumulant generating function can be expressed as follows:

$$\mathcal{F}_x(y) = \frac{y}{2} - \lim_{N\to\infty}\lim_{\epsilon\to 0^+}\frac{1}{N}\mathrm{Ln}\langle [Z(x_\epsilon)]^{\frac{iy}{\pi}}[\overline{Z(x_\epsilon)}]^{-\frac{iy}{\pi}}\rangle_{\boldsymbol{M}}, \tag{A4}$$

with $\langle\cdots\rangle_{\boldsymbol{M}}$ denoting the ensemble average over the distribution of $\boldsymbol{M}$ [see Eqs. (1) and (2)]. To calculate $\mathcal{F}_x(y)$ for large N, we use the replica method in the form of the identity

$$\mathcal{F}_x(y) = \frac{y}{2} - \lim_{N\to\infty}\lim_{\epsilon\to 0^+}\lim_{n_\pm\to\pm\frac{iy}{\pi}}\frac{1}{N}\mathrm{Ln}\mathcal{Q}_N(n_\pm,x_\epsilon), \tag{A5}$$

where we have introduced the object

$$\mathcal{Q}_N(n_\pm,x_\epsilon) = \langle [Z(x_\epsilon)]^{n_+}[\overline{Z(x_\epsilon)}]^{n_-}\rangle_{\boldsymbol{M}}. \tag{A6}$$

Initially, the exponents $n_\pm$ are assumed to be positive integers. After the ensemble average in the above equation has been calculated for $N\to\infty$, we perform the replica limit $n_\pm \to \pm\frac{iy}{\pi}$ and make an analytical continuation of $n_\pm$ to the imaginary axis.

Using the Fresnel integral representation of Eq. (A2), we write $\mathcal{Q}_N(n_\pm,x_\epsilon)$ as follows:

$$\begin{aligned}\mathcal{Q}_N(n_\pm,x_\epsilon) = \int\left(\prod_{a=1}^{n_+} d\boldsymbol{z}_a \prod_{a=1}^{n_-} d\boldsymbol{y}_a\right)\\ \times \exp\left[-\frac{ix_\epsilon}{2}\sum_{a=1}^{n_+}\sum_{i=1}^{N} z_{ia}^2 + \frac{i\overline{x_\epsilon}}{2}\sum_{a=1}^{n_-}\sum_{i=1}^{N} y_{ia}^2\right]\\ \times\left\langle \exp\left[\frac{i}{2}\sum_{i,j=1}^{N} M_{ij}(\underline{z}_i\cdot\underline{z}_j - \underline{y}_i\cdot\underline{y}_j)\right]\right\rangle_M,\end{aligned} \tag{A7}$$

in which $\boldsymbol{z}_a \equiv (z_{1a},\ldots z_{Na})$, $\boldsymbol{y}_a \equiv (y_{1a},\ldots y_{Na})$, $\underline{z}_i = (z_{i1},\ldots,z_{in_+})$, and $\underline{y}_i = (y_{i1},\ldots,y_{in_-})$, with $i,j = 1,\ldots,N$. Equation (A7) holds in general for a real symmetric matrix $\boldsymbol{M}$. After computing the ensemble average in Eq. (A7) with the distribution of $\boldsymbol{M}$, defined by Eqs. (1) and (2), and introducing the order-parameter function

$$\rho(\underline{z},\underline{y}) = \frac{1}{N}\sum_{i=1}^{N}\delta(\underline{z}-\underline{z}_i)\delta(\underline{y}-\underline{y}_i), \tag{A8}$$

the quantity $\mathcal{Q}_N(n_\pm,x_\epsilon)$ can be expressed as a path integral over two functions, $\rho(\underline{z},\underline{y})$ and $k(\underline{z},\underline{y})$, defined on the replica space $(\underline{z},\underline{y}) \in \mathbb{R}^{n_+}\times\mathbb{R}^{n_-}$,

$$\mathcal{Q}_N(n_\pm,x_\epsilon) = \int D[\rho,k]e^{-N\mathcal{S}[\rho,k]}, \tag{A9}$$

where the action reads

$$\begin{aligned}\mathcal{S}[\rho,k] \equiv -i\int d\underline{z}\,d\underline{y}\,k(\underline{z},\underline{y})\rho(\underline{z},\underline{y})\\ -\mathrm{Ln}\left[\int d\underline{z}\,d\underline{y}\,\exp\left(-\frac{ix_\epsilon}{2}\underline{z}^2+\frac{i\overline{x_\epsilon}}{2}\underline{y}^2 - ik(\underline{z},\underline{y})\right)\right]\\ -\alpha\mathrm{Ln}\int D[\underline{m},\underline{L}]e^{\frac{i}{2}(\underline{m}^2-\underline{L}^2)}\mathcal{M}(\underline{m},\underline{L}|\rho)\end{aligned} \tag{A10}$$

with

$$\mathcal{M}(\underline{m},\underline{L}|\rho) \equiv \sum_{k=0}^{\infty} \frac{e^{-d}d^k}{k!} \int \left(\prod_{\ell=1}^{k} d\underline{z}_\ell d\underline{y}_\ell \rho(\underline{z}_\ell,\underline{y}_\ell) d\xi_\ell P_\xi(\xi_\ell) \right) \times \delta\left(\underline{m} - \frac{1}{\sqrt{d}} \sum_{\ell=1}^{k} \xi_\ell \underline{z}_\ell \right) \delta\left(\underline{L} - \frac{1}{\sqrt{d}} \sum_{\ell=1}^{k} \xi_\ell \underline{y}_\ell \right). \tag{A11}$$

The quantity $k(\underline{z},\underline{y})$ is the function conjugate to the order parameter $\rho(\underline{z},\underline{y})$. The next step consists in evaluating the path integral by applying the saddle-point method, which captures the leading contribution to the integral in the limit $N \to \infty$. Plugging Eq. (A9) in Eq. (A5) and taking the limit $N \to \infty$, we obtain

$$\mathcal{F}_x(y) = \frac{y}{2} + \lim_{\epsilon \to 0^+} \lim_{n_\pm \to \pm \frac{iy}{\pi}} \mathcal{S}_0[\rho,k], \tag{A12}$$

where $\mathcal{S}_0[\rho,k]$ is the action $S[\rho,k]$ evaluated at its saddle point, at which the functions $\rho(\underline{z},\underline{y})$ and $k(\underline{z},\underline{y})$ obey the stationary equations:

$$\frac{\delta S[\rho,k]}{\delta \rho(\underline{z},\underline{y})} = \frac{\delta S[\rho,k]}{\delta k(\underline{z},\underline{y})} = 0. \tag{A13}$$

After some straightforward algebra, the saddle-point equations take the following form:

$$k(\underline{z},\underline{y}) = i\alpha d \int d\xi\ P_\xi(\xi) \frac{\int D[\underline{m},\underline{L}] e^{\frac{i}{2}(\underline{m}^2 - \underline{L}^2)} \mathcal{M}(\underline{m} - \xi \underline{z}/\sqrt{d}, L - \xi \underline{y}/\sqrt{d} \big| \rho)}{\int D[\underline{m},\underline{L}] e^{\frac{i}{2}(\underline{m}^2 - \underline{L}^2)} \mathcal{M}(\underline{m},\underline{L}|\rho)}, \tag{A14}$$

$$\rho(\underline{z},\underline{y}) = \frac{\exp\left[-\frac{ix_\epsilon}{2} \underline{z}^2 + \frac{i\overline{x_\epsilon}}{2} \underline{y}^2 - ik(\underline{z},\underline{y}) \right]}{\int d\underline{z}\, d\underline{y}\ \exp\left[-\frac{ix_\epsilon}{2} \underline{x}^2 + \frac{i\overline{x_\epsilon}}{2} \underline{y}^2 - ik(\underline{z},\underline{y}) \right]}. \tag{A15}$$

The replica limit $n_\pm \to \pm \frac{iy}{\pi}$ in the saddle-point equations and in the expression (A12) will be dealt with by assuming that $\rho(\underline{z},\underline{y})$ and $k(\underline{z},\underline{y})$ are replica symmetric (RS) in the subspaces $\mathbb{R}^{n_\pm}$.

1. Replica symmetric *Ansatz*

Even though it is possible to express Eq. (A10) solely in terms of the order-parameter function $\rho(\underline{z},\underline{y})$, in the following derivation we introduce, for the sake of clarity, a replica symmetric *Ansatz* for each function $k(\underline{z},\underline{y})$, $\rho(\underline{z},\underline{y})$, and $\mathcal{M}(\underline{m},\underline{L}|\rho)$. As shown in previous works dealing with the spectral properties of random graphs [18,19,30], the parametrization of the order parameter in terms of a superposition of Gaussian functions leads to exact results. Thus, with a modest amount of foresight, we propose the *Ansätze*

$$\mathcal{M}(\underline{m},\underline{L}|\rho) = \int d\sigma\ w_\sigma(\sigma) \left[\prod_{a=1}^{n_+} \frac{\exp\left(i \frac{m_a^2}{2\sigma}\right)}{\sqrt{2\pi i \sigma}} \right] \times \left[\prod_{a=1}^{n_-} \frac{\exp\left(-i \frac{L_a^2}{2\overline{\sigma}}\right)}{\sqrt{-2\pi i \overline{\sigma}}} \right], \tag{A16}$$

$$\rho(\underline{z},\underline{y}) = \int d\Delta\, w_\rho(\Delta) \left[\prod_{a=1}^{n_+} \frac{\exp\left(i \frac{z_a^2}{2\Delta}\right)}{\sqrt{2\pi i \Delta}} \right] \times \left[\prod_{a=1}^{n_-} \frac{\exp\left(-i \frac{y_a^2}{2\overline{\Delta}}\right)}{\sqrt{-2\pi i \overline{\Delta}}} \right], \tag{A17}$$

$$k(\underline{z},\underline{y}) = i\mathcal{A} \int d\Gamma\, w_k(\Gamma) \left[\prod_{a=1}^{n_+} \sqrt{\frac{\Gamma}{2\pi i}} \exp\left(i\Gamma \frac{z_a^2}{2} \right) \right] \times \left[\prod_{a=1}^{n_-} \sqrt{\frac{\overline{\Gamma}}{-2\pi i}} \exp\left(-i\overline{\Gamma} \frac{y_a^2}{2} \right) \right]. \tag{A18}$$

It is important to keep track of the normalization factors, which is why we have written them explicitly in the above equations. We also assume that the distributions $w_\rho(\Delta)$, $w_k(\Delta)$, and $w_\sigma(\sigma)$ are normalized. Equations (A16)–(A18) remain invariant under the interchange of replica indexes within each subspace. Inserting the above Gaussian assumptions in Eqs. (A11), (A14), and (A15), and performing the replica limit $n_\pm \to \pm \frac{iy}{\pi}$, one arrives at the set of self-consistent Eqs. (12)–(14). The expression for the cumulant generating function, explicitly shown in Eq. (10), is derived by substituting Eqs. (A16)–(A18) in Eq. (A12) and then taking the limit $n_\pm \to \pm \frac{iy}{\pi}$.

The limit $\epsilon \to 0^+$ is implicit in Eq. (10) as well as in the self-consistent Eqs. (12)–(14). From a mathematical viewpoint, this limit corresponds to recovering a Dirac δ distribution from a Cauchy distribution [18,19]. However, as we will see below, choosing a small value of ϵ is enough to obtain excellent numerical results, and the actual limit does not need to be taken.

APPENDIX B: WEIGHTED POPULATION DYNAMICS

To solve numerically Eqs. (12)–(14), we use the weighted population dynamics algorithm, whose main idea has been put forward in [30]. This numerical approach consists in representing each density $w_\rho(\Delta)$, $w_k(\Gamma)$, and $w_\sigma(\sigma)$ by a large collection or population containing $\mathcal{L}$ random variables, which are updated according to the algorithm explained below. After a sufficient number of updating steps, the empirical distribution of each population converges to a fixed-point distribution that solves its corresponding self-consistent equation. The calculation of averages involving $w_\rho(\Delta)$, $w_k(\Gamma)$, and $w_\sigma(\sigma)$ is performed by taking the arithmetic mean with the corresponding population of random variables.

Thus, by choosing a large value of $\mathcal{L}$, the unknown distributions of Eqs. (12)–(14) are parametrized as follows:

$$w_\rho(\Delta) \leftrightarrow \{\Delta^{(\alpha)}\}_{\alpha=1}^{\mathcal{L}},$$
$$w_k(\Gamma) \leftrightarrow \{\Gamma^{(\alpha)}\}_{\alpha=1}^{\mathcal{L}},$$
$$w_\sigma(\sigma) \leftrightarrow \{\sigma^{(\alpha)}\}_{\alpha=1}^{\mathcal{L}}.$$

Once one has set up initial values for the three populations $\{\Delta^{(\alpha)},\Gamma^{(\alpha)},\sigma^{(\alpha)}\}_{\alpha=1}^{\mathcal{L}}$, one performs the following steps at a single iteration of the algorithm:

(i) Estimate the constant $\mathcal{A}$, defined in Eq. (16), as follows:

$$\frac{\alpha d}{\mathcal{A}} \simeq \frac{1}{\mathcal{L}}\sum_{\alpha=1}^{\mathcal{L}} \exp\left[-\frac{iy}{2\pi}\mathrm{Ln}\left(\frac{1+\sigma^{(\alpha)}}{1+\overline{\sigma^{(\alpha)}}}\right)\right].$$

(ii) Draw a random number $\ell \sim \text{Poisson}(\mathcal{A})$ and select the variables $\{\Gamma^{(\alpha_s)}\}_{s=1}^{\ell}$ uniformly and randomly from the population $\{\Gamma^{(\alpha)}\}_{\alpha=1}^{\mathcal{L}}$.

(iii) Using the ℓ variables chosen in the previous step, calculate the following quantities:

$$I = \left\lfloor e^{\frac{iy}{2\pi}W_\epsilon(\Gamma_1,\ldots,\Gamma_\ell)}\right\rfloor,$$
$$R = e^{\frac{iy}{2\pi}W_\epsilon(\Gamma_1,\ldots,\Gamma_\ell)} - I, \tag{B1}$$

where $\lfloor \ldots \rfloor$ is the floor function and W_ϵ is defined in Eq. (11).

(iv) Pick up an element $\Delta^{(\alpha_0)}$ uniformly and randomly from the population $\{\Delta^{(\alpha)}\}_{\alpha=1}^{\mathcal{L}}$ and update its value as

$$\Delta^{(\alpha_0)} \to \Delta^{(\alpha_0)} = \frac{1}{\sum_{s=1}^{\ell}\Gamma^{(\alpha_s)} - x_\epsilon},$$

with rate R. Then increase the population to size $\mathcal{L}+I$ by adding I extra values of $\Delta^{(\alpha_0)}$ to the original population. Finally, filter the new population back to its original size by choosing randomly $\mathcal{L}$ elements from the $\mathcal{L}+I$ available.

(v) Draw a random number $k \sim \text{Poisson}(d)$ and select $\{\Delta^{(\alpha_\ell)}\}_{\ell=1}^{k}$ uniformly and randomly from the population $\{\Delta^{(\alpha)}\}_{\alpha=1}^{\mathcal{L}}$. Besides that, draw independently k random numbers $\{\xi_\ell\}_{\ell=1}^{k}$ from the distribution $P_\xi(\xi)$.

(vi) Pick up an element $\sigma^{(\alpha_0)}$ uniformly and randomly from $\{\sigma^{(\alpha)}\}_{\alpha=1}^{\mathcal{L}}$ and update its value according to

$$\sigma^{(\alpha_0)} \to \sigma^{(\alpha_0)} = \frac{1}{d}\sum_{\ell=1}^{k}\xi_\ell^2\Delta^{(\alpha_\ell)}.$$

(vii) Draw a single random number ξ from $P_\xi(\xi)$ and choose a pair of elements $\sigma^{(\alpha_1)}$ and $\Gamma^{(\alpha_0)}$ uniformly and randomly from $\{\sigma^{(\alpha)}\}_{\alpha=1}^{\mathcal{L}}$ and $\{\Gamma^{(\alpha)}\}_{\alpha=1}^{\mathcal{L}}$, respectively. Update $\Gamma^{(\alpha_0)}$ as follows:

$$\Gamma^{(\alpha_0)} \to \Gamma^{(\alpha_0)} = \frac{\xi^2}{d(1+\sigma^{(\alpha_1)})}.$$

(viii) Go back to step (i) and repeat steps (i)–(vii) until the empirical distributions of $\{\Delta^{(\alpha)}\}_{\alpha=1}^{\mathcal{L}}$, $\{\Gamma^{(\alpha)}\}_{\alpha=1}^{\mathcal{L}}$, and $\{\sigma^{(\alpha)}\}_{\alpha=1}^{\mathcal{L}}$ attain stationary profiles. We use the standard convention that a single Monte Carlo step consists in repeating $\mathcal{L}$ times the steps (i)–(vii).

For each choice of parameters α, x, d, and y, we fix $\epsilon = 10^{-8}$ and $\mathcal{L}$ between 10^6 and 10^7. Such a value of ϵ is sufficiently small such that the limit $\epsilon \to 0^+$ is attained. We have usually run the algorithm between 300 and 600 Monte Carlo steps, which is more than sufficient to reach convergence. In some cases, in order to further improve the accuracy of our estimates, we have performed averages over independent runs of the algorithm.

One of our aims consists in deriving results for the rate function $\Psi_x(k)$. However, from the expression (9), we are required to find the value of y such that, for a fixed k, the following equation is fulfilled:

$$k = \kappa_1(y) \equiv \frac{\partial \mathcal{F}_x(y)}{\partial y}. \tag{B2}$$

Here $\kappa_1(y)$ corresponds to the weighted first cumulant, which can be expressed in terms of averages of the random variables $\{\mathcal{I}_\ell\}_{\ell=1}^{3}$, defined in Eqs. (17)–(19). Although it is certainly possible to find numerically y obeying Eq. (B2), it is more efficient to evaluate the rate function $\Psi_x(k)$ parametrically in y: for a given value of y, we determine the corresponding value

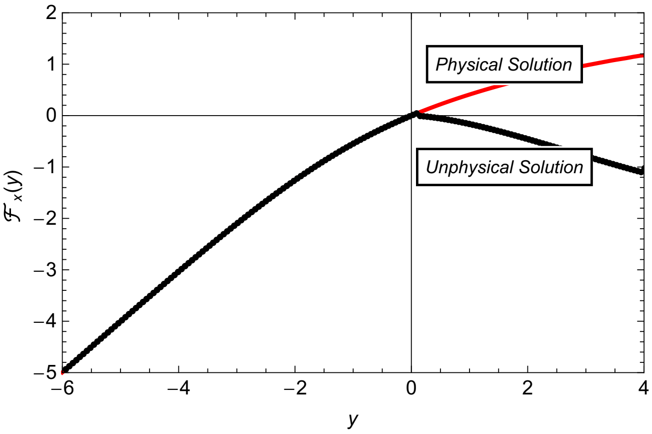


FIG. 5. Population dynamics results for $\mathcal{F}_x(y)$ as a function of y in the case of $x = 1.01$, $\epsilon = 10^{-8}$, $d = 1$, and $\alpha = 2$. This result has been obtained using $\mathcal{L} = 10^6$ and 300 Monte Carlo steps. For $y > 0$ there are two possible values of $\mathcal{F}_x(y)$, with the physical branch corresponding to the red curve.

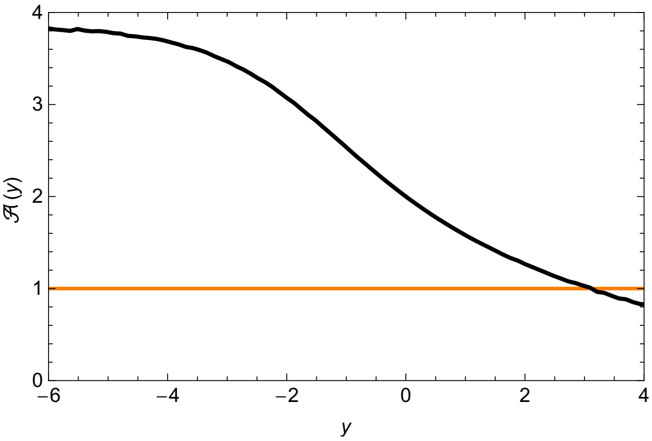


FIG. 6. Population dynamics results for $\mathcal{A}(y)$, defined by Eq. (16), as a function of y in the case of $x = 1.01$, $\epsilon = 10^{-8}$, $d = 1$, and $\alpha = 2$. This result has been obtained using $\mathcal{L} = 10^6$ and 300 Monte Carlo steps.

of $k = \kappa_1(y)$ using Eq. (B2). The rate function for such a value of k is simply given by $\Psi_x(k) = \mathcal{F}_x(y) - ky$.

When applying the weighted population dynamics algorithm, we noted that, for positive values of y, the cumulant generating function $\mathcal{F}_x(y)$ has two extrema as a functional of the distributions, that is, we find two distinct fixed-point solutions of Eqs. (12)–(14) depending on the choice of the initial distributions. We illustrate this feature in Fig. 5, where we show both the physical and the unphysical behavior of $\mathcal{F}_x(y)$ as a function of y. The physical branch is obtained by choosing the fixed-point solution of Eqs. (12)–(14) that yields $\frac{\partial \mathcal{F}_x(y)}{\partial y} \geqslant 0$. We have found that the initial conditions to obtain the physical solution are such that $\mathrm{Im}(\Gamma) = \mathrm{Im}(\Delta) = \mathrm{Im}(\sigma) = 0$.

To conclude this appendix, let us briefly comment on a limitation of our numerical method. The factor $\mathcal{A}$, given by Eq. (16), plays the role of a rescaled average degree αd and it is a function of y. In Fig. 6 we present population dynamics results for $\mathcal{A}$ as a function of y, and, as we can see, there is a range of y for which $\mathcal{A}$ is smaller than the percolation threshold. In this situation, we have found that population dynamics does not provide reliable numerical estimates for the densities, and we had to disregard the corresponding results.

[1] J. Fan, F. Han, and H. Liu, Natl. Sci. Rev. **1**, 293 (2014).

[2] I. T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics (Springer, New York, 2010).

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, Springer Texts in Statistics (Springer New York, 2014).

[4] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, Science **308**, 523 (2005).

[5] A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane, Proc. Natl. Acad. Sci. (USA) **97**, 12182 (2000).

[6] Y. Deng, Y.-H. Jiang, Y. Yang, Z. He, F. Luo, and J. Zhou, BMC Bioinform. **13**, 113 (2012).

[7] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau, PLOS Comput. Biol. **11**, e1004226 (2015).

[8] J. Fan, Y. Liao, and H. Liu, Econometrics J. **19**, C1 (2016).

[9] J. Wishart, Biometrika **20A**, 32 (1928).

[10] A. Gupta and D. Nagar, *Matrix Variate Distributions*, Monographs and Surveys in Pure and Applied Mathematics (Taylor & Francis, Boca Raton, FL, 1999).

[11] F. J. Dyson, J. Math. Phys. **3**, 140 (1962).

[12] F. J. Dyson, J. Math. Phys. **3**, 157 (1962).

[13] F. J. Dyson, J. Math. Phys. **3**, 166 (1962).

[14] P. Vivo, S. N. Majumdar, and O. Bohigas, J. Phys. A **40**, 4317 (2007).

[15] E. Katzav and I. Pérez Castillo, Phys. Rev. E **82**, 040104 (2010).

[16] S. N. Majumdar and P. Vivo, Phys. Rev. Lett. **108**, 200601 (2012).

[17] S. N. Majumdar and M. Vergassola, Phys. Rev. Lett. **102**, 060601 (2009).

[18] T. Rogers, I. P. Castillo, R. Kühn, and K. Takeda, Phys. Rev. E **78**, 031116 (2008).

[19] R. Kühn, J. Phys. A **41**, 295002 (2008).

[20] T. Rogers and I. P. Castillo, Phys. Rev. E **79**, 012101 (2009).

[21] T. Rogers, C. P. Vicente, K. Takeda, and I. P. Castillo, J. Phys. A **43**, 195002 (2010).

[22] F. L. Metz, I. Neri, and D. Bollé, Phys. Rev. E **82**, 031135 (2010).

[23] F. L. Metz, I. Neri, and D. Bollé, Phys. Rev. E **84**, 055101 (2011).

[24] I. Neri and F. L. Metz, Phys. Rev. Lett. **109**, 030602 (2012).

[25] D. Bollé, F. L. Metz and I. Neri, in *Spectral Analysis, Differential Equations and Mathematical Physics: A Festschrift in Honor of Fritz Gesztesy's 60th Birthday*, edited by H. Holden *et al.*, Proceedings of Symposia in Pure Mathematics (AMS, Providence, Rhode Island, 2013), Vol. 87, pp. 35–58.

[26] F. L. Metz, G. Parisi, and L. Leuzzi, Phys. Rev. E **90**, 052109 (2014).

[27] I. Neri and F. L. Metz, Phys. Rev. Lett. **117**, 224101 (2016).

[28] These techniques were developed in [30,30,38]. See also [39] for a similar technique introduced in a different context.

[29] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond*, Lecture Notes in Physics Series (World Scientific, Singapore, 1987).

[30] F. L. Metz and I. P. Castillo, Phys. Rev. Lett. **117**, 104101 (2016).

[31] B. Bollobás, *Random Graphs*, 2nd ed. (Cambridge University Press, Cambridge, 2001), Cambridge Books Online.

[32] M. Mehta, *Random Matrices*, Pure and Applied Mathematics (Elsevier Science, Amsterdam, 2004).

[33] H. Touchette, Phys. Rep. **478**, 1 (2009).

[34] M. Bauer and O. Golinelli, J. Stat. Phys. **103**, 301 (2001).

[35] F. L. Metz and D. A. Stariolo, Phys. Rev. E **92**, 042153 (2015).

[36] B. Fornberg, Math. Comput. **51**, 699 (1988).

[37] R. Kühn, Phys. Rev. E **93**, 042110 (2016).

[38] F. L. Metz and I. P. Castillo, Phys. Rev. B **96**, 064202 (2017).

[39] A. Coolen, J. Phys. Conf. Ser. **699**, 012022 (2016).