Ερι drαft

Eigenvalue and Eigenvector Statistics in Time Series Analysis

PAOLO BARUCCA[1], MARIO KIEBURG[2] and ALEXANDER OSSIPOV[3]

[1] *Department of Computer Science, University College London, London WC1E 6EA, United Kingdom*
[2] *Faculty of Physics, Bielefeld University, P.O. Box 100131, D-33501 Bielefeld, Germany*
[3] *School of Mathematical Sciences, University of Nottingham, Nottingham NG7 2RD, United Kingdom*

Abstract –The study of correlated time series is ubiquitous in statistical analysis, and the matrix decomposition of the cross-correlations between time series is a universal tool to extract the principal patterns of behavior in a wide range of complex systems. Despite this fact, no general result is known for the statistics of eigenvectors of the cross-correlations of correlated time series. Here we use supersymmetric theory to provide novel analytical results that will serve as a benchmark for the study of correlated signals for a vast community of researchers.

The theory of complex systems ultimately deals with the identification of patterns of simple behaviours accounting for the emergence of universal statistics in the time series measured in a vast range of disciplines, including condensed matter physics, medicine, finance, signal transmission, biology, and more recently computational social sciences [1]. A time series is a series of values scanned over time of a given observable of a system [2] such as the sea level [3], the temperature of a lake [4], the neuron activity in electroencephalography (EEG) [5,6], the response in a unit of volume of a magnetic resonance imaging experiment [7], the gross domestic product of a country [8], the price or return of a stock [9,10], the volume of an order in the market [11,12], the infected individuals in a region affected by an epidemics [13], and the online activity of a user [14].

The basic analysis that is ubiquitously performed when dealing with multiple time series are covariance and correlation analysis, especially with the aim of identifying the main factors accounting for time variability and parsimoniously representing the state space of the system, through denoising and dimensionality reduction. The generality of this statistical approach constitutes the basis for Principal Component Analysis (PCA) [15,16], clustering analysis, and many other data mining algorithms [17]. In these techniques one distinguishes between eigenvalue and eigenvector statistics and both of them carry important information as we know, for example, from the theory of quantum disordered systems. Therefore it is even more surprising that only few results are available for the joint statistics of eigenvalues and eigenvectors, when dealing with the covariance and correlation matrices of noisy time series.

The spectral density of the eigenvalues is up to now the major quantity where the theory provides robust and general results [18–21]. For instance, the Marčenko–Pastur distribution (MPD) [22] usually serves as a blueprint for describing the influence of white noise in the time series on the spectral density. Any deviation from the MPD, for instance outliers, can be considered as system specific information so that the MPD serves as a filter. The remaining eigenvalues should then go into PCA, since they obviously carry system relevant information. However, some eigenvalues encoding those relevant information might be obscured by the bulk of the spectrum described by the MPD. Then PCA may remove relevant data that should be taken into account. To distinguish those system specific eigenvalues from the eigenvalues of the MPD one needs to take into consideration the eigenvector statistics. An important step in this direction is made in the present letter. We derive an analytical formula for the first moment of a fixed component of the eigenvectors conditioned to a chosen eigenvalue. This quantity is known as the inverse participation number (IPN) in disordered systems studies and has been applied there since the early 80s [23]. Moreover, we state a conjecture on their general moments and distributions for a correlation matrix of noisy time series. Our results provide insights and pave the way for a much more

informative spectral decomposition in time series analysis, allowing not only to focus on the spectral density but also on the individual contribution of each component to the spectrum, leading to a much deeper understanding of a system's dynamics.

Random Matrix Model. – Specifically, we study the statistics of the eigenvectors and the eigenvalues of the matrix $C = WW^T$, with $W \in \mathbb{R}^{p \times n}$ representing p time series of length n or, in the case of PCA, p descriptors with n variants, and W^T being the transpose of W. Thus C can be interpreted as the correlation matrix between the time series aggregated in W or the correlation between the descriptors respectively. The real rectangular matrix W in our model is composed of four matrices [22]

$$W = \sqrt{C_L}(W_0 + W_1)\sqrt{C_R}, \qquad (1)$$

where $W_0 \in \mathbb{R}^{p \times n}$ is a deterministic real matrix and $W_1 \in \mathbb{R}^{p \times n}$ is a Gaussian random matrix distributed by

$$P(W_1) = (2\pi\sigma^2)^{-pn/2} \exp\left[-\frac{1}{2\sigma^2}\operatorname{tr} W_1 W_1^T\right], \sigma > 0. \quad (2)$$

where $\sigma > 0$ is the variance. The two real symmetric matrices $C_L = C_L^T \in \mathbb{R}^{p \times p}$ and $C_R = C_R^T \in \mathbb{R}^{n \times n}$ are positive definite and represent a spatio-temporal correlation between the various time series. Here, the matrix C_L can be identified with a time correlation and the matrix C_R with the spatial correlations, i.e., we have $\{C_L\}_{jj} = \{C_R\}_{kk} = 1$. Additionally, at difference with many common models, we include an offset W_0. This term captures the fact that the noise W_1 cannot only enter multiplicative in time series but also additive. Although, one would usually restrict the discussion to either additive or multiplicative noise, we would like to discuss the more general case that has both situations as its limit. To keep the interpretation of $C = WW^T$ and $W^T W$ as a correlation matrix the matrix W_0 has to satisfy

$$\sum_{j'=1}^{p} \{\sqrt{C_L}W_0\}_{j',k} = \sum_{k'=1}^{n} \{W_0\sqrt{C_R}\}_{j,k'} = 0 \qquad (3)$$

and

$$\begin{cases} \frac{1}{n}\sqrt{C_L}W_0 C_R W_0^T \sqrt{C_L} + \sigma^2 C_L \end{cases}_{jj} = 1, \\ \begin{cases} \frac{1}{p}\sqrt{C_R}W_0^T C_L W_0 \sqrt{C_R} + \sigma^2 C_R \end{cases}_{kk} = 1 \qquad (4)$$

for all $k = 1, \ldots, n$ and $j = 1, \ldots, p$. These conditions mean that the first moment of W has to vanish and the covariance matrices of the columns and of the rows of W are unity, meaning that W as model time series matrix is properly normalised. Otherwise, if we only impose condition (3) the model can represent covariance matrices. Conditions (3) and (4) can, however, be relaxed if one considers applications motivated by physics, rather than

a time series, such as the Dirac operator in QCD since then W ceases to model times series nor $C = WW^T$ has to be a correlation matrix. In summary, W is a non-centred and doubly correlated Gaussian random matrix. This form allows the model to capture in detail the case of factor models ubiquitous in statistics and econometrics [24].

Though our model is quite general, it is still not the most general Gaussian random matrix model. We assume that the spatio-temporal correlations of the multivariate time series factorize in the two matrices, C_L and C_R. Therefore time-dependent spatial correlations, like the two epoch model [25], are not considered here.

The random matrix model defined above can be also considered as a simple deformation of the standard real Wishart ensemble of random matrices, in which the orthogonal invariance, i.e. the invariance of the probability distribution function of W under the orthogonal transformation $W \to OWO^T$, is broken in several ways. Such non-invariant deformations of the standard random matrix ensembles were introduced and studied in different contexts including wireless communication [26], vibration analysis [27], signal processing [28] and neural networks [29]. As the eigenvectors in such ensembles are characterised by some non-trivial statistics and contain an additional information compared to the eigenvalues, there is a growing interest to their statistical properties. While there are some recent results about the statistics of the eigenvectors in the deformed Gaussian Orthogonal and Unitary ensembles [30–36], we are not aware of similar results for the Wishart ensemble except for Ref. [34], in which the ergodicity of the eigenvectors was proven for the special case $C_L = \mathbb{1}_p$, $C_R = \mathbb{1}_n$.

In the following, we will not simply focus on the computation of the spectral density of the eigenvalues, analysed in [37,38] with the same supersymmetric (SUSY) approach as in the present work, but also calculate a detailed eigenvector statistics of the matrix $C = U\Lambda U^T$, whose eigenvalues are represented by the diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ and the eigenvectors are represented by the columns of the matrix $U \in \mathrm{O}(p)$, where U_{ab} are the matrix entries of U. Due to the absence of the orthogonal invariance in our model, different eigenvector components are characterised by different distribution functions and therefore should be studied separately. Their distribution functions generally depend also on the position of the corresponding eigenvalue in the spectrum. This provides a motivation for studying the statistics of a particular eigenvector component at a particular point in the spectrum. The full information about such statistics is contained in the conditional density [39]

$$\mathcal{I}_b(\mu|\lambda) = \frac{1}{p\rho(\lambda)}\sum_{a=1}^{p} \langle\delta(\mu - |U_{ab}|^2)\delta(\lambda - \lambda_a)\rangle, \quad (5)$$

where the index $b = 1, \ldots, p$ refers to a particular eigenvector component, μ is a value of $|U_{ab}|^2$, λ is a given point

in the spectrum and

$$\rho(\lambda) = \frac{1}{p} \sum_{a=1}^{p} \langle \delta(\lambda - \lambda_a) \rangle \qquad (6)$$

is the mean density of the eigenvalues and $\langle . \rangle$ stands for the ensemble average over the distribution of W_1. In the case of a factorisation of the eigenvector and eigenvalue statistics, as in the Wishart ensemble, one finds the chi-square distribution

$$\mathcal{I}_b^{(\mathrm{Haar})}(\mu|\lambda) = \sqrt{\frac{p}{2\pi\mu}} \exp\left[-\frac{p\mu}{2}\right], \qquad (7)$$

which is equal to the Porter–Thomas distribution [40] of a uniformly distributed real eigenvector on a unit sphere. The distribution (7) is independent of the component b and the eigenvalue λ due to the Haar distributed eigenvectors. This simplification cannot be expected to hold in our non-trivial model as well as in a realistic situation. The computation of (5) or its arbitrary moments

$$I_{q,b}(\lambda) = \langle \mu^q \rangle = \frac{1}{p\rho(\lambda)} \sum_{a=1}^{p} \langle |U_{ab}|^{2q} \delta(\lambda_a - \lambda) \rangle, \qquad (8)$$

where q is a positive integer, is technically a very challenging problem. In the present letter, we focus on the analytical derivation of the first moment $I_{1,b}(\lambda)$ and make a conjecture about an arbitrary moment $I_{q>1,b}(\lambda)$ and $\mathcal{I}_b(\mu|\lambda)$ in the conclusions.

The moments of the eigenvectors are also a standard tool to characterise properties of complex quantum systems and are used to distinguish different phases in condensed matter physics [23, 41]. Hence, we expect that it may give valuable insights for time series as well.

Before we start with the analytical calculation of $I_{1,b}$, we want to point out that the eigenvector components U_{ab} are basis dependent. Thus the conditional distribution $\mathcal{I}_b(\mu|\lambda)$ strongly depends on the reference frame. In this work such a frame is chosen as the eigenbasis of C_L, allowing us to investigate the broadening of the eigenvectors due to the white noise W_1 and its strength σ. Another natural and valuable reference frame could be the eigenbasis of $\sqrt{C_L} W_0 C_R W_0^T \sqrt{C_L}$ which we do not consider here for simplicity.

Eigenvector Statistics with SUSY. – The first moment of the eigenvectors, see (8) for $q = 1$, can be computed by taking the imaginary part and the limit of a regularization $\epsilon \to 0$ of the quantity

$$I'_{1,b}(\lambda) = -\frac{1}{\sqrt{\lambda_+}} \left\langle \left\{ \begin{bmatrix} \sqrt{\lambda_+}\mathbb{1}_p & W \\ W^T & \sqrt{\lambda_+}\mathbb{1}_n \end{bmatrix}^{-1} \right\}_{bb} \right\rangle, \qquad (9)$$

where $\lambda_+ = \lambda + i\varepsilon$. Defining the $(p+n)$-dimensional unit vector e_b with unity at the position b and zero otherwise, this quantity can be generated by differentiating

$$Z_b(\lambda) = \left\langle \exp\left[i\alpha^2 e_b^T \begin{bmatrix} \sqrt{\lambda_+}\mathbb{1}_p & W \\ W^T & \sqrt{\lambda_+}\mathbb{1}_n \end{bmatrix}^{-1} e_b \right] \right\rangle, \qquad (10)$$

with respect to the auxiliary parameter α^2, at $\alpha = 0$. α is chosen to be real to guarantee convergence later on. We follow the standard steps of the SUSY method [37, 38]. For an introduction to superanalysis and superalgebra we refer to [42]. Here, we will only sketch our computations and refer to our Supplementary Material [43] for further details. First, we represent the generating function $Z_b(\lambda)$ by a supersymmetric Gaussian integral by using the idea that $\int \exp[-\mathrm{Str}\, V\Xi V^\dagger + i\mathrm{Str}\,(VK^\dagger + KV^\dagger)]d[V] \propto \exp[-\mathrm{Str}\, K\Xi^{-1}K]$ for some specific (super-)matrices V, K and Ξ. In our case, Ξ will be the ordinary matrix in (10) and K will be the embedding of the vector e_b in terms of a $(p+n) \times (2|2)$ rectangular supermatrix, see [43] for the details. In this way we obtain a non-centred Gaussian integral in W_1 that can be carried out. Because the resulting integral involves an exponential with a quartic term in the auxiliary supermatrix V we need to apply a hyperbolic Hubbard-Stratonovich transformation [44]. The Hubbard-Stratonovich transformation reformulates the integral over V again to a Gaussian form that can be integrated out. In this way, we derive the following supersymmetric representation for $I_{1,b}(\lambda)$,

$$I_{1,b}(\lambda) = \frac{\int d[T] \exp\left[F(T_+, T_-)\right] G_{1b,1b} \sqrt{\mathrm{Sdet}\, G}}{\sqrt{\lambda_+} \int d[T] \exp\left[F(T_+, T_-)\right]}, \qquad (11)$$

where $F(T_+, T_-) = -\frac{1}{2\sigma^2}\mathrm{Str}\,(T_+^2 + T_-^2) - \mathrm{Str}\, T_+ L$, $L = \mathrm{diag}\,(-\mathbb{1}_2; \mathbb{1}_2)$. The $(2|2) \times (2|2)$ supermatrices $T_\pm$ are the auxiliary supermatrices from the Hubbard-Stratonovich transformation. They are symmetric in the boson-boson block and self-dual in the fermion-fermion block and their eigenvalues run along complex contours that are detailed in the Supplemental Material [43]. The supersymmetric Green function G has the form

$$G = \begin{bmatrix} G^{(11)} & \sqrt{C_L}W_0\sqrt{C_R} \otimes LJ \\ \sqrt{C_R}W_0^T\sqrt{C_L} \otimes LJ & G^{(22)} \end{bmatrix}^{-1},$$

$$G^{(11)} = \sqrt{\lambda_+}\mathbb{1}_p \otimes LJ - C_L \otimes (T_+ - iT_- + \sigma^2 L)LJ,$$

$$G^{(22)} = \sqrt{\lambda_+}\mathbb{1}_n \otimes LJ - C_R \otimes (T_+ + iT_- + \sigma^2 L)LJ,$$

$$(12)$$

with $J = \mathrm{diag}\,(\mathbb{1}_2; \tau_2)$, where τ_2 is the second Pauli matrix. The representation (11) is exact, but rather involved and technical. An expression for the mean level density can be obtained by summing over $b = 1, \ldots, p$ and should be compared with the corresponding result in [37, 38, 45]. The above expression simplifies a lot in the limit $n, p \to \infty$, which is considered next. As many other results in the random matrix theory our results can be interpreted as leading asymptotic results as $n, p \to \infty$, i.e., n and p are large, but finite as it is often the case in the time of "big data".

Macroscopic level density and limiting statistics. – In most applications, one is interested in the limit $n, p \to \infty$ because the data set of the time series is extremely large. In this limit the integral in Eq. (11)

can be evaluated using the saddle-point approximation. To derive the saddle-point equation, it is convenient to introduce the supermatrices $S = T_+ - iT_- + \sigma^2 L$ and $R = T_+ + iT_- + \sigma^2 L$, which can be considered as independent. Again we can only sketch the derivation and postpone the details to our Supplementary Material [43]. After we have independently differentiated with respect to these two supermatrices, we obtain two coupled equations. The saddle-point solution contributing most to the integral is given by the diagonal matrices $S_0 = s_0 \mathbf{1}_{2|2}$ and $R_0 = r_0 \mathbf{1}_{2|2}$ with the complex parameters s_0 and r_0 that satisfy the coupled equations [43]

$$\frac{r_0}{\sigma^2} = \text{tr} \left[A_L - W_0 A_R^{-1} W_0^T \right]^{-1},$$
$$\frac{s_0}{\sigma^2} = \text{tr} \left[A_R - W_0^T A_L^{-1} W_0 \right]^{-1}, \text{ with} \quad (13)$$
$$A_L = \sqrt{\lambda} C_L^{-1} - s_0 \mathbf{1}_p \text{ and } A_R = \sqrt{\lambda} C_R^{-1} - r_0 \mathbf{1}_n.$$

The mean level density can be expressed in the saddle solution, too, since the sum over b of $I'_{1,b}$, see (11), yields the Green function. It is up to a normalisation constant given by

$$\rho(\lambda) \propto \text{Im} \left[\text{tr} \left(Q^{-1} \text{diag} \left(\mathbf{1}_p, 0 \right) \right) \right],$$
$$Q = \begin{bmatrix} \sqrt{\lambda} \mathbf{1}_p - s_0 C_L & \sqrt{C_L} W_0 \sqrt{C_R} \\ \sqrt{C_R} W_0^T \sqrt{C_L} & \sqrt{\lambda} \mathbf{1}_n - r_0 C_R \end{bmatrix}, \quad (14)$$

where we assume $p \leq n$ without loss of generality. The case $p > n$ only yields an additional Dirac delta function at the origin. The term $\text{tr} \left(Q^{-1} \text{diag} \left(\mathbf{1}_p, 0 \right) \right)$ is essentially the $\sum_b G_{1b,1b}$ at the saddle point.

Formula (14) reduces to the MPD [22] in the case of the Wishart ensemble, i.e., $C_L = \mathbf{1}_p$, $C_R = \mathbf{1}_n$ and $W_0 = 0$. We illustrate the result for $\rho(\lambda)$ in Fig. 1 for the one-factor model, which is described in the next subsection.

The result for $I_{1,b}(\lambda)$ can be expressed in terms of the same matrix Q and reads

$$I_{1,b}(\lambda) = \frac{\text{Im} \left[\text{tr} \left(Q^{-1} \text{diag} \left(\widehat{E}_b, 0 \right) \right) \right]}{\text{Im} \left[\text{tr} \left(Q^{-1} \text{diag} \left(\mathbf{1}_p, 0 \right) \right) \right]}, \quad (15)$$

which constitutes the main result of the present letter. $\widehat{E}_b$ is a diagonal matrix with unity at the b-th diagonal entry and zeros otherwise. The normalisation is fixed by the condition $\sum_{b=1}^p I_{1,b}(\lambda) = 1$. We note that for a Haar distributed vector one has $I_{1,b}^{(\text{Haar})}(\lambda) = 1/p$.

One-factor model. – To illustrate our findings we apply our general results to the one-factor model supplemented with Gaussian noise. Specifically, we set $W_0 = wx^T$, where w and x are column vectors of length p and n, respectively. The correlation matrices are chosen to be diagonal $C_L = \text{diag} \left(l_1^{-1}, \ldots, l_p^{-1} \right)$ and $C_R = \mathbf{1}_n$. The vector x represents a common factor, e.g., the market mode in financial time series analysis, and the component w_j quantifies the relative weight of the common factor on the jth time series, before normalization.
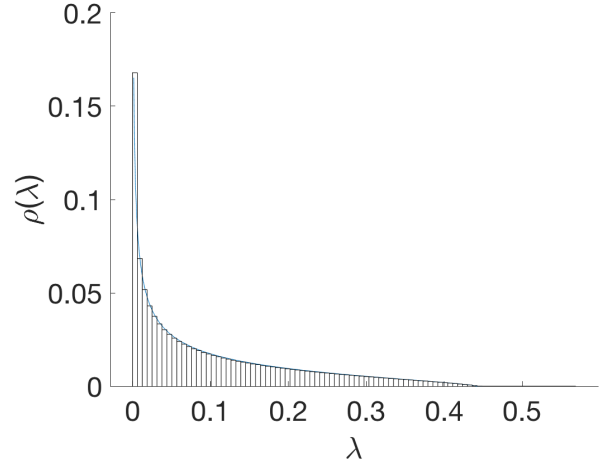

Fig. 1: Eigenvalue density for the one-factor model: analytical result (solid line, combination of Eqs. (14),(16), and (17)) and Monte-Carlo simulation (histogram, $p = 2000$, $n = 2000$ and sample size is 1000). $(C_L)_{ij} = l_i^{-1} \delta_{ij}$, $C_R = \mathbf{1}_n$, $(W_0)_{it} = w_i x_t$, $\sigma = 0.3/\sqrt{n}$, where the l_is and w_is are drawn only once from a log-normal distribution with mean 1 and variance $e^{\sigma_{L/w}^2} - 1$, with $\sigma_L = 0.3$ and $\sigma_w = 0.1$, respectively, and then kept fixed. The vector $\{x_t = A \cos(f t)\}$ is a cosine wave with frequency $f = 1/50$ and amplitude $A = 1/\sqrt{n}$.

We plug the matrices of the one-factor model into the saddle-point equation (13) and simplify the resulting expression via the Sherman-Morrison identity for the inverse matrices [46], i.e. $(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} uv^T A^{-1}}{1 + v^T A^{-1} u}$. This leads to the coupled equations

$$\frac{r_0}{\sigma^2} = \sum_j \frac{1}{\sqrt{\lambda} l_j - s_0}$$
$$+ \frac{\sum_t x_t^2 \sum_j w_j^2/(\sqrt{\lambda} l_j - s_0)^2}{\sqrt{\lambda} - r_0 - \sum_t x_t^2 \sum_j w_j^2/(\sqrt{\lambda} l_j - s_0)},$$
$$\frac{s_0}{\sigma^2} = \frac{1}{\sqrt{\lambda} - r_0} \left(n \right.$$
$$\left. + \frac{\sum_t x_t^2 \sum_j w_j^2/(\sqrt{\lambda} l_j - s_0)}{\sqrt{\lambda} - r_0 - \sum_t x_t^2 \sum_j w_j^2/(\sqrt{\lambda} l_j - s_0)} \right). \quad (16)$$

Solving these saddle-point equations we can derive the spectral density and the moments of the eigenvectors, simply by plugging the following matrix elements in Eqs. (14)-(15)

$$(Q^{-1})_{ii} = \frac{l_i}{\sqrt{\lambda} l_i - s_0} + \frac{1}{(\sqrt{\lambda} l_i - s_0)^2}$$
$$\times \frac{l_i w_i^2 \sum_t x_t^2}{\sqrt{\lambda} - r_0 - \sum_t x_t^2 \sum_j w_j^2/(\sqrt{\lambda} l_j - s_0)}. \quad (17)$$

We illustrate these results in Figs. 1 and 2, where we also compare them with Monte-Carlo simulations. The deviations from a uniformly distributed eigenvector, whose first
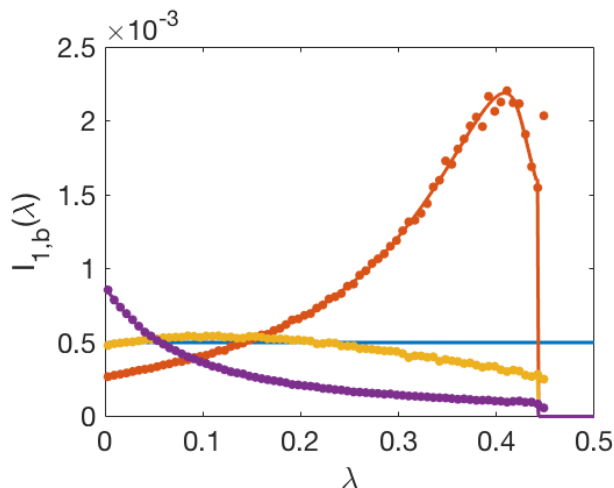
Fig. 2: Moments $I_{1,b}(\lambda)$ of the eigenvectors (15) for the one-factor model corresponding to different components: analytical result (solid line) and Monte Carlo simulation (points), for the same parameters as in Fig. 1. Moments of the components corresponding to the 50th (orange, $(1/l)_{50} = 1.88$), 1000th (yellow, $(1/l)_{1000} = 1.044$), and 1950-th (purple, $(1/l)_{1950} = 0.57$) values of $1/l$. The blue line corresponds to $I_{1,b}^{(\text{Haar})}(\lambda) = 1/p$.

moment is the constant $I_{1,b}^{(\text{Haar})} = 1/p$, can be readily seen for some components of the eigenvectors. They indicate that the corresponding eigenvalues still carry a lot of information on the matrix C_L, although these eigenvalues are evidently inside the bulk of the spectrum, cf., Fig 1. This simple example demonstrates the strength of the combined statistics of eigenvalues and eigenvectors.

Conclusions. – The general result in Eq. (15) provides a powerful analytical methodology to quantify the expected value of the square of specific components in a given eigenvalue interval for a wide range of random matrices. We tested numerically these analytical results in detail for the one-factor model (see Figs. 1-2). Our general formulation allows an arbitrary number of factors to be added in the matrix W_0. Although our analytical results were derived in the limit $n, p \to \infty$, they show a very good agreement with the results of numerical simulations at finite n and p. The rate of convergence to the limiting statistics will generally depend on the input W_0, C_L, and C_R.

In the present work we derived analytically a closed result only for the first moment $I_{1,b}(\lambda)$ of a particular eigenvector component under the condition of a fixed eigenvalue. However we conjecture that all higher moments are related to the first moment as follows:

$$I_{q,b}(\lambda) = \frac{(2q)!}{2^q\, q!}[I_{1,b}(\lambda)]^q, \tag{18}$$

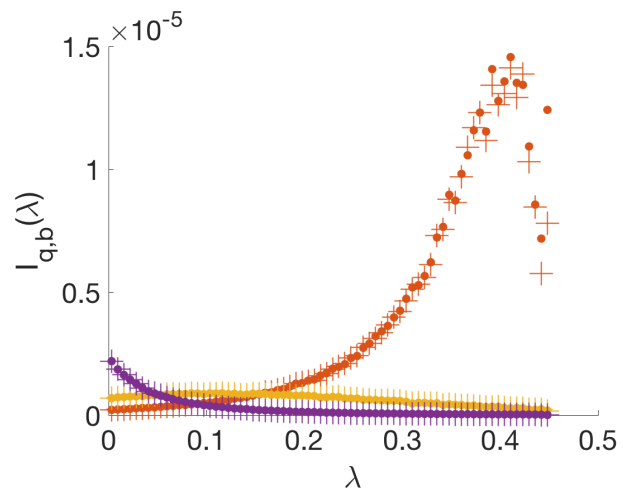which corresponds to a locally rescaled chi-square distri-

Fig. 3: Moments $I_{2,b}(\lambda)$ (dots) of the eigenvectors for the one-factor model and $3[I_{1,b}(\lambda)]^q$ (crosses), see (18) for $q = 2$, as functions of λ, for the same parameters as in Fig. 2.

bution

$$\mathcal{I}_b(\mu|\lambda) = \frac{1}{\sqrt{2\pi I_{1,b}(\lambda)\mu}} \exp\left[-\frac{\mu}{2I_{1,b}(\lambda)}\right]. \tag{19}$$

A similar result has been also found for the conditioned eigenvector statistics of the deformed Gaussian Unitary Ensemble (GUE) in [31, 32]. The only difference is the prefactor in (18), which is equal to $(2q)!/(2^q\, q!)$ in our case and given by $q!$ for the complex eigenvectors in the deformed GUE [31,32]. These numerical values result from the averaged moments of real and complex normalized vectors, respectively. We have tested this conjecture numerically for $q = 2$ and found a nice agreement, see Fig. 3. A similar observation for the distribution of the components of a fixed eigenvector over the diagonal matrix entries has been made in [47] in the field of nuclear physics.

Our analytical results are of general relevance for the spectral decomposition of time series and could lead to unprecedented understanding of the full statistics of the eigen-components in signal analysis. A strong deviation of the moment $I_{1,b}(\lambda)$ from the λ-independent constant $1/p$ hints at an eigenvector-eigenvalue pair that contains system specific information. This knowledge can improve PCA and other techniques to reduce highly dimensional data without loosing relevant information.

$*\ *\ *$

at the Simons Center for Geometry and Physics (Stony Brook University).

REFERENCES

[1] STROGATZ S. H., *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering* (CRC Press) 2018.

[2] CHATFIELD C., *Introduction to multivariate analysis* (Routledge) 2018.

[3] RAHMSTORF S., *Science*, **315** (2007) 368.

[4] SHARMA S., GRAY D. K., READ J. S., OREILLY C. M., SCHNEIDER P., QUDRAT A., GRIES C., STEFANOFF S., HAMPTON S. E., HOOK S. *et al.*, *Scientific Data*, **2** (2015) 150008.

[5] ŠEBA P., *Physical review letters*, **91** (2003) 198104.

[6] MÜLLER M., JIMÉNEZ Y. L., RUMMEL C., BAIER G., GALKA A., STEPHANI U. and MUHLE H., *Physical Review E*, **74** (2006) 041119.

[7] KWONG K. K., BELLIVEAU J. W., CHESLER D. A., GOLDBERG I. E., WEISSKOFF R. M., PONCELET B. P., KENNEDY D. N., HOPPEL B. E., COHEN M. S. and TURNER R., *Proceedings of the National Academy of Sciences*, **89** (1992) 5675.

[8] LEE C.-C., *Energy economics*, **27** (2005) 415.

[9] PLEROU V., GOPIKRISHNAN P., ROSENOW B., AMARAL L. A. N. and STANLEY H. E., *Physical review letters*, **83** (1999) 1471.

[10] BARUCCA P., *Physical Review E*, **90** (2014) 062129.

[11] BOUCHAUD J.-P., FARMER J. D. and LILLO F., *How markets slowly digest changes in supply and demand* in *Handbook of financial markets: dynamics and evolution* (Elsevier) 2009 pp. 57–160.

[12] CHIARELLA C., IORI G. and PERELLÓ J., *Journal of Economic Dynamics and Control*, **33** (2009) 525.

[13] GRENFELL B. T., BJØRNSTAD O. N. and KAPPEY J., *Nature*, **414** (2001) 716.

[14] O'CONNOR B., BALASUBRAMANYAN R., ROUTLEDGE B. R., SMITH N. A. *et al.*, *Icwsm*, **11** (2010) 1.

[15] PEARSON K., *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2** (1901) 559.

[16] JOLLIFFE I., *Principal component analysis* in *International encyclopedia of statistical science* (Springer) 2011 pp. 1094–1096.

[17] LLOYD S., *IEEE transactions on information theory*, **28** (1982) 129.

[18] LALOUX L., CIZEAU P., BOUCHAUD J.-P. and POTTERS M., *Physical review letters*, **83** (1999) 1467.

[19] LILLO F. and MANTEGNA R., *Physical Review E*, **72** (2005) 016219.

[20] ALLEZ R., BOUCHAUD J.-P., MAJUMDAR S. N. and VIVO P., *Journal of Physics A: Mathematical and Theoretical*, **46** (2012) 015001.

[21] MAJUMDAR S. N. and VIVO P., *Physical review letters*, **108** (2012) 200601.

[22] MARČENKO V. A. and PASTUR L. A., *Math. USSR-Sbornik*, **1** (1967) 457.

[23] WEGNER F., *Z. Phys. B*, **36** (1980) 209.

[24] ROSS S. A., *Journal of Economic Theory*, **13** (1976) 341.

[25] AKEMANN G., CHECINSKI T. and KIEBURG M., *J. Phys. A*, **49** (2016) 315201.

[26] COUILLET R. and DEBBAH M., *Random matrix methods for wireless communications* (Cambridge University Press) 2011.

[27] SOIZE C., *Journal of Sound and Vibration*, **263** (2003) 893.

[28] NADAKUDITI R. and EDELMAN A., *IEEE Trans. Signal Process.*, **56** (2008) 2625.

[29] AHMADIAN Y., FUMAROLA F. and MILLER K. D., *Phys. Rev. E*, **91** (2015) 012820.

[30] ALLEZ R., BUN J. and BOUCHAUD J.-P., *arXiv preprint arXiv:1412.7108*, (2014) .

[31] TRUONG K. and OSSIPOV A., *Journal of Physics A: Mathematical and Theoretical*, **49** (2016) 145005.

[32] TRUONG K. and OSSIPOV A., *Europhysics Letters*, **116** (2016) 37002.

[33] BUN J., BOUCHAUD J. P. and POTTERS M., *Physics Reports*, **666** (2017) 1.

[34] BOURGADE P. and YAU H.-T., *Communications in Mathematical Physics*, **350** (2017) 231.

[35] BENIGNI L., *preprint arXiv:1711.07103*, (2017) .

[36] TRUONG K. and OSSIPOV A., *Journal of Physics A: Mathematical and Theoretical*, **51** (2018) 065001.

[37] RECHER C., KIEBURG M. and GUHR T., *Phys. Rev. Lett.*, **105** (2010) 244101.

[38] RECHER C., KIEBURG M., GUHR T. and ZIRNBAUER M. R., *J. Statistical Phys.*, **148** (2012) 981.

[39] MIRLIN A. D., *Physics Reports*, **326** (2000) 259 .

[40] PORTER C. E. and THOMAS R. G., *Phys. Rev.*, **104** (1956) 483.

[41] EVERS F. and MIRLIN A. D., *Rev. Mod. Phys.*, **80** (2008) 1355.

[42] BEREZIN F., *Introduction to Superanalysis* (Reidel) 1987.

[43] *See Supplemental Material [url]*, () .

[44] FYODOROV Y. V., WEI Y. and ZIRNBAUER M., *Journal of Mathematical Physics*, **49** (2008) 053507.

[45] WALTNER D., WIRTZ T. and GUHR T., *Journal of Physics A: Mathematical and Theoretical*, **48** (2015) 175204.

[46] SHERMAN J. and MORRISON W. J., *The Annals of Mathematical Statistics*, **21** (1950) 124.

[47] VERBAARSCHOT J. J. M. and BUSSARD P. J., *Phys. Lett. B*, **87** (1979) 155.

Derivation of the supersymmetric integral representation for the moments of the eigenvectors. –
The quantity $I'_{1,b}$ defined in Eq. (9) can be computed by differentiating the generating function
(10)

$$Z_b(\lambda) = \left\langle \exp\left[i\alpha^2 e_b^T \begin{bmatrix} \sqrt{\lambda_+}\mathbf{1}_p & W \\ W^T & \sqrt{\lambda_+}\mathbf{1}_n \end{bmatrix}^{-1} e_b \right] \right\rangle \tag{20}$$

with respect to $i\alpha^2$ and setting $\alpha = 0$. The $(p+n)$-dimensional unit vector e_b has unity at the position b and zero otherwise. The normalization is given as $\lim_{\lambda \to \infty} Z_b(\lambda) = 1$. In order to construct a representation of $Z_b(\lambda)$ in terms of the supersymmetric integral we use the identity

$$\exp\left[i\alpha^2 e_b^T \begin{bmatrix} \sqrt{\lambda_+}\mathbf{1}_p & W \\ W^T & \sqrt{\lambda_+}\mathbf{1}_n \end{bmatrix}^{-1} e_b \right] \int d[\psi]d[\phi] \exp\left[-\mathrm{Str}\,(\phi,\psi)^T(\phi,\psi)J \right]$$

$$= \int d[\psi]d[\phi] \exp\left[i\mathrm{Str}\,(\phi,\psi)^T \begin{bmatrix} \sqrt{\lambda_+}\mathbf{1}_p & W \\ W^T & \sqrt{\lambda_+}\mathbf{1}_n \end{bmatrix} (\phi,\psi)J + 2\alpha\mathrm{Str}\,(\phi,\psi)^T(e_b,0,0,0) \right], \tag{21}$$

where we employed the matrix ψ which is an $(n+p)\times 2$ dimensional matrix of real Grassmann variables and ϕ is an $(n+p)\times 2$ dimensional ordinary real matrix. The two matrices are introduced in order to cancel the resulting determinants from the Gaussian integral. To ensure integrability we have introduced the constant matrix $J = \mathrm{diag}\,(\mathbf{1}_2; \tau_2)$, where τ_2 is the second Pauli matrix. It naturally drops out when performing the Gaussian integral (21).
To simplify the notation, we define the diagonal $(2|2) \times (2|2)$ supermatrix $L = \mathrm{diag}\,(-\mathbf{1}_2; \mathbf{1}_2)$ and the $(n+p) \times (2|2)$ rectangular supermatrix $E_b = (\alpha e_b, 0; 0, 0)$. Moreover, we rearrange the matrices ψ and ϕ in the $p \times (2|2)$ supermatrix V_L and the $n \times (2|2)$ supermatrix V_R as follows

$$(\phi, \psi) = \begin{pmatrix} V_L \\ V_R \end{pmatrix}. \tag{22}$$

Both matrices are two real rectangular supermatrices $V_L = V_L^*$ and $V_R = V_R^*$ with dimensions $p \times (2|2)$ and $n \times (2|2)$ respectively. The first two columns of V_L and V_R are real variables while the last two columns are Grassmann variables. In this way, we find

$$Z_b(\lambda) = \frac{\left\langle \int d[V_R, V_L] \exp\left[S(V_R, V_L, W) \right] \right\rangle}{\int d[V_R, V_L] \exp\left[S_0(V_R, V_L) \right]},$$

$$S(V_R, V_L, W) = -i\sqrt{\lambda_+}\mathrm{Str}\, LJ(V_L^T V_L + V_R^T V_R) - i\mathrm{Str}\, LJ(V_L^T W V_R + V_R^T W^T V_L) + 2\mathrm{Str}\,(V_L, V_R)^T E_b, \tag{23}$$

$$S_0(V_R, V_L) = -\mathrm{Str}\, J(V_L^T V_L + V_R^T V_R),$$

which is exactly (21) rewritten in the new notation. Note that we have used the circularity of the supertrace and trace, i.e., $\mathrm{tr}\, AB^T = \mathrm{Str}\, B^T A$ if A and B are two arbitrary $n \times (2|2)$ matrices.
The average over W_1 yields

$$Z_b(\lambda) = \frac{\int d[V_R, V_L] \exp\left[S_1(V_R, V_L) + S_2(V_R, V_L) \right]}{\int d[V_R, V_L] \exp\left[S_0(V_R, V_L) \right]},$$

$$S_1(V_R, V_L) = -i\sqrt{\lambda_+}\mathrm{Str}\, LJ(V_L^T V_L + V_R^T V_R) - i\mathrm{Str}\, LJ(V_L^T \sqrt{C_L} W_0 \sqrt{C_R} V_R + V_R^T \sqrt{C_R} W_0^T \sqrt{C_L} V_L) \tag{24}$$
$$+ 2\mathrm{Str}\,(V_L, V_R)^T E_b,$$

$$S_2(V_R, V_L) = -2\sigma^2 \mathrm{Str}\, V_L^T C_L V_L L J V_R^T C_R V_R L J,$$

where we have shifted W_1 by $iV_R L J V_L^T$ after completing the squares in the Gaussian. Since the action contains the quartic term $S_2(V_R, V_L)$ in the matrices V_L and V_R, the next step is to perform the Hubbard-Stratonovich transformation, which allows one to decouple such terms. Up to the normalization the result reads

$$Z_b(\lambda) \propto \int d[V_R, V_L] \int d[T] \exp\left[S_1(V_R, V_L) + S_3(V_R, V_L, T_+, T_-) \right],$$

$$S_3(V_R, V_L, T_+, T_-) = i\sigma^2 \mathrm{Str}\, J(V_L^T C_L V_L + V_R^T C_R V_R) - \frac{1}{2\sigma^2}\mathrm{Str}\,(T_+^2 + T_-^2) \tag{25}$$

$$+ i\mathrm{Str}\, T_+ JL(V_L^T C_L V_L + V_R^T C_R V_R + iJ) + \mathrm{Str}\, T_- JL(V_L^T C_L V_L - V_R^T C_R V_R).$$

When completing the squares of the auxiliary supermatrices T_+ and T_- one can readily check that $S_3(V_R, V_L, T_+, T_-)$ yields again the term $S_2(V_R, V_L)$.

The parametrization of the two $(2|2) \times (2|2)$ supermatrices $T_\pm$ needs to be chosen carefully to guarantee the convergence of the integral. They are given by

$$T_+ = \left[\begin{array}{cc} B_1 + iC(B_2) & \eta_1 \tau_2 \\ -\eta_1^T & iF_1 \end{array} \right] \text{ and } T_- = \left[\begin{array}{cc} B_2 & \eta_2 \tau_2 \\ -\eta_2^T & iF_2 \end{array} \right] \tag{26}$$

equipped with the flat Berezinian measure

$$d[T] = d[B_1]d[B_2]d[F_1]d[F_2]d[\eta_1]d[\eta_2]. \tag{27}$$

The ordinary matrices B_1 and B_2 are negative definite and symmetric and can be diagonalized with orthogonal matrices $O_1, O_2 \in O(2)$ as follows

$$B_1 = -O_1 b_1 O_1^{-1} \text{ and } B_2 = -O_2 b_2 O_2^{-1}, \tag{28}$$

with b_1, b_2 two positive definite diagonal matrices. The matrix $C(B_2)$ has the form

$$C(B_2) = -O_2 \sqrt{\mathbb{1}_2 + b_2^2} O_2^{-1}. \tag{29}$$

The matrices F_1 and F_2 are Hermitian self-dual matrices and η_1 and η_2 are two 2×2 rectangular matrices whose entries are independent real Grassmann variables. Such a non-trivial parameterisation is chosen in order to guarantee the convergence of the integral. Nonetheless, the number of independent ordinary variables as well as the number of Grassmann variables are 8 so that the Gaussian integral in (25) employed in the Hubbard-Stratonovich transformation can be still performed, especially we can shift T_+ by $iJL(V_L^T C_L V_L + V_R^T C_R V_R + iJ)$ and T_- by $JL(V_L^T C_L V_L - V_R^T C_R V_R)$ to regain the term $S_2(V_R, V_L)$ in (24).

The shift of B_1 in T_+ by the imaginary part $\sqrt{\mathbb{1}_2 + B_2^2}$ solves a convergence problem in the Gaussian terms in (25). In particular the Gaussian integrals over the supermatrices V_L and V_R are absolutely convergent and yield

$$Z_b(\lambda) \propto \int d[T] \exp\left[-\frac{1}{2\sigma^2} \text{Str}\,(T_+^2 + T_-^2) - \text{Str}\,T_+ L - i\alpha^2 G_{1b,1b} \right] \sqrt{\text{Sdet}\,G}, \tag{30}$$

where G is defined as in (12). The supermatrix G^{-1} can be read off from the quadratic terms in V_R and V_L in the terms $S_1(V_R, V_L)$ and $S_3(V_R, V_L)$. The term $i\alpha^2 G_{1b,1b}$ in the exponent results from completing squares with the linear term $2\text{Str}\,(V_L, V_R)^T E_b$ in $S_1(V_R, V_L)$ and the superdeterminant $\sqrt{\text{Sdet}\,G}$ originates from the Gaussian integral itself. Due to the multiplication of V_R and V_L with matrices and supermatricces from both sides, $G_{\mu a, \nu b}$ has four indices with $\mu, \nu = 1, \ldots, 4$ and $a, b = 1, \ldots, n + p$.

To fix the normalization we take $\lambda \to \infty$ and notice that G becomes approximately $\lambda_+^{-1/2} \mathbb{1}_{n+p} \otimes LJ$. Therefore we end up with the intermediate result

$$Z_b(\lambda) = \frac{\int d[T] \exp\left[-\frac{1}{2\sigma^2} \text{Str}\,(T_+^2 + T_-^2) - \text{Str}\,T_+ L - i\alpha^2 G_{1b,1b} \right] \sqrt{\text{Sdet}\,G}}{\int d[T] \exp\left[-\frac{1}{2\sigma^2} \text{Str}\,(T_+^2 + T_-^2) - \text{Str}\,T_+ L \right]}. \tag{31}$$

Coming back to our original problem, we notice that we are interested in the first derivative with respect to $i\alpha^2$ at $\alpha = 0$. In particular, the quantity $I_{1,b}(\lambda)$ is given by

$$I_{1,b}(\lambda) = \frac{1}{\sqrt{\lambda_+}} \frac{\int d[T] \exp\left[-\frac{1}{2\sigma^2} \text{Str}\,(T_+^2 + T_-^2) - \text{Str}\,T_+ L \right] G_{1b,1b} \sqrt{\text{Sdet}\,G}}{\int d[T] \exp\left[-\frac{1}{2\sigma^2} \text{Str}\,(T_+^2 + T_-^2) - \text{Str}\,T_+ L \right]}, \tag{32}$$

which coincides with Eq. (11). A similar strategy can be used in principle in order to calculate the general moment $I_{q,b}(\lambda)$. However, in that case one would need to introduce a different generating function containing two inverse matrices instead of one. This would double the number of the integration variables in the Gaussian integral and lead to appearance of $(4|4) \times (4|4)$ supermatrices, which would make the corresponding counterpart of Eq. (32) much more complicated.

Saddle-point equation. – For deriving the saddle-point equation we only need to consider the exponential function and the superdeterminant in the integral (32). The term $G_{1b,1b}$ is only a polynomial prefactor which does not influence the saddle-point solution. It is easier to study the saddle-point by introducing the supermatrices $S = T_+ - iT_- + \sigma^2 L$ and $R = T_+ + iT_- + \sigma^2 L$, which can be considered to be independent. Then the action, i.e., the function, that needs to be minimised, is

$$\frac{1}{2\sigma^2(n+p)}\text{Str}\,SR + \frac{1}{2(n+p)}\text{Str}\ln\left[\begin{array}{cc} \sqrt{\lambda_+}C_L^{-1} \otimes \mathbb{1}_{2|2} - \mathbb{1}_p \otimes S & W_0 \otimes \mathbb{1}_{2|2} \\ W_0^T \otimes \mathbb{1}_{2|2} & \sqrt{\lambda_+}C_R^{-1} \otimes \mathbb{1}_{2|2} - \mathbb{1}_n \otimes R \end{array}\right], \tag{33}$$

where the second term is the superdeterminant. Differentiating it with respect to S and R yields two coupled equations

$$R - \sigma^2 \text{tr}\,_1\left[\sqrt{\lambda_+}C_L^{-1} \otimes \mathbb{1}_{2|2} - (W_0 \otimes \mathbb{1}_{2|2})(\sqrt{\lambda_+}C_R^{-1} \otimes \mathbb{1}_{2|2} - \mathbb{1}_n \otimes R)^{-1}(W_0^T \otimes \mathbb{1}_{2|2}) - \mathbb{1}_p \otimes S\right]^{-1} = 0,$$

$$S - \sigma^2 \text{tr}\,_1\left[\sqrt{\lambda_+}C_R^{-1} \otimes \mathbb{1}_{2|2} - (W_0^T \otimes \mathbb{1}_{2|2})(\sqrt{\lambda_+}C_L^{-1} \otimes \mathbb{1}_{2|2} - \mathbb{1}_p \otimes S)^{-1}(W_0 \otimes \mathbb{1}_{2|2}) - \mathbb{1}_n \otimes R\right]^{-1} = 0. \tag{34}$$

Here, we have employed the relation $\text{Sdet}\begin{pmatrix} \hat{A} & \hat{B} \\ \hat{C} & \hat{D} \end{pmatrix} = \text{Sdet}(\hat{A} - \hat{B}\hat{D}^{-1}\hat{C})\text{Sdet}\,D$ and the expansion $\text{Str}\ln(A+\lambda B) = \text{Str}\ln(A) + \lambda\text{Str}\,A^{-1}B + o(\lambda)$ for a small parameter λ and two supermatrices A and B. Both relations can be found in [42]. The operator $\text{tr}\,_1$ is the partial trace over the first tensor space which is here the space of ordinary $n \times n$ and $p \times p$ matrices, respectively. It appears due to the fact that S and R only act on the superspace but not on the ordinary part.

The saddle-point equation is rotation invariant, i.e., when (S_0, R_0) is a solution then this is also true for $(R_0 S_0 R_0^{-1}, R_0)$ as well as $(S_0, S_0 R_0 S_0^{-1})$ and any kind of combination. This can be seen by multiplying both equations from the left and the right with R and R^{-1}, which is equivalent to replacing S by RSR^{-1}. Assuming that the saddle-point solution (S_0, R_0) is unique, we conclude that S_0 and R_0 must commute. The uniqueness of the solution should follow from the contour of the integration, which was shifted by the term $i\epsilon$. This contour cannot cross the poles. Additionally, the Berezinian (the Jacobian in superspace), that is $|b_{1j} - b_{2j}|/[(b_{1j} - if_j)^2(b_{2j} - if_j)^2]$ for $j = 1, 2$, is not suppressed only when the multiplicity of the eigenvalues in the Fermion-Fermion blocks is equal to those in the Boson-Boson block. The Fermion-Fermion blocks are doubly degenerate due to their Hermitian self-duality. Thus also the Boson-Boson blocks are doubly degenerate, which implies for $(2|2) \times (2|2)$ supermatrices that we can diagonalize S and R simultaneously and the solution has to be diagonal and degenerate, i.e., $S_0 = s_0\mathbb{1}_{2|2}$ and $R_0 = r_0\mathbb{1}_{2|2}$. Substituting this ansatz into Eq. (34), we obtain two scalar equations which are Eq. (13) and explicitly read

$$\frac{r_0}{\sigma^2} = \text{tr}\left[\sqrt{\lambda_+}C_L^{-1} - W_0(\sqrt{\lambda_-}C_R^{-1} - r_0\mathbb{1}_n)^{-1}W_0^T - s_0\mathbb{1}_p\right]^{-1},$$

$$\frac{s_0}{\sigma^2} = \text{tr}\left[\sqrt{\lambda_+}C_R^{-1} - W_0^T(\sqrt{\lambda_-}C_L^{-1} - s_0\mathbb{1}_p)^{-1}W_0 - r_0\mathbb{1}_n\right]^{-1}. \tag{35}$$

The ϵ regularization only determines which saddle-point has to be chosen, especially which sign the imaginary part carries. Assuming the correct sign of the imaginary part we neglected this regularization in Eq. (13).