Localized short-range correlations in the spectrum of the equal-time correlation matrix

Markus Müller,[*] Yurytzy López Jiménez, Christian Rummel, and Gerold Baier
Facultad de Ciencias, Universidad Autónoma del Estado de Morelos, 62209 Cuernavaca, Morelos, México

Andreas Galka
Institut für Experimentelle und Angewandte Physik, Christian-Albrechts-Universität, 24098 Kiel, Germany
and Institute of Statistical Mathematics (ISM), Minami-Azabu 4-6-7, Minato-ku, Tokyo 106-8569, Japan

Ulrich Stephani and Hiltrud Muhle
Klinik für Neuropädiatrie, Christian-Albrechts-Universität, 24105 Kiel, Germany

(Received 15 May 2006; published 24 October 2006)

We suggest a procedure to identify those parts of the spectrum of the equal-time correlation matrix **C** where relevant information about correlations of a multivariate time series is induced. Using an ensemble average over each of the distances between eigenvalues, all nearest-neighbor distributions can be calculated individually. We present numerical examples, where (a) information about cross correlations is found in the so-called "bulk" of eigenvalues (which generally is thought to contain only random correlations) and where (b) the information extracted from the lower edge of the spectrum of **C** is statistically more significant than that extracted from the upper edge. We apply the analysis to electroencephalographic recordings with epileptic events.

DOI: 10.1103/PhysRevE.74.041119 PACS number(s): 02.50.Sk, 05.45.Tp, 89.75.−k, 05.10.−a

I. INTRODUCTION

In recent years, the application of tools known from random matrix theory (RMT) to time series analysis has become more and more popular. The application of RMT techniques to a variety of multivariate data sets like financial data [1–6], electroencephalographic [7], magnetoencephalographic recordings [8], climate data [9], internet traffic [10], and others has been reported. One of the main goals of the employment of RMT measures in time series analysis is to separate in the spectrum of the equal-time correlation matrix **C** genuine information about the correlation structure from random correlations (generated by the finite size of the time window used to construct **C**) and noise. In general, the assumption was made that those eigenvalues λ and respective eigenvectors $\vec{v}$ of the correlation matrix, which can be described by random matrix ensembles are dominated by random correlations and noise, whereas deviations from the RMT behavior indicate true information about the correlation structure. The natural null hypothesis for the comparison of empirical results is Wishart ensembles (WE), i.e., ensembles of correlation matrices constructed from signals of independent Gaussian white noise. WE are characterized by two parameters, the length T and the dimension M of the multivariate data set.

In Ref. [1], properties of the empirical correlation matrix constructed from financial data have been studied. A clear separation of a few large eigenvalues from the remaining ones was observed. It was shown, that apart from the few largest λ, the level density $\rho(\lambda)$ of the empirical correlation matrix can be approximately fitted by the analytical formula for the WE, which is valid in the limit $T,M\to\infty$ with $T/M = \text{cst.} > 1$. In Ref. [5] it was demonstrated (see Fig. 3 of Ref. [5]) that the observed deviations from the level density of the corresponding WE are not caused by the finite size of the empirical data set. It was argued that these deviations are due to the influence of the few well-separated large eigenvalues.

The results of Ref. [1] have been confirmed by a series of subsequent papers, analyzing not only the level density but also applying more sophisticated RMT tools, which measure the correlation properties of the spectrum of eigenvalues, such as, e.g., the nearest-neighbor distribution $P(s)$ or the number variance $\Sigma^2(l)$ (see, e.g., Refs. [11,12]). In all those papers empirical results have been compared to the analytically known ones of the Gaussian orthogonal ensemble (GOE). Strictly speaking, an ensemble of (empirical or random) correlation matrices does not belong to the GOE, but as the differences of the statistical properties between WE and GOE decay rapidly as one goes away from zero in the spectrum of the correlation matrix [5], the GOE seems to be an appropriate choice for a null hypothesis for these measures.

Independently of the type of data considered, the application of RMT tools seemed to confirm the statement that the "*bulk*" of eigenvalues (i.e., all those below a few, well-separated ones) are strongly contaminated by noise and consequently do not contain any valuable information. Within the statistical errors, the nearest-neighbor statistics as well as the results for the number variance calculated from the empirical data coincide very well with the universal properties of the GOE. These results led many authors to the generally accepted conclusion that only the largest eigenvalues (and their respective eigenvectors) contain relevant information, while the remaining part of the spectrum, the bulk, is mainly dominated by noise. Such statements are in agreement with the philosophy of the principal component analysis (PCA), where only a certain number of the largest eigenstates of the covariance matrix are used to project the data onto its principal components [13].

[*] Electronic address: muellerm@buzon.uaem.mx

However, the results presented in Ref. [1,3] and subsequent publications met a certain criticism. Besides the fact that in Ref. [1] no formal statistical test was applied in order to corroborate the final conclusion, in Refs. [14,15] it was analytically shown and confirmed by numerical examples that the level density of the so-called bulk is certainly *not* in the Wishart class and, hence, might contain information about the correlation structure. Similar conclusions are drawn in Ref. [16]. Moreover, in Refs. [2,5,6] significant deviations from the GOE at the *lower edge* of the spectrum of the empirical correlation matrix were reported: The dominant components of the eigenvectors corresponding to the lowest eigenvalues stem from pairs of time series that have the largest cross-correlation coefficients C_{ij} of the whole sample.

This observation can be understood by the fact that genuine information about the correlations is induced in the spectrum of the correlation matrix via (nonrandom) repulsions of eigenstates between both edges [6,17]: The special features of the correlation structure of the multivariate data set determine how many states at the edges repel and how strong the displacement of the eigenvalues will be [17]. On the other hand, if there are large groups of correlated time series, the same mechanism can induce a (possibly subtle) repulsion between eigenvalues that are far from the boundaries of the spectrum. Such information can be washed out by the spectral average, which is usually employed while calculating the RMT correlation measures.

The aim of the present paper is to confirm the usefulness of the RMT measures for the application to data analysis from a slightly different point of view. By simply normalizing each of the level distances by its ensemble average, one can calculate the nearest-neighbor distribution for each of the distances *individually*. This way of unfolding permits us to distinguish precisely those parts of the spectrum that are dominated by noise from those that carry valuable information about the correlation structure of a system. The proposed method does not depend on the spectral region in which the information is induced. Furthermore, we present examples where the information extracted at the lower part of the spectrum is statistically more significant than the results obtained from the largest eigenvalues and give clear evidence that genuine information within the central part of the spectrum of **C** is washed out when spectral averages are employed. Finally, we show that the results presented so far are not only theoretical considerations but may have impact on the analysis of such experimental data where temporal changes of the correlation structure are mainly visible at the lower edge of the spectrum of the correlation matrix.

II. BASIC METHODOLOGY

The equal-time correlation matrix is constructed from an M-dimensional multivariate data set $X_i(k)$ of length T, where $i=1,\ldots,M$ and $k=1,\ldots,T$. First the raw time series is normalized according to

$$\tilde{X}_i(k) = \frac{X_i(k) - \bar{X}_i}{\sigma_i}, \tag{1}$$

where $\bar{X}_i$ and σ_i, respectively, denote the mean value and standard deviation of $X_i(k)$. Then the equal-time correlation matrix **C** is computed by [18]

$$C_{ij} = \frac{1}{T}\sum_{k=1}^{T} \tilde{X}_i(k)\tilde{X}_j(k). \tag{2}$$

With λ_i and $\vec{v}_i$ we denote the eigenvalues and eigenvectors of the correlation matrix **C**,

$$\mathbf{C}\vec{v}_i = \lambda_i \vec{v}_i, \tag{3}$$

where the eigenvalues λ_i shall be ordered according to size, such that $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_M$. The distribution of the eigenvalues is directly related to the amount of correlations of the multivariate data set [17].

The normalization, Eq. (1), has the effect of removing amplitude information and a possible offset from each of the $X_i(k)$, such that the measured cross correlations exclusively reflect relationships between the shapes and phases of the signals. In addition, the normalization provides a well-defined scale, i.e., the matrix elements of **C** vary between +1 (completely correlated series) and −1 (completely anticorrelated series). Furthermore, the normalization ensures that

$$\mathrm{Tr}(\mathbf{C}) = \sum_i C_{ii} = \sum_i \lambda_i = M. \tag{4}$$

Therefore, if one of the eigenvalues alters, its change has to be compensated by some others such that condition (4) is still fulfilled.

Properties of interest in RMT theory are correlations within the spectrum of eigenvalues of a given (Hamiltonian) matrix. The two most prominent measures are the nearest-neighbor distribution $P(s)$ and the number variance $\Sigma^2(l)$ [11,12,19,20]. The calculation of these measures requires the transformation of the level density $\rho(\lambda)$ to a uniform distribution such that the average level distance is unity everywhere; this procedure is known as "*unfolding*" [12,20]. To this end, one usually calculates the so-called accumulated level density

$$N(\lambda) = \int_{-\infty}^{\lambda} \rho(\lambda')d\lambda', \tag{5}$$

which counts the number of states in the interval $[-\infty,\lambda]$. It can be split into a smooth and a fluctuating part

$$N(\lambda) = N_{\text{smooth}}(\lambda) + N_{\text{fluct}}(\lambda). \tag{6}$$

Because the fluctuating part of the level density

$$\rho_{\text{fluct}}(\lambda) = \frac{dN_{\text{fluct}}(\lambda)}{d\lambda} \tag{7}$$

vanishes on the average, the mean level density is solely given by

$$\rho_{\text{smooth}}(\lambda) = \frac{dN_{\text{smooth}}(\lambda)}{d\lambda}. \tag{8}$$

Once the smooth part $N_{\text{smooth}}(\lambda)$ of the accumulated level density is known, the eigenvalues Λ_i of the unfolded spectrum can be calculated by the transformation

$$\Lambda_i = N_{\text{smooth}}(\lambda_i). \tag{9}$$

Hence, the crucial problem of unfolding a spectrum is to find the correct form of $\rho_{\text{smooth}}(\lambda)$. In cases where the analytical formula for the smooth part of the level density is unknown, one has to perform a fit of the numerical probability distribution of the eigenvalues. One common procedure is a polynomial fit to the numerically obtained accumulated level density $N(\lambda)$. This fit function is then used to perform the unfolding transformation (9).

A different approach was followed in Refs. [2,5]. It was assumed that the theoretically known expression for the level density of the WE can be used for $\rho_{\text{smooth}}(\lambda)$ in the unfolding procedure. In the limits $M\rightarrow\infty$ and $T\rightarrow\infty$, such that $Q=T/M$ stays constant, the probability distribution of the eigenvalues of the WE is given by [21]

$$P_W = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}. \tag{10}$$

In Eq. (10) $\lambda_\pm$ are given by

$$\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}} \tag{11}$$

and $\sigma^2 = 1/M\Sigma_{i=1}^{M}\sigma_i^2$ denotes the total variance of the $X_i(k)$. For finite values of M and T, deviations occur at both edges of the spectrum [22]. Using $\lambda_\pm$ [5] or the variance σ^2 of the signal [1] as fit parameters and integrating Eq. (8) the smooth part of the accumulated level density $N_{\text{smooth}}(\lambda)$ can be obtained, which in turn is used for the unfolding procedure (9). In principle, neither $\lambda_\pm$ nor the variance of the multivariate data sets are free parameters; however, the resulting fit might describe the level density of the empirical correlation matrix satisfactorily well.

After the unfolding procedure, correlation functions such as $P(s)$ or $\Sigma^2(l)$ can be extracted numerically and consequently used for comparison with the analytical results, as derived from random matrix ensembles. Independent of the way by which the fit procedure (as required by the unfolding) is designed, the subsequent calculation of the RMT correlation measures resembles an average over a certain central interval of the unfolded level density (spectral average [20]); the boundaries usually have to be neglected. This causes a twofold problem: First, as correlations in many cases induce level repulsions at the edges of the eigenvalue spectrum, exactly these interesting repelling levels might be left out when calculating the correlation measures in the usual way. Second, even if the level repulsion occurs in the central part of the spectrum, the extractable information can be localized in a markedly narrow part of the whole spectrum—on occasion in the distance between two states only. In such cases, small localized deviations from the universal properties of the GOE might be washed out due to the averaging process.

In order to be able to check how the nonrandom level repulsion caused by a specific correlation structure of the multivariate data set influences the results of the RMT measures, we propose an *individual unfolding* for each of the $M-1$ distances by

$$s_i = \frac{\lambda_{i+1} - \lambda_i}{\langle\lambda_{i+1} - \lambda_i\rangle}, \tag{12}$$

where $\langle\lambda_{i+1}-\lambda_i\rangle$ is the ensemble average of the distance between two neighboring eigenvalues. The normalization (12) allows a clean unfolding along the whole spectrum yielding $\langle s_i\rangle=1$ for all $i=1,\ldots,M-1$ without the application of any kind of fit procedure. Numerical inaccuracies, which might occur especially at the edges of the spectrum, are avoided completely. Even more important at this place is that this procedure permits us to calculate the nearest-neighbor distribution for each of the distances separately. Hence, one is able to detect precisely in which part of the spectrum deviations from the RMT predictions appear. Before applying Eq. (12) to numerical and empirical test data, we have successfully checked the performance of this unfolding at ensembles of GOE and GSE matrices as well as Poisson distributions.

III. MODEL SYSTEMS

As a first model system we consider a standard class of dynamical systems; the sum of N_f sine waves with mutually incommensurate frequencies,

$$X_i(k) = \sum_{j=1}^{N_f} \sin(2\pi f_j k + \delta_{ij}), \quad i = 1, \ldots, M. \tag{13}$$

Here we choose equal amplitudes for all sines. Data sets generated by such systems sample N_f-tori [23], i.e., geometrical objects of topological dimension equal to N_f. Because the correlation structure of this system class can be perfectly controlled, it serves as an ideal test system.

In our application of Eq. (13) the set of frequencies f_j is sampled randomly from a uniform distribution on the interval $[0,f_{\max}]$ with $f_{\max}=20$ frequency units, where a time unit is arbitrarily defined as 256 data points. The same set of frequencies is chosen for all i, whereas the set of random partial phases δ_{ij} is assigned from the interval $[0,2\pi]$ to each time series X_i independently. In all examples of N_f-tori we choose $N_f=1000$ frequencies and $M=20$ time series. For the length of the time series we take $T=2048$ corresponding to 8 time units. We create an ensemble of $N=100\,000$ independent realizations of $X_i(k)$ and calculate the correlation matrices (2).

The correlation structure of the multivariate data set (13) is controlled by the partial phases δ_{ij}. If the δ_{ij} are uniformly distributed between 0 and 2π, the X_i are mutually uncorrelated. Any deviation from the uniform distribution will cause phase-shape correlations between the time series. In order to generate correlations, we choose for some $X_i(k)$, Gaussian distributed δ_{ij} modulo 2π with prespecified mean μ_δ and variance σ_δ.

In order to treat an example of a real dynamical system, we employ data sets from coupled Rössler systems [17],

$$\dot{X}_{1,2} = -\omega_{1,2}Y_{1,2} - Z_{1,2} + \eta(X_{2,1} - X_{1,2}),$$

$$\dot{Y}_{1,2} = \omega_{1,2}X_{1,2} + \beta Y_{1,2} + \eta(Y_{2,1} - Y_{1,2}),$$

$$\dot{Z}_{1,2} = 0.1 + Z_{1,2}(X_{1,2} - 8.5), \qquad (14)$$

with $\omega_1 = 0.98$, $\omega_2 = 1.03$, and $\beta = 0.28$. η denotes the coupling strength between two oscillators. The Rössler system is of special interest in the present context, because in Ref. [24] it was argued, that due to the funnel characteristics of the underlying attractor, the detection of relationships like phase synchronization between two coupled units is particularly difficult. Therefore the Rössler system is an ideal testing ground for the method developed in this paper. Instead of using only two coupled units like in Ref. [24], we consider a set of $M = 20$ Rössler equations (14), where only two are coupled as described by Eq. (14) with a nonzero value of η, while the $M-2$ remaining systems stay uncoupled ($\eta = 0$). We generate a multivariate time series by sampling the X components of each system.

IV. RESULTS OBTAINED FOR SIMULATED SYSTEMS

As a first example of a system, where the correlation pattern can be perfectly controlled, we consider N_f-tori [Eq. (13)]. For the completely uncorrelated case, viz., when the partial phases of all signals are uniformly distributed, the correlation matrix of the N_f-tori are close to Wishart matrices of the same dimension. Consequently, the nearest-neighbor distribution of all level distances of the uncorrelated N_f-tori show GOE behavior with high accuracy (results not presented in a figure). For only two outermost distances we observe slight deviations to the theoretical curve for the GOE. This border effect is of the same order of magnitude as for an equivalent ensemble of GOE or Wishart matrices (when $T \gg M \gg 1$).

First we consider N_f-tori where only 2 of $M = 20$ series are correlated via Gaussian distributed partial phases with the same mean $\mu_\delta = \pi/2$ and standard deviation $\sigma_\delta = 3\pi/2$. For the creation of the remaining $X_i(k)$, uniformly distributed random partial phases are used. After diagonalizing the correlation matrix (2) we perform the individual unfolding [Eq. (12)] in order to calculate the nearest-neighbor distribution for each distance. The results at the edges are shown in Fig. 1. As one can see, a clear deviation from the universal results of the GOE appears for s_1 and s_{19}. Even for comparatively poor statistics of only 1000 matrices (histograms drawn with a gray line) the deviations appearing for $P(s_1)$ and $P(s_{19})$ are obvious. The distances between the two lowest and largest eigenvalues show a comparatively narrow distribution around unity whereas the behavior of all other distances $s_2 \cdots s_{18}$ can be perfectly described by the universal properties of the GOE.

A similar test system has already been studied in Ref. [17]. It has been found that independently of which of the two $X_i(k)$ correlate a repulsion between the largest and lowest eigenvalue λ_i occurred. As Fig. 1 confirms, not only the average magnitude of the corresponding level distances increases, but also its stiffness. Note that for comparison in this and all other figures of this paper, we show the nearest-neighbor distribution for the GOE (and not the Wigner surmise).

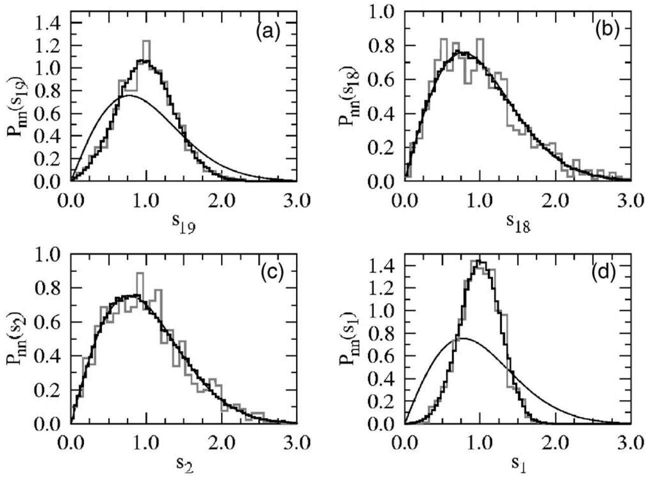

FIG. 1. Nearest-neighbor distribution obtained by individual unfolding for N_f-tori where two from 20 time series are correlated due to Gaussian distributed partial phases with mean $\mu_\delta = \pi/2$ and standard deviation $\sigma_\delta = \pi/2$. Shown are the results for (a) s_{19}, (b) s_{18}, (c) s_2, and (d) s_1. The theoretical result for the GOE is drawn as a solid line. The histograms drawn by a black and gray line has been calculated for an ensemble of 100 000 and 1000 matrices, respectively.

Two aspects of the results shown in Fig. 1 are remarkable. First, a sharp cut between the noisy and nonrandom parts of the spectrum is clearly visible in the spacing statistics. The diagonalization of the correlation matrix provides a kind of filter, which distinguishes precisely between random and nonrandom eigenstates. Second, within the given statistics no border effects are visible, not even in the distribution for s_2 and s_{18}. In the example shown in Fig. 1 the behavior of the random part of the spectrum is described perfectly well by the universal behavior of the central part of the GOE.

In a second example we investigate the properties of the Rössler system (14) where two of 20 units are coupled with $\eta = 0.049$. In Fig. 2 we show the nearest-neighbor distribution for s_{19} and s_1. Like in the former example, all other distances behave exactly like the GOE in the center. What is remarkable in this case is that the deviations from GOE are much more pronounced in s_1 than at the upper edge of the spectrum of **C**. For an ensemble of only 1000 matrices [Fig. 2(a)], the Kolmogorov-Smirnov test (KS test) [25] gives a 100% coincidence with the GOE for s_{19} and even for 10^5 matrices [Fig. 2(b)] the KS test gives a probability of 99.7% that the empirical distribution is of GOE-type. Whereas in the latter case a careful visual inspection of the figure reveals a slight shift of the histogram to the left, such an effect is not visible for the poorer statistics. For the distance s_1 [Fig. 2(c)] the KS test rejects completely the GOE hypothesis, even for poor statistics of only a few hundred matrices. It is evident that in the given example, the information extracted at the lower edge is statistically much more significant than for the largest eigenvalue. The same tendency, although not as pro-

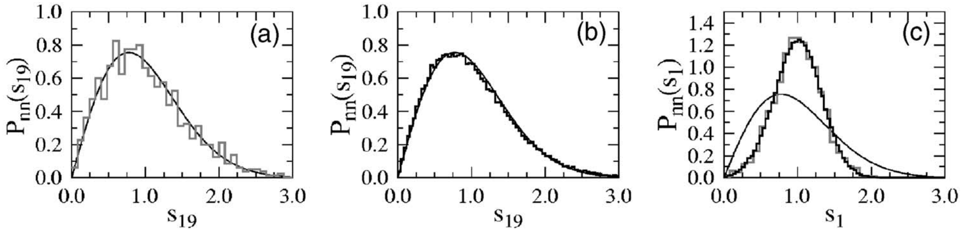

FIG. 2. Nearest-neighbor distribution obtained by individual unfolding for the Rössler system where two from 20 time series are correlated via diffusive coupling with strength η=0.049. Shown are the results for distance s_{19} considering an ensemble of (a) 1000 and (b) 100 000 matrices and (c) the nearest-neighbor distribution for distance s_1 for 1000 (gray histogram) and 100 000 (black histogram). The theoretical result for the GOE is drawn as a solid line.

nounced as in the given example, can be already seen in the former case of the N_f-tori (see Fig. 1). Also, here the distribution for s_1 is slightly narrower than that for s_{19}.

The differences between $P(s_{19})$ and $P(s_1)$ can be explained by the special role of the eigenvector $\vec{v}_M$ of **C** that corresponds to the largest eigenvalue. This eigenvector is in some sense the most collective one [26–28] and may act as a kind of "doorway state" [29]. Even in the extreme case where only a small subset of signals $X_i(k)$ correlate, the eigenvector $\vec{v}_M$ of **C** takes up contributions not only within the correlated subspace but also from all other basis states [17]. As a consequence, $\vec{v}_M$ is always influenced by both the random and nonrandom parts of the correlations of the whole data set. For the same reason, also tiny correlations between all $X_i(k)$ might cause a huge separation of the largest eigenvalue from the rest of the spectrum [14] and for the case of financial data, the behavior of $\vec{v}_M$ has been interpreted as the dynamics of the entire market [2,4,5]. The eigenvector $\vec{v}_1$ corresponding to the smallest eigenvalue, on the other hand, is almost exclusively determined by genuine correlations (at least in situations where only a few time series correlate) and therefore nonrandomly oriented in the correlating subspace (see Fig. 3 of Ref. [17]). As a consequence, the nearest-neighbor distribution of s_1 deviates more from the GOE than s_{M-1}, or s_{19} in our case.

Next we study a case where the signature of correlations between the $X_i(k)$ is implemented in the central part of the spectrum of **C**. In Ref. [17] it was shown that correlations between $K<M$ signals $X_i(k)$ affect K eigenvalues such that the $K-1$ smallest ones decrease whereas the largest eigenvalue will be displaced to higher values. In this case we expect the distances $P(s_{K-1})$ and $P(s_{M-1})$ to deviate from the GOE. Again we employ the N_f-tori in order to simulate such a situation where $K=8<M=20$ signals are correlated with $\mu_\delta=\sigma_\delta=\pi/2$. Then we normalize the distances to its ensemble average [Eq. (12)] and calculate the nearest-neighbor distributions. Some results are shown in Fig. 3. As expected, differences to the GOE are only visible for $P(s_7)$ [Fig. 3(b)] and $P(s_{19})$ (not shown). For the remaining ones a KS test gives a 100% coincidence for the numerical results with the GOE, whereas the differences between $P(s_7)$ and the GOE are still visible even with poor statistics.

Again, as in the examples discussed before, the part of the spectrum where genuine information is induced is well separated from the random part. Even the distances in the direct neighborhood of s_7 and s_{19} reproduce the GOE behavior with high accuracy. The KS test provides a 100% probability for the coincidence with the GOE result for all spacings, with the exception of s_7 and s_{19}. If we calculate a spectral average, over all distances with the exception of the one to the well-separated largest level, which is equivalent to the unfolding procedure applied in Refs. [1–10], the averaged nearest-neighbor distribution becomes close to the GOE prediction (Fig. 4). Calculating the spectral average, the effect of the nonrandom displacement of the eigenstates 1 to 7 is washed out and the tiny differences between the GOE and the spectral averaged nearest-neighbor distributions of the correlated N_f-tori are visible for large ensembles only—a requirement that is hardly fulfilled for the majority of experimentally accessible data. For the given case, the KS test gives a 100% probability for the coincidence of the spectral averaged spacing statistics with the GOE. A very similar result is obtained for the case of eight coupled Rössler systems.

V. ANALYSIS OF EXPERIMENTAL DATA

Up to now, our arguments were based on model-generated data. In order to put the results in a more general context we

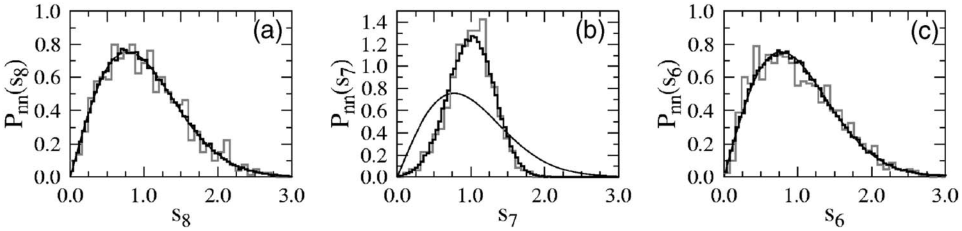

FIG. 3. Nearest-neighbor distribution obtained by individual unfolding for N_f-tori where eight from 20 time series are correlated with $\sigma_\delta=\pi/2$. Shown are the results for (a) s_8, (b) s_7, and (c) s_6. The theoretical result for the GOE is drawn as a solid line. The histograms drawn by a black and gray line has been calculated for an ensemble of 100 000 and 1000 matrices, respectively.

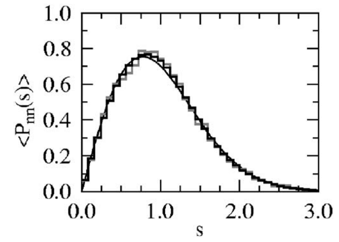

FIG. 4. Spectral averaged nearest-neighbor distribution for N_f-tori where eight from 20 time series are correlated. The theoretical result for the GOE is drawn as a solid line. The histograms drawn by a black and gray line has been calculated for an ensemble of 100 000 and 1000 matrices, respectively.

demonstrate its relevance by analyzing real-world data. In this paper we consider an electroencephalogram (EEG) of a ten-year-old male patient suffering from generalized epilepsy (absence seizures). Figure 5(a) presents a segment of 125 s of the electric potential from one of the $M=19$ scalp electrodes (F3 over the left frontal lobe in clinical terminology). The time series were referenced in a source density derivation according to the Hjorth method [30]. The data were sampled with a 12-bit A/D conversion at a sampling rate of 256 Hz and the 50 Hz ac component was filtered. While the EEG of normal brain dynamics looks irregular and only partially correlated, during an epileptic seizure of absence type the EEG displays a strongly correlated rhythm of comparatively large amplitude and well-defined frequency close to 3 Hz. The epileptic activity of the examined example starts almost simultaneously in all electrodes at about 499 s and lasts up to 512 s. At approximately 405 and 485 s artifacts caused by muscular movement can be seen.

When analyzing experimental data such as EEG recordings with the methods proposed in the present paper, two questions arise. First, correlation matrices have to be constructed from possibly nonstationary data. To this end one can shift a window over the time series or parts of it. In the present paper we have chosen $T=75$ data points (corresponding to approximately 0.29 s) for the window size and a maximal overlap of $T-1$ points. The second question is how to define an appropriate null hypothesis for the comparison with the empirical nearest-neighbor distributions. While for the WE the GOE seems to be adequate, for the present empirical ensembles of correlation matrices this null hypothesis cannot be used. The brain activity changes continuously even in the resting state and therefore also a continuously changing correlation structure has to be expected. Moreover, static correlations different from those of the GOE are induced due to the chosen reference point [31], in our case the Hjorth transformation. For such reasons it seems appropriate to create the null hypothesis for the present data from the experimental recordings itself.

An adequate null hypothesis can be an average over all possible dynamical states of the brain during the resting state, avoiding activities like eye movement, swallowing, and other artifacts. Statistically significant deviations from this average are then believed to indicate relevant correlations distinct from those of the usual resting behavior. For this purpose we defined an artifact-free reference interval of the same EEG recording between 200 and 250 s. Shifting T over this segment an ensemble of 12 726 correlation matrices is created. For each of the 18 distances between the eigenvalues we calculate the nearest-neighbor distribution, which is then used as the null hypothesis for further analysis.

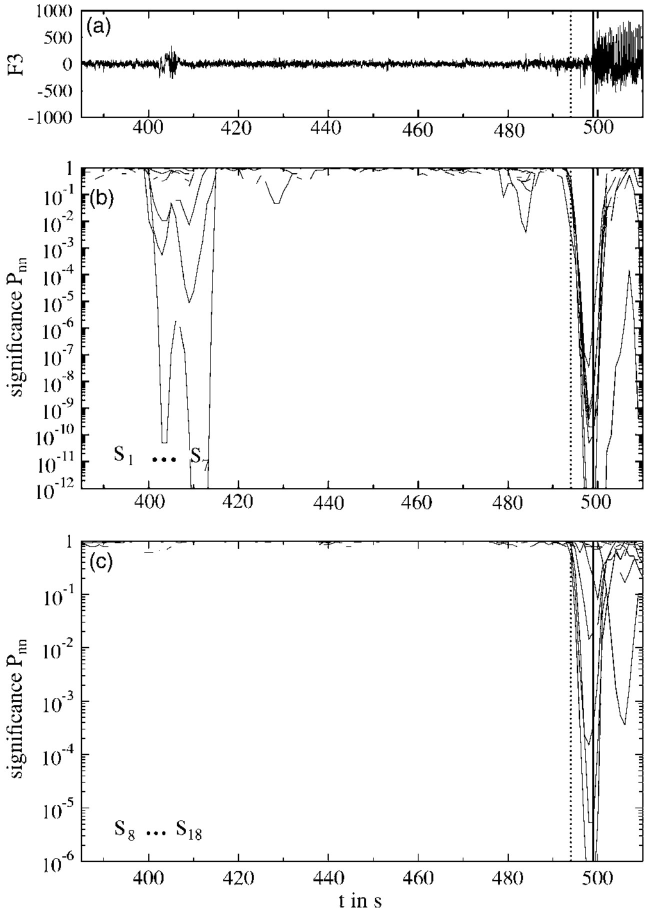

FIG. 5. (a) A segment of 125 s from an EEG recording at electrode F3. From about 499 to 512 s the large amplitude activity of a generalized epileptic seizure can be seen. At approximately 405 and 485 s artifacts caused by muscular movement occur. (b) KS probability for the coincidence of the nearest-neighbor distributions of the first seven distances with the null hypothesis and (c) the same as in (b) for the distances s_8 to s_{18}. The null hypothesis was calculated as an average over an artifact-free period between 200 to 250 s of the same recording. The solid vertical line marks the onset of the seizure. The dashed vertical line indicates when the moving window starts to overlap with the seizure period.

The analysis of the EEG data is performed as follows. Within an interval of $I=10$ s of the data set an ensemble of 2485 correlation matrices are constructed by moving T with maximal overlap over the time interval I. Note that a period of the order of 10 s of EEG data is assumed to reflect almost stationary behavior of the electrical brain activity (for example, cf. Ref. [32]). Then, I is moved with a step width of 1 s along the data set; for each step calculating the nearest-neighbor distribution of the 18 distances after using Eq. (12) for the individual unfolding. Comparison with the null hypothesis is performed employing a KS test. The results are shown in Fig. 5(b) for the distances between the eight small-

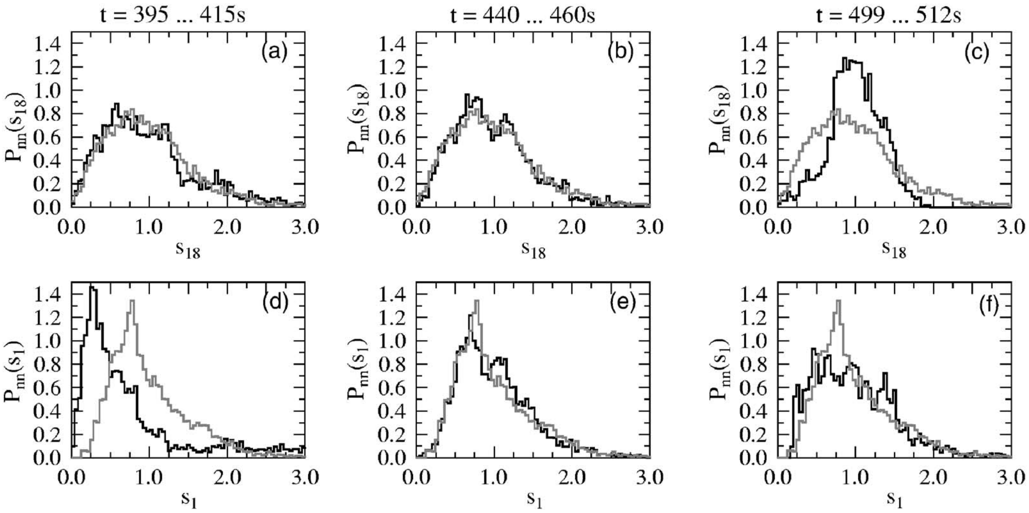

FIG. 6. Nearest-neighbor distributions obtained by individual unfolding for the EEG recording shown in Fig. 5 for distance s_{18} [panels (a), (b), and (c)] and distance s_1 [panels (d), (e), and (f)]. The null hypothesis is displayed as a gray line. The histograms drawn with a black line present an average over the period 395 to 415 s [panels (a) and (d)], which includes a muscle artifact, an artifact-free period between 440 and 460 s [panels (b) and (e)], and during the seizure period betwen 499 and 512 s [panels (c) and (f)].

est eigenvalues and Fig. 5(c) for those between the 12 largest eigenvalues. Note the different scales of the logarithmic representation.

The nearest-neighbor distribution of the largest eigenvalues [Fig. 5(c)], reflecting the overall behavior of the system, is close to the null hypothesis during almost the whole time course. The KS test gives a result close to 100% probability for coincidence. Only when the interval I starts to overlap with the period of epileptic activity (indicated by the vertical dashed line) drastic deviations occur as expected because of the change of the global correlation structure during the seizure. In contrast, the dynamics of the smallest eigenvalues turns out to be more sensitive for the detection of subtle changes of the brain dynamics [Fig. 5(b)]. During the artifact at 405 s the distribution of the distances between a number of smallest eigenvalues show large deviations from the corresponding distributions calculated for the reference interval. A similar behavior can be observed for the small eye movement artifact at about 485 s. Although the effect is not as pronounced as in the previous case, the nearest-neighbor distributions measured at the lower edge of the spectrum indicate nicely the temporal position of the artifact.

In general, artifacts caused by swallowing, eye movement, etc. go along with correlations measured predominantly by only a subset of all electrodes. Consequently, such an event is more comparable to situations discussed in Figs. 1 and 2, i.e., where only a small fraction of the signals correlate. As shown by means of simulated systems, this information can be measured more sensitively at the lower part of the spectrum. Hence, the nearest-neighbor distributions for the s_i between large eigenvalues are close to the null hypothesis during almost the whole time course of the EEG and react drastically only during the seizure, whereas the distributions measured for small eigenvalues show significant deviations for more subtle changes of the correlation structure.

In Fig. 6 some examples of nearest-neighbor distributions, calculated in different segments of the EEG are presented in comparison with the null hypothesis. The first row shows the results for the distance s_{18} between the two largest eigenvalues and the second row the same for s_1. The first column displays the behavior during the artifact at 405 s, the second one during an artifact-free period and the third column reflects the features of the epileptic seizure. The distribution for s_{18} is close to its null hypothesis during almost the whole time course. Even the artifact at 405 s does not significantly alter the distribution. This picture is changed when the epileptic activity initiates. As the generalized seizure presents a highly collective dynamical state, it is not surprising that it is clearly reflected in the dynamics of the largest eigenstates, which measure the global collective behavior of the whole system. During the seizure period the distribution of s_{18} is considerably narrower than the null hypothesis and shows a pronounced peak at $s_{18}=1$. This fact indicates the expected increase of correlations in the entire system. On the contrary, the distribution of s_1 shows its most prominent deviation during the artifact. Not even the epileptic activity causes major changes in the distribution of s_1.

VI. DISCUSSION AND CONCLUSIONS

While the nondiagonal elements C_{ij} of the correlation matrix have a direct interpretation as cross-correlation coefficients between signal X_i and X_j (disregarding the influence of the random correlations), the eigenvalues are in general a complicated mixture of all these coefficients and their interpretation is at first not obvious. In Ref. [17] it was shown that changes of the nonrandom cross correlations between the different time series of a multivariate dataset imply nonrandom level repulsions between the eigenvalues of $\mathbf{C}$. Which eigenstates repel depends on the strength and type of correlations and the number of correlated time series. In this way, information about the correlation structure, or its dynamics, is induced in comparably small sections of the whole spectrum—occasionally between two states only. Hence, employing a spectral average during the unfolding transformation, this information is washed out or might even be lost completely as discussed in Fig. 4. For this reason we proposed an unfolding by simply normalizing the distances by its ensemble average. Then the nearest-neighbor distribution can be calculated for each of the distances individually. This technique permits us to detect precisely in which part of the spectrum correlations are induced.

With the help of simulated systems like N_f-tori or Rössler systems we gave clear numerical evidence for these statements. The nonrandom repulsion between eigenvalues increases the stiffness of the distances, consequently leading to significant deviations from the universal properties of the GOE for the nearest-neighbor distribution. In the examples discussed in this paper, especially the sharp cut between those parts of the spectrum influenced by noise and random correlation and those parts that carry genuine information about the correlations is remarkable.

We could prove that in cases where only a small fraction of all signals $X_i, i=1,\ldots,M$ (de)correlate, this information is implemented with higher significance at the lower part of the spectrum. This fact could be explained by the collective character of the large eigenstates, which carry components not only from the correlated subspace, but also from the random part [17].

Finally, we could prove the applicability of the method by analyzing nonstationary experimental data using a sliding window in which the empirical ensembles are created. Here, an average over the data itself is used as a kind of null hypothesis, instead of comparing results with the GOE. The reason for that is that in general the correlations of EEG data measured from a person at resting state are not of GOE type. This is true not only because of correlations caused by the electrical brain activity but also because of possible stationary correlations induced by the choice of the EEG reference.

By using a KS test when measuring the deviation from our null hypothesis, we could detect nonstationary dynamics like artifacts in the behavior of the lowest eigenvalues, while such events are almost invisible in the upper half of the spectrum. The deviations from the null hypothesis for the level statistics at the lower part of the spectrum we attribute to the fact that artifacts cause correlations between only a fraction of all signals. As discussed in Figs. 1 and 2, these more subtle changes in the global correlation structure of the whole data set can be measured with higher sensitivity and better statistical significance in the changes of the smallest eigenvalues of **C**. Due to the collective character of the largest λ, these tiny features are hidden within the overall behavior of the system.

The real-world example illustrates in a clear manner that the way of individual unfolding and calculating RMT measures such as the nearest-neighbor distribution individually for each of the level distances, can give detailed insight into the characteristics of the correlation structure of a multivariate data set. We believe that the technique presented in this paper is highly promising even in cases where nonstationary systems are in consideration and one is unable to create large matrix ensembles.

ACKNOWLEDGMENTS

This work was supported by CONACyT, Mexico (Project No. 40885-F) and by the Deutsche Forschungsgemeinschaft (DFG) via Grant No. RU 1401/1-1.

[1] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Phys. Rev. Lett. **83**, 1467 (1999).

[2] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, and H. E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).

[3] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, and H. E. Stanley, Physica A **287**, 374 (2000).

[4] S. Drozdz, F. Gruemmer, F. Ruf, and J. Speth, Physica A **287**, 440 (2000); **294**, 226 (2001); S. Drozdz, J. Kwapien, F. Gruemmer, F. Ruf, and J. Speth, *ibid.* **299**, 144 (2001); J. Kwapien, S. Drozdz, and J. Speth, *ibid.* **330**, 605 (2003).

[5] V. Plerou, P. Gopikrishnan, B. Rosenow, Luis A. Nunes Amaral, T. Guhr, and H. E. Stanley, Phys. Rev. E **65**, 066126 (2002).

[6] A. Utsugi, K. Ino, and M. Oshikawa, Phys. Rev. E **70**, 026110 (2004).

[7] P. Seba, Phys. Rev. Lett. **91**, 198104 (2003).

[8] J. Kwapien, S. Drozdz, and A. A. Ioannides, Phys. Rev. E **62**, 5557 (2000).

[9] M. S. Santhanam and P. K. Patra, Phys. Rev. E **64**, 016102 (2001).

[10] M. Barthélemy, B. Gondran, and E. Guichard, Phys. Rev. E **66**, 056110 (2002).

[11] M. L. Mehta, *Random Matrices* (Academic Press, New York, 1990).

[12] T. Guhr, A. Müller-Groeling, and H. A. Weidenmüller, Phys. Rep. **299**, 189 (1998).

[13] I. T. Jolliffe, *Principal Component Analysis* (Springer-Verlag, Berlin, 1986); M. Kendall, *Multivariate Analysis* (Charles Griffin & Co., London, 1975).

[14] Y. Malevergne and D. Sornette, Physica A **331**, 660 (2004).

[15] Z. Burda and J. Jurkiewicz, Physica A **344**, 67 (2004).

[16] J. Kwapień, S. Drozdz, and P. Oświecimka, Physica A **359**, 589 (2006).

[17] M. Müller, G. Baier, A. Galka, U. Stephani, and H. Muhle, Phys. Rev. E **71**, 046116 (2005).

[18] R. J. Muirhead, *Aspects of Multivariate Statistical Theory* (Wiley, New York, 1982).

[19] F. Haake, in *Quantum Signatures of Chaos*, Springer Series in Synergetics Vol. 54, 2nd ed. (Springer, Berlin, 2001).

[20] J. Flores, M. Horoi, M. Müller, and T. H. Seligman, Phys. Rev. E **63**, 026204 (2000).

[21] A. M. Sengupta and P. P. Mitra, Phys. Rev. E **60**, 3389 (1999).

[22] M. J. Bowick and E. Brézin, Phys. Lett. B **268**, 21 (1991); J. Feinberg and A. Zee, J. Stat. Phys. **87**, 473 (1997).

[23] A. Galka and G. Pfister, Int. J. Bifurcation Chaos Appl. Sci. Eng. **13**, 723 (2003).

[24] M. G. Rosenblum *et al.*, Phys. Rev. Lett. **89**, 264102 (2002).

[25] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flan-

nery, *Numerical Recipes in C, The Art of Scientific Computing*, 2nd ed. (Cambridge University Press, Cambridge, England, 1992).

[26] W. Iskra, M. Müller, and I. Rotter, J. Phys. G **19**, 2045 (1993).

[27] W. Iskra, M. Müller, and I. Rotter, J. Phys. G **20**, 775 (1994).

[28] C. Jung, M. Müller, and I. Rotter, Phys. Rev. E **60**, 114 (1999).

[29] E. Persson and I. Rotter, Phys. Rev. C **59**, 164 (1999).

[30] B. Hjorth, Am. J. EEG Technol. **20**, 121 (1980).

[31] C. Rummel, M. Müller, and P. Valdez (unpublished).

[32] C. Rieke, R. G. Andrzejak, F. Mormann, and K. Lehnertz, Phys. Rev. E **69**, 046111 (2004); C. Rieke, R. G. Andrzejak, F. Mormann, T. Kreuz, P. David, C. E. Elger, and K. Lehnertz, IEEE Trans. Biomed. Eng. **50**, 634 (2003).