Correlation dimension of high-dimensional and high-definition experimental time series

Cite as: Chaos 33, 123114 (2023); doi: 10.1063/5.0168400
Submitted: 18 July 2023 · Accepted: 13 November 2023 ·
Published Online: 11 December 2023

Valeri A. Makarov,[1,a] Ricardo Muñoz,[1,2] Oscar Herreras,[2] and Julia Makarova[2,b]

AFFILIATIONS

[1] Department of Applied Mathematics and Mathematical Analysis, Universidad Complutense de Madrid, Plaza de las Ciencias 3, Madrid 28040, Spain
[2] Department of Translational Neuroscience, Cajal Institute, CSIC, Av. Doctor Arce 37, Madrid 28002, Spain

Note: This paper is part of the Focus Issue: Data-Driven Models and Analysis of Complex Systems.
[a] **Author to whom correspondence should be addressed:** vmakarov@ucm.es. **URL:** http://blogs.mat.ucm.es/vmakarov.
[b] **Electronic mail:** julia.samuseva@cajal.csic.es

ABSTRACT

The correlation dimension (CD) is a nonlinear measure of the complexity of invariant sets. First introduced for describing low-dimensional chaotic attractors, it has been later extended to the analysis of experimental electroencephalographic (EEG), magnetoencephalographic (MEG), and local field potential (LFP) recordings. However, its direct application to high-dimensional (dozens of signals) and high-definition (kHz sampling rate) 2HD data revealed a controversy in the results. We show that the need for an exponentially long data sample is the main difficulty in dealing with 2HD data. Then, we provide a novel method for estimating CD that enables orders of magnitude reduction of the required sample size. The approach decomposes raw data into statistically independent components and estimates the CD for each of them separately. In addition, the method allows ongoing insights into the interplay between the complexity of the contributing components, which can be related to different anatomical pathways and brain regions. The latter opens new approaches to a deeper interpretation of experimental data. Finally, we illustrate the method with synthetic data and LFPs recorded in the hippocampus of a rat.

Modern EEG, MEG, or local field potential (LFP) recordings provide high-dimensional and high-definition (2HD) data. Their nonlinear analysis is an attractive but long-standing challenging problem. On the one hand, it requires long samples, but on the other, the data stationarity and availability limit the potential of modern methods, especially in biology and medicine. Here, we study the correlation dimension (CD) that captures the intrinsic complexity of time series and is used, e.g., for describing brain states or detecting seizures. Yet, there is a controversy in the results obtained in medical applications. We discuss the possible reasons for the observed discrepancy and provide a novel method for estimating CD from 2HD data. The method uses a linear approach to preprocessing data, recently shown to help deal with LFPs, and divides the problem into smaller pieces, tractable by the nonlinear CD analysis. Such an approach enables the reduction of the required data sample size by orders of magnitude and, hence, improves the time resolution. It opens new opportunities for a deeper understanding of information processing in the brain.

I. INTRODUCTION

The correlation dimension (CD) is a nonlinear measure of the data complexity.[1] It is, in particular, helpful in detecting qualitative changes in the brain dynamics, localizing regions causing abnormal oscillations and even predicting seizures (see, e.g., Refs. 2–4). Although the CD is routinely used in different applications, the results are controversial. For example, one can argue that the CD must decrease during a seizure or preictal period due to excessive synchronization of brain waves.[5] However, opposite results have also been reported,[3,6] which led to the conclusion that the CD measure is insufficient for clinical application.[3,7]

The disparity observed in the literature can be ascribed to the amount of data used to estimate the CD and the data stationarity. Modern experimental techniques, such as electroencephalographic (EEG), magnetoencephalographic (MEG), or local field potential (LFP) recordings, enable massive acquisition of high-dimensional and high-definition (2HD) data. Although it could be deemed a panacea, the analysis of 2HD signals is a challenging problem, in

particular, due to the curse of dimensionality coined by Bellman.[8] The increasing data dimension requires an exponential growth of samples for statistically significant assessment of different measures, which frequently contradicts the data stationarity and availability, especially in medicine. Moreover, it challenges numerical methods and available computational power. Thus, reducing the number of samples required to reach the statistical significance of the CD is a critical issue.

The CD describes the complexity of invariant sets and is closely related to the topological dimension. Loosely speaking, the latter is the number of independent variables required to define a neighborhood of a point in a set. For instance, any curve has a topological dimension equal to one, whereas a surface is a two-dimensional (2D) object. However, sometimes, such a definition can produce misunderstanding. For example, the topological dimension of a Koch snowflake is one, but the length of the "curve" between any two points on it is infinite. In some sense, it is "too big" to be a one-dimensional object. The box-counting[9] and correlation dimensions solve this issue. The fractal dimension of a Koch snowflake is $\log(4)/\log(3) \approx 1.26$. The advantage of the CD over the box-counting method is its relative simplicity of calculations.[10,11] Moreover, in contrast to the box-counting, the CD takes into account not only the geometrical properties of a set but also the frequency with which a typical trajectory visits different regions of the set (for details, see, e.g., Ref. 12). This makes it better suited for analysis of experimental data.

To illustrate the problem of estimating the CD from 2HD data, we have built semi-synthetic LFPs [Fig. 1(a)] with the known CD of $\nu^* \approx 3.89$ (see below for details). Figure 1(b), blue curve, shows the CD evaluated by a standard method over data samples of different sizes (see below for details). The estimation exhibits a strong negative bias at small sample sizes, e.g., providing $\nu \approx 2.08$ for a sample of 10^3 points, i.e., 46% of the relative error. The CD estimate was reasonable only for long data sets (above 5×10^5 points). Note that the LFPs had a relatively low CD in this example, facilitating the calculations. Real 2HD data can have much higher dimensions, making the CD estimation from raw LFPs even worse.

The difficulties in a reliable estimation of the CD have been known for a long time, which stimulated a discussion on the sufficient conditions for mapping original data sets to Euclidean spaces of different dimensions (see, e.g., Ref. 13), and a search for improved measures (see, e.g., Refs. 11 and 14). However, such measures have been tested on low-dimensional data (usually chaotic attractors) and have limited application to EEG, MEG, or LFPs, where the number of data channels can reach hundreds. In such cases, the phase space dimension becomes a prohibitive factor for direct calculations.

Nevertheless, the analysis of 2HD data can benefit from statistical pre-processing, such as independent component analysis of LFPs, introduced in Refs. 15 and 16. It has been shown to be exceptionally useful for identifying neuronal sources and disentanglement of LFPs (see, for a review, Ref. 17).

Here, we introduce a novel method for estimating the CD from 2HD experimental data and provide its mathematical justification. The method reduces the data size required to a confident estimate of the CD by several orders of magnitude. It separates raw recordings in independent, relatively low-dimensional invariant sets and then separately estimates each set's CD. Finally, as proved below, the CD of the original data is a sum of the obtained CDs. With the data shown in Fig. 1(a), the method provides an accurate estimate of the CD (<5% of the relative error) with no more than 3×10^3 samples [Fig. 1(b), red curve]. Such an improvement can foster CD use in medical and biological applications.

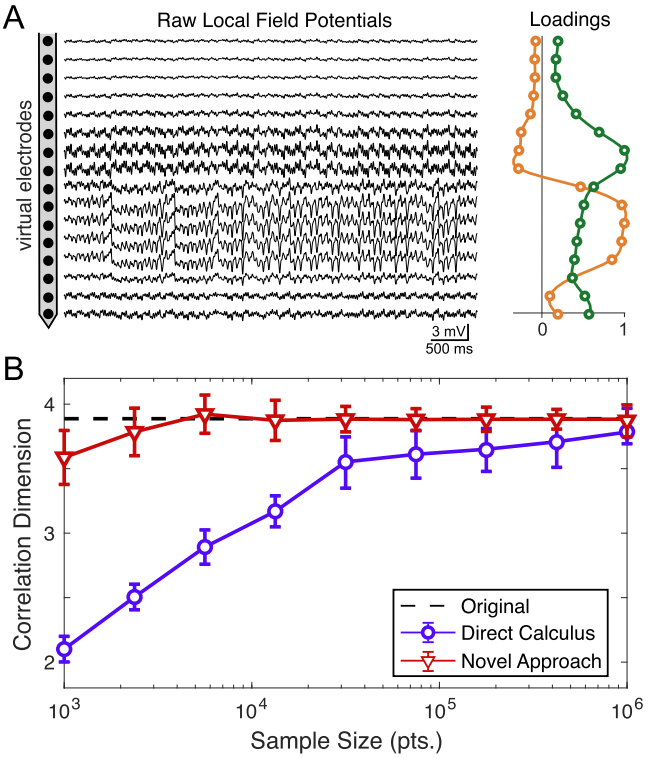

FIG. 1. The problem of estimating the correlation dimension from raw LFPs. (a) An epoch of semi-synthetic LFPs. The raw LFPs (left) are built from the double scroll and Lorenz attractors using experimental spatial loadings (right). (b) The correlation dimension estimated from raw LFPs (blue) and using the proposed approach (red) vs the data sample lengths. The black dashed line marks the theoretical value.

II. ESTIMATING THE CORRELATION DIMENSION OF 2HD DATA

A. Problem formulation

1. Standard approach

Consider a 2HD signal $x(t) \in \mathbb{R}^n$, where $t \in \mathbb{N}$ is the discrete time, and n is the number of recording electrodes [e.g., in Fig. 1(a), $n = 16$]. In practice, we have access to a data matrix $X = (x(t))_{t=1}^{L}$, where $L \gg 1$ is the sample size.

Thus, given a 2HD data matrix X, we want to estimate the correlation dimension ν_x of the invariant set generating $x(t)$. Note that although being high-dimensional ($n \gg 1$), $x(t)$ consists of measurable variables only and, hence, may not provide complete information on the latent space generating $x(t)$. Thus, the dimension of the space embedding $x(t)$ could be much higher than n. As mentioned

in Sec. I, the standard approach to this problem (see below) leads to a significant bias [Fig. 1(b)].

2. Linear mixture model

In applications such as MEG, EEG, or LFPs, some spatiotemporal sources are mixed in space and time, and such a mixture is usually linear or can be assumed linear.[18,19] For example, in LFP recordings, sources are the transmembrane currents elicited in spatially ordered neurons upon activating groups of converging axons (for more details, see the review,[20] and references therein). These currents produce field potentials that linearly mix in the volume conductor. The resulting field potential φ is given by[19]

$$\nabla \cdot (\sigma \nabla \varphi) = -J, \tag{1}$$

where σ is the electrical conductivity of the brain tissue and J is the current density. We note that φ and J are mesoscopic variables built by conglomerates of fragments of individual currents in a myriad of neurons. Moreover, each source globally maintains a stable nonlinear geometry depending on different anatomical and functional factors. Equation (1) establishes a linear mixture of sources, i.e., if $\varphi_{1,2}$ are solutions for two sources $J_{1,2}$, then $\varphi_1 + \varphi_2$ is a solution for $J_1 + J_2$. Although the mixture is linear, each source can have a nonlinear spatial distribution and complex time dynamics.

Thus, we can assume that the signal $x(t)$ is a mixture of m sources whose temporal evolution is given by $s(t) \in \mathbb{R}^m$, such that $m \leq n$. For example, for LFP recordings in the hippocampus, m is usually between five and seven, depending on the experimental conditions.[15–17] Therefore, we have the data model,

$$x(t) = Ws(t), \tag{2}$$

where $W = (w_1, \ldots, w_m) \in \mathbb{R}^{n\times m}$ is the mixing matrix defined by the brain area's anatomy, and the vector $w_i \in \mathbb{R}^n$ $(i = 1, 2, \ldots, m)$ represents the loading (spatial weights) the ith source contributes to the mixture at each electrode [Fig. 1(a), left, shows two loadings]. We note that in experimental conditions $w_i \neq 0$ and $w_i \neq w_j$ $(i \neq j)$, i.e., W is of full rank, $\mathrm{rank}(W) = m$. Moreover, neither W nor $s(t)$ is known.

Although model (2) is quite general, it enables an efficient method for estimating the CD, as shown below.

B. Standard approach to CD estimation

Theoretically, the CD of a 2HD recording can be estimated directly using the embedding approach (see, for details, Refs. 13, 21, and 22), although such an estimation can be significantly biased [Fig. 1(b)].

The standard approach goes through two steps:

1. Reconstruct the state space from the time series.
2. Estimate the CD of the reconstructed set by the Grassberger–Procaccia method.

However, due to a high CD of experimental 2HD data, such an approach requires an exponentially high amount of data [Fig. 1(b)], usually unavailable, and besides quickly saturates modern computers (see, e.g., Ref. 23 for discussion).

1. State space reconstruction

The state space reconstruction is based on the Takens embedding.[22] Using the original time series $x(t) \in \mathbb{R}^n$ $(n \geq 1)$ taken d times with increasing delays, we build a new $(n \times d)$-dimensional vector,

$$y(t) = \left(x^T(t), x^T(t-\tau), \ldots, x^T(t-(d-1)\tau)\right)^T \in \mathbb{R}^{nd}, \tag{3}$$

which represents the invariant set in the state space. Given certain conditions, the reconstructed manifold is, in the sense of diffeomorphism, equivalent to the original one. Thus, the reconstruction preserves the correlation dimension.

The reconstruction (3) depends on two parameters: the embedding delay $\tau \in \mathbb{N}$ and dimension $d \in \mathbb{N}$. To choose τ, we use the auto-mutual information (AMI).[24] Assuming $x(t)$ is scalar, the AMI is

$$I(x(t), x(t+\tau)) = \sum_{i,j} p_{ij}(\tau) \log\left(\frac{p_{ij}(\tau)}{p_i p_j}\right), \tag{4}$$

where p_i is the probability that $x(t)$ is in bin i of the histogram constructed from $x(t)$, and $p_{ij}(\tau)$ is the probability that $x(t)$ is in bin i and $x(t+\tau)$ is in bin j. The optimal time lag corresponds to the first minimum of the AMI, where $x(t)$ and $x(t+\tau)$ are the most independent possible. For an nD time series, the AMI function (4) and the optimal delay are evaluated component-wise. Then, the best delay is set to the average of the individual delays.[25]

The best minimal dimension d is selected by the false nearest neighbor (FNN) analysis.[26] Let denote $\tilde{y}(t; d)$ as the first nearest neighbor of $y(t; d)$ for the d-dimensional embedding [Eq. (3)]. Then, the squared Euclidean distance between these points is $R_d^2(t) = \|y(t; d) - \tilde{y}(t; d)\|_2^2$. We now calculate the same distance but in the $(d+1)$-dimensional embedding $R_{d+1}^2(t)$ and the absolute difference $\Delta R_d^2 = |R_{d+1}^2(t) - R_d^2(t)|$. A pair of points is called FNN if it satisfies either of the following conditions:

$$\frac{\Delta R_d^2(t)}{R_d^2(t)} > R_{\mathrm{tol}}^2, \quad \frac{\Delta R_d^2(t)}{\mathrm{tr}(C_x)} > A_{\mathrm{tol}}, \tag{5}$$

where R_{tol}^2 and A_{tol} are thresholds (set to $R_{\mathrm{tol}}^2 = 10$ and $A_{\mathrm{tol}} = 2$ in numerics[26]), and C_x is the covariance matrix of the data. The best minimal dimension is set to the value when the ratio of FNNs falls below a threshold.

2. CD estimation

Once the state space has been reconstructed, we apply the Grassberger–Procaccia method.[1] The correlation integral for a compact set $\mathscr{X}$ is given by

$$C(r) = \int p(x)p(x')H(r - \|x - x'\|)\,\mathrm{d}x\mathrm{d}x', \tag{6}$$

where H is the Heaviside function and r is the box size. In practice, it can be estimated from a trajectory (time series) $x(t)$ $(t = 1, 2, \ldots, L)$ asymptotic to $\mathscr{X}$,

$$C(L, r) = \frac{2}{L(L-1)} \sum_{1\leq t<s\leq L} H\left(r - \|x(t) - x(s)\|\right). \tag{7}$$

Then, $C(r) = \lim_{L\to\infty} C(L, r)$.

Assuming that $C(r) = \phi(r)r^{\nu}$, where ϕ, in general, is a function of r that tends to a constant or oscillates around a constant as $r \to 0$ (see Ref. 14 for details), we define the correlation dimension,

$$\nu = \lim_{r\to 0} \frac{\log C(r)}{\log r}. \tag{8}$$

In practice, the CD is obtained by fitting a straight line to the plot of $\log C(L, r)$ vs $\log r$. We used this procedure to estimate the CD of raw LFPs (Fig. 1).

C. Theoretical foundations of the novel method

Now, let us provide a theoretical justification of the novel method used later for building a numerical algorithm.

1. CD is invariant under linear mapping

Assume that an invariant set $\mathscr{X} \subset \mathbb{R}^k$ is linearly mapped to set $\mathscr{Y} \subset \mathbb{R}^l$ in a higher dimensional space ($l \geq k$),

$$\begin{aligned} &\mathscr{L}: \mathscr{X} \to \mathscr{Y}, \\ &x \mapsto y = Mx, \end{aligned} \tag{9}$$

where $M \in \mathbb{R}^{l\times k}$ is a full rank matrix. Let ν_x and ν_y be the CDs of $\mathscr{X}$ and $\mathscr{Y}$, respectively. Then, we have the following result.

Lemma 1. *Under conditions provided by Eq. (9), the correlation dimensions of the set $\mathscr{X}$ and its image $\mathscr{Y}$ are equal $\nu_y = \nu_x$.*

Proof. Since M is of full rank, its pseudoinverse is $A = (M^TM)^{-1}M^T$, which is also a full rank matrix. Thus, besides $y = Mx$ mapping linearly points from $\mathscr{X}$ to $\mathscr{Y}$, we have the inverse linear mapping: $y \mapsto x = Ay$. Therefore, the set $\mathscr{Y}$is homeomorphic to $\mathscr{X}$, and, hence, $\nu_x = \nu_y$.

Indeed, for a pair of points in $\mathscr{Y}$, we have

$$\|y - y'\| = \|M(x - x')\| \leq \|M\|\|x - x'\|, \tag{10}$$

where $\|\cdot\|$ are standard vector and vector-induced matrix norms. Extending (10) to the inverse map, we get

$$\frac{1}{\|A\|}\|x - x'\| \leq \|y - y'\| \leq \|M\|\|x - x'\|. \tag{11}$$

The latter inequality in (11) gives rise to

$$\begin{aligned} C_{\mathscr{Y}}(r) &= \int p(y)p(y')H(r - \|y - y'\|)\,\mathrm{d}y\mathrm{d}y' \\ &\geq \int p(x)p(x')H(r - \|M\|\|x - x'\|)\,\mathrm{d}x\mathrm{d}x' \\ &= C_{\mathscr{X}}\left(\frac{r}{\|M\|}\right). \end{aligned} \tag{12}$$

Repeating the same arguments for the first inequality in (11), we get

$$C_{\mathscr{X}}(\|A\|r) \leq C_{\mathscr{Y}}(r) \leq C_{\mathscr{X}}\left(\frac{r}{\|M\|}\right). \tag{13}$$

Using the CD definition (8), we have

$$\begin{aligned} \lim_{r\to 0} \frac{\log C_{\mathscr{X}}(\|A\|r)}{\log r} &= \lim_{r\to 0} \frac{\log C_{\mathscr{X}}(\|A\|r)}{\log(\|A\|r)} \\ &= \lim_{r\to 0} \frac{\log C_{\mathscr{X}}(\frac{r}{\|M\|})}{\log \frac{r}{\|M\|}} = \nu_x. \end{aligned} \tag{14}$$

Finally, $\nu_x \leq \nu_y \leq \nu_x$, which ends the proof. □

2. CD of Cartesian products

Let $\mathscr{X}_1 \subset \mathbb{R}^{k_1}$ and $\mathscr{X}_2 \subset \mathbb{R}^{k_2}$ be two invariant sets with the correlation dimensions ν_1 and ν_2, respectively. We now construct their Cartesian product,

$$\mathscr{X} := \mathscr{X}_1 \times \mathscr{X}_2. \tag{15}$$

Then, we have the following result.

Lemma 2. *Assume we take points from the Cartesian product (15) $x = (x_1, x_2) \in \mathscr{X}$ statistically independently, i.e., $p(x_1, x_2) = p(x_1)p(x_2)$. Then, the correlation dimension of the compound set is $\nu = \nu_1 + \nu_2$.*

Proof. For convenience, we use maximum norm $\|\cdot\| := \|\cdot\|_\infty$ in the correlation integral (6). We now note that

$$H(r - \|x\|) = H(r - \|x_1\|)H(r - \|x_2\|). \tag{16}$$

Then, using the statistical independence, we get

$$\begin{aligned} C_{\mathscr{X}}(r) = &\int p(x_1)p(x_1')H(r - \|x_1 - x_1'\|)\,\mathrm{d}x_1\mathrm{d}x_1' \\ &\times \int p(x_2)p(x_2')H(r - \|x_2 - x_2'\|)\,\mathrm{d}x_2\mathrm{d}x_2' = C_{\mathscr{X}_1}(r)C_{\mathscr{X}_2}(r). \end{aligned} \tag{17}$$

Thus, the correlation integral is factorized, and the CDs are summed. □

Corollary 1. *The product set (15) and Lemma 2 can be extended into general Cartesian products $\mathscr{X} := \mathscr{X}_1 \times \cdots \times \mathscr{X}_m$. Then, $\nu = \sum_{i=1}^m \nu_i$.*

3. CD of the decomposition into independent components

Let $\mathscr{S}_i \subset \mathbb{R}^{k_i}$ ($i = 1, \ldots, m$, $k_i \in \mathbb{N}$) be compact sets. We now build a global set $\mathscr{X}$by the linear mapping,

$$\begin{aligned} &\mathscr{L}: \mathscr{S}_1 \times \cdots \times \mathscr{S}_m \to \mathscr{X} \subset \mathbb{R}^N, \\ &(s_1, \ldots, s_m) \mapsto x = W\left(s_1^T, \ldots, s_m^T\right)^T, \end{aligned} \tag{18}$$

where $W = (w_1, \ldots, w_m) \in \mathbb{R}^{N\times M}$ is the global mixing matrix composed of matrices for individual components $w_i \in \mathbb{R}^{N\times k_i}$ such that $M = \sum_{i=1}^m k_i \leq N$ and $\mathrm{rank}(W) = M$. Then, we have the following result.

Theorem 1. *Consider model (18). Let ν_i, $i = 1, 2, \ldots, m$ be the correlation dimensions of the invariant sets $\{\mathscr{S}_i\}$. Given that data are sampled independently from the sets $\{\mathscr{S}_i\}$, the correlation dimension of the set $\mathscr{X}$ is*

$$\nu_{\mathscr{X}} = \sum_{i=1}^m \nu_i. \tag{19}$$

Proof. Using Lemma 1, we deduce that the correlation dimensions of the sets $\mathscr{X}$ and of the Cartesian product $\mathscr{S}_1 \times \cdots \times \mathscr{S}_m$ are equal. Then, using Lemma 2, we conclude that the Cartesian product has the dimension $\sum_{i=1}^{m} \nu_i$. □

D. Efficient algorithm for estimating CD of 2HD data

Let $X \in \mathbb{R}^{n\times L}$ be an experimental data matrix. Without loss of generality, we suppose that the recording microelectrode array is linear. Then, the data model (2) can be written in the form

$$X = WS, \tag{20}$$

where $W \in \mathbb{R}^{n\times m}$ is the matrix of (nonlinear) spatial loadings, $S = (s_1, \ldots, s_m)^T \in \mathbb{R}^{m\times L}$ is the activation matrix. Then, Theorem 1 says that the correlation dimension of the observable X is the sum of the correlation dimensions of the components of S.

Thus, to find the correlation dimension, we go through the following steps:

1. Apply the independent component analysis to X [see Eq. (20)] and find the mixing matrix W and the matrix of activations S.
2. For each component s_i of the matrix S reconstruct the invariant set in the phase space by using Taken's embedding $y_i(t) = (s_i(t), s_i(t-\tau), \ldots, s_i(t-(d-1)\tau)) \in \mathbb{R}^d$. Find the optimal time delay $\tau \in \mathbb{N}$ by minimizing the AMI and the embedding dimension $d \in \mathbb{N}$ by the FNN method (see Sec. II B 1).
3. Apply the Grassberger–Procaccia algorithm to each reconstructed invariant set $\{y_i(t)\}_{t=1}^{L}$ $(i = 1, \ldots, m)$ and estimate the CD $\nu_i(y_i)$ (see Sec. II B 2).
4. Evaluate the correlation dimension of the original 2HD data,

$$\nu(X) = \sum_{i=1}^{m} \nu_i(y_i).$$

In the literature, there are a number of algorithms to perform ICA (for review, see Ref. 27). At step 1, we use the kernel-density algorithm[28] adapted to LFPs (for more details, see the review[29]) and implemented in the ICAofLFPs package running in MATLAB (available at http://blogs.mat.ucm.es/vmakarov/downloads/).

E. Short-term CD

Experimental data usually have a degree of non-stationarity, e.g., due to changes in the brain state. The experimentalist may then want to assess how the CD changes over time.

The standard approach in such cases is to evaluate a measure over a small enough time interval T and repeat the calculation by sliding the interval over time. However, there is a trade-off between better statistics for larger T and better data stationarity for smaller T. In addition, each measure sets a minimal limit on T. The method proposed here reduces the requirement on the sample size by orders of magnitude [Fig. 1(b)], thus significantly improving the time resolution (i.e., the sensitivity to changes in the CD of LFPs).

Let $X = (x_{ij})_{i=1,j=1}^{n,L}$ be the data matrix. We then define a set of shorter matrices $X_t = (x_{ij})_{i=1,j=t}^{n,t+T}$, where T is the length of the sliding window and t is the "current" time (one can shift it by $T/2$ for better centering). Then, the CD is evaluated over all matrices X_t separately, and we get the time-dependent measure $\nu(t)$.

III. NUMERICAL ASSESSMENT OF THE METHOD

We illustrate the method on synthetic LFPs and electrophysiological data.

A. Synthetic and semi-synthetic LFPs

We simulate LFPs by mixing several generators. Each LFP generator is composed of a time course (activation) and a nonlinear spatial loading.

1. Time courses

We generate time courses by employing three chaotic attractors: Lorenz, Double scroll, and Rossler, and also a quasi-periodic orbit embedded into a 3D space.

The Lorenz dynamical system is given by[30]

$$\begin{aligned} \dot{x}_1 &= \sigma(x_2 - x_1), \\ \dot{x}_2 &= r x_1 - x_2 - x_1 x_3, \\ \dot{x}_3 &= x_1 x_2 - b_L x_3, \end{aligned} \tag{21}$$

where we used typical parameter values: $\sigma = 10$, $r = 28$, and $b_L = 8/3$.

The Rossler dynamical system is given by[30]

$$\begin{aligned} \dot{x}_1 &= -x_2 - x_3, \\ \dot{x}_2 &= x_1 + a_R x_2, \\ \dot{x}_3 &= b_R + x_3(x_1 - c), \end{aligned} \tag{22}$$

with $a_R = 0.2$, $b_R = 0.2$, and $c = 5.7$.

The double scroll dynamical system is given by[31]

$$\begin{aligned} \dot{x}_1 &= a_{DS}(x_2 - \phi(x_1)), \\ \dot{x}_2 &= x_1 - x_2 + x_3, \\ \dot{x}_3 &= -b_{DS} x_2, \end{aligned} \tag{23}$$

where $\phi(x) = (1 + m_1)x + \frac{1}{2}(m_0 - m_1)(|x+1| - |x-1|)$, with $a_{DS} = 15.6$, $m_0 = -8/7$, $m_1 = -5/7$, and $b_{DS} = 27$.

The quasi-periodic time series is given by

$$\begin{aligned} x_1(t) &= (A + \cos(2\pi\omega t))\cos(8\pi t), \\ x_2(t) &= (A + \cos(2\pi\omega t))\sin(8\pi t), \\ x_3(t) &= \sin(8\pi t), \end{aligned} \tag{24}$$

with $\omega = \frac{1}{\sqrt{2}}$ and $A = 3$. It represents a trajectory that densely covers a torus.

The correlation dimensions for these attractors are $\nu_L = 2.044$, $\nu_R = 1.877$, and $\nu_{DS} = 1.829$, which agree with the data provided in the literature (see, e.g., Ref. 11). The CD of the quasi-periodic time series is $\nu_{QP} = 2$.

Now, let $(x(t))_{t=1}^{L} \in \mathbb{R}^{3\times L}$ be a matrix representing an invariant set obtained by numerical integration of one of the systems (21)–(23) or simulation of (24), with the mean $\bar{x} = \frac{1}{L}\sum_{t=1}^{L} x(t)$ and variance $\sigma_x^2 = \frac{1}{L}\sum_{t=1}^{L} \|x(t) - \bar{x}\|^2$. The generator's activation $s(t) \in \mathbb{R}$ and, hence, not all information of the latent variables $x(t) \in \mathbb{R}^3$ is accessible "experimentally," i.e., the original 3D latent space must

be mapped into 1D observation space. To model such a mapping, we generate a random vector $B \in \mathbb{R}^3$ simulating the observation procedure. Then, the generator's time course is

$$s(t) = \frac{B^T}{\sigma_x \|B\|}(x(t) - \bar{x}),\ t = 1, 2, \ldots, L. \tag{25}$$

Note that we can generate many different copies of B, thus, obtaining different $s(t)$ for the same dynamical system. This property is used below to gain statistical measures for different methods.

2. Spatial loadings

Each generator has a specific loading or a vector of spatial weights $w \in \mathbb{R}^n$. In experimental conditions, it is defined by the anatomy, the electrode characteristics, and its placement in the neuronal tissue. For instance, local generators have strongly nonlinear w, whereas generators, volume conducted from other brain regions, have linear spatial weights (for more details, see the review[17]). Here, we used two approaches: (1) Loadings obtained experimentally in the rat hippocampus [as shown in Fig. 1(a), right, for details, see Ref. 16] and (2) Random loadings taken i.i.d. from the uniform distribution $U[-1, 1]^n$. For numerics, we used $n = 16$, i.e., simulating recordings with 16 electrodes.

3. Simulated LFPs

Finally, we build LFPs using m generators with the corresponding time courses and spatial loadings,

$$x(t) = \sum_{i=1}^{m} w_i s_i(t). \tag{26}$$

The CD of, thus, obtained LFPs is equal to the sum of the CDs of the sources $\{s_i\}_{i=1}^m$ and, hence, to the CDs of the original attractors and quasi-periodic time series Eqs. (21)–(24).

B. The method at work

First, let us illustrate the proposed method on semi-synthetic data. We generated 16-channel LFPs consisting of $L = 10^4$ samples [Fig. 2(a)] by using the procedure described in Sec. III A employing the Lorenz and Double Scrool attractors. Assuming the standard

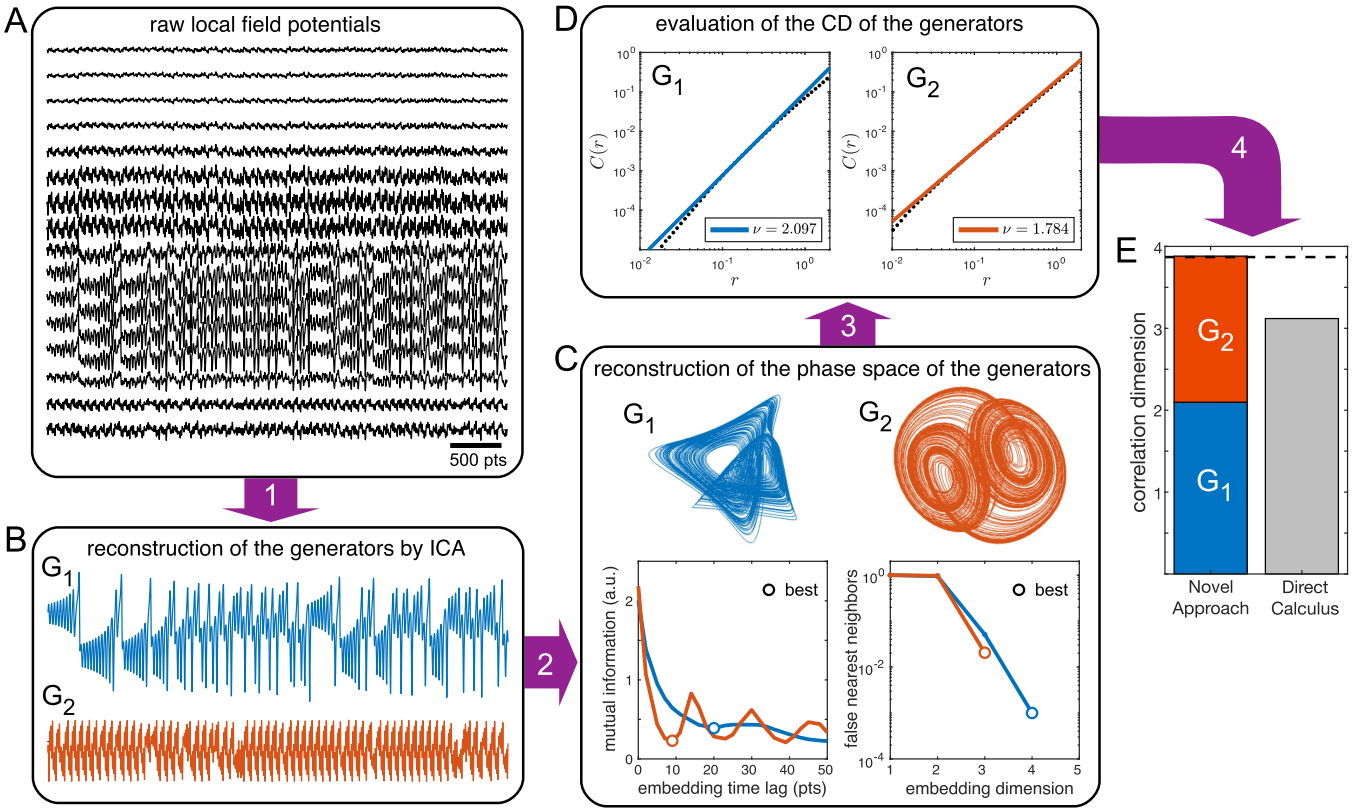

FIG. 2. Example of the method applied to the semisynthetic LFPs. *Step 1*: LFPs created from the Lorenz and double scroll attractors (a) are separated into two generators (G_1 and G_2) by the independent component analysis (b). *Step 2*: The phase space is reconstructed for each generator separately (c). The optimal time lags are 20 and 9; the embedding dimensions are 4 and 3 for G1 and G2, respectively. *Step 3*: The CD for each invariant set is estimated (d), which provides $\nu_1 = 2.097$ and $\nu_2 = 1.784$. *Step 4*: The final correlation dimension is $\nu_1 + \nu_2 = 3.881$ (e), which is close to the theoretical value of 3.873 (dashed black line). For comparison, the direct evaluation of the CD from raw LFPs provides (grey bar) $\nu_{LFP} \approx 3.119$.

sampling frequency of 1 kHz, the data set corresponds to a 10 s epoch. Then, we estimated the CD by applying the novel method described in Sec. II D.

1. The independent component analysis (ICAofLFPs package) revealed the presence of two generators, as expected (G_1 and G_2 correspond to the Lorenz and Double Scrool attractors, respectively). Their loadings were close to the original loadings of the Schaffer and LM generators [Fig. 1(a), right]. Figure 2(b) shows the found generators' time courses (activations).
2. We use the generators' activations to reconstruct invariant sets in the phase spaces of the generators [Fig. 2(c)]. Note that this and the following steps are done separately for each generator. For the reconstruction, we defined the optimal time lag by evaluating the AMI and the embedding dimension by evaluating the FNN ratio. The procedure gave $\tau_1 = 20$, $d_1 = 4$ for G_1 and $\tau_2 = 9$, $d_2 = 3$ for G_2. Figure 2(c) (top) shows the reconstructed invariant sets (a 3D projection for G_1).
3. We apply the Grassberger–Procaccia algorithm to estimate the correlation dimensions for each generator [Fig. 2(d)]. It was done by estimating the correlation integral $C(r)$ as a function of the ball size r and fitting a straight line. The line slope provided the CDs: $\nu_1 = 2.097$ and $\nu_2 = 1.784$. Note that these values are close to the CDs of the Lorenz (2.044) and Double Scroll (1.829) attractors.
4. By adding up the found CDs, we obtained the correlation dimensions of the original 2HD data set: $\nu = 3.881$ [Fig. 2(e), the blue-red bar]. This figure agrees with the theoretical value of 3.873.

To compare the method's performance to the standard approach (Sec. II B), we applied it to the same 2HD data. The calculation yielded a CD of 3.119 [Fig. 2(e), the gray bar], significantly below the theoretical value (>19%). Such a discrepancy is due to the intrinsic difficulties arising when we deal with data in high-dimensional spaces.[23] On the one hand, the high initial data dimension ($n = 16$) increases the optimal lag ($\tau = 29$) and the variance. In turn, the embedding dimension for reconstructing the state space goes to 48, significantly increasing the distance between two samples in the set. On the other hand, the high CD of the data ($\nu \approx 4$) induces a fast drop in the number of points within a ball of radius r, which is necessary to evaluate the correlation integral, which reduces the calculation precision. Thus, an accurate estimation of the correlation integral requires the number of samples orders of magnitude higher than with the proposed method.

C. Method performance vs sample size

As mentioned in Sec. I, one of the main objectives is to reduce the amount of data required to estimate the correlation dimension accurately. Let us now study this problem.

We repeated the calculations described in Sec. III B for different values of the sample size L, spanning the range from 10^3 to 10^6 points. We used the Monte Carlo method to estimate the CD's mean value and standard deviation.

Figure 1(b) shows the results. We observe that the proposed method performs well with $L \approx 5 \times 10^3$ points, whereas the standard direct approach to the correlation dimension requires a sample size above 10^6. Thus, the novel method offers at least a two-order magnitude gain on LFPs contributed by only two generators, as we predicted in Sec. III B.

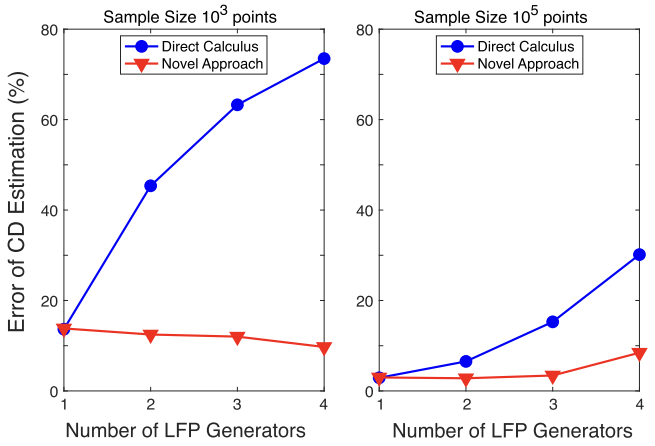

FIG. 3. Performance of the CD estimation for different complexity of LFPs. An increase in the number of LFP generators leads to an increase in LFP complexity. For a single generator (minimal complexity), the direct estimation of the CD from LFP generators (blue curves) and by the novel method (red curves) gives the same error in accordance with Lemma 1. Left and right figures correspond to sample sizes of 10^3 and 10^5 points, respectively (i.e., 1 and 100 s LFPs recorded at 1 kHz sampling rate).

D. Method performance for different data complexity

We now study how the performance of the CD estimation depends on the number of generators contributing to LFPs, i.e., on the complexity of LFPs.

We built 16-channel LFPs using the procedure described in Sec. III A. The maximal sample size was set to $L = 10^3$ or 10^5 points. In each experiment, the LFPs had random loadings $\{\nu_i\}$ [Eq. (26)] and had been contributed by a different number of generators from one to four. The generators have been selected using all possible combinations of the Lorenz, Double Scroll, and Rossler attractors and a quasi-periodic time series. Thus, different LFPs had different complexities, going from the minimal (contributed by a single generator) to the maximal (contributed by the four generators). Note that four generators live in a 12-dimensional space, approaching the maximum of 16 available channels.

Figure 3 shows the results. The error of estimating the correlation dimension by the direct method increases strongly with the increased complexity of LFPs, reaching 70% for $L = 10^3$ pts and 30% for $L = 10^5$ pts (Fig. 3, blue curves) for four generators, as expected.

The estimation of the CD by the proposed approach was consistently more precise, except for a single generator when the errors were the same as with the direct method. The latter is expected and follows from Lemma 1. A higher number of generators slightly improves the performance for a small sample size (Fig. 3, left, red curve). However, longer time series invert the tendency (Fig. 3, right,

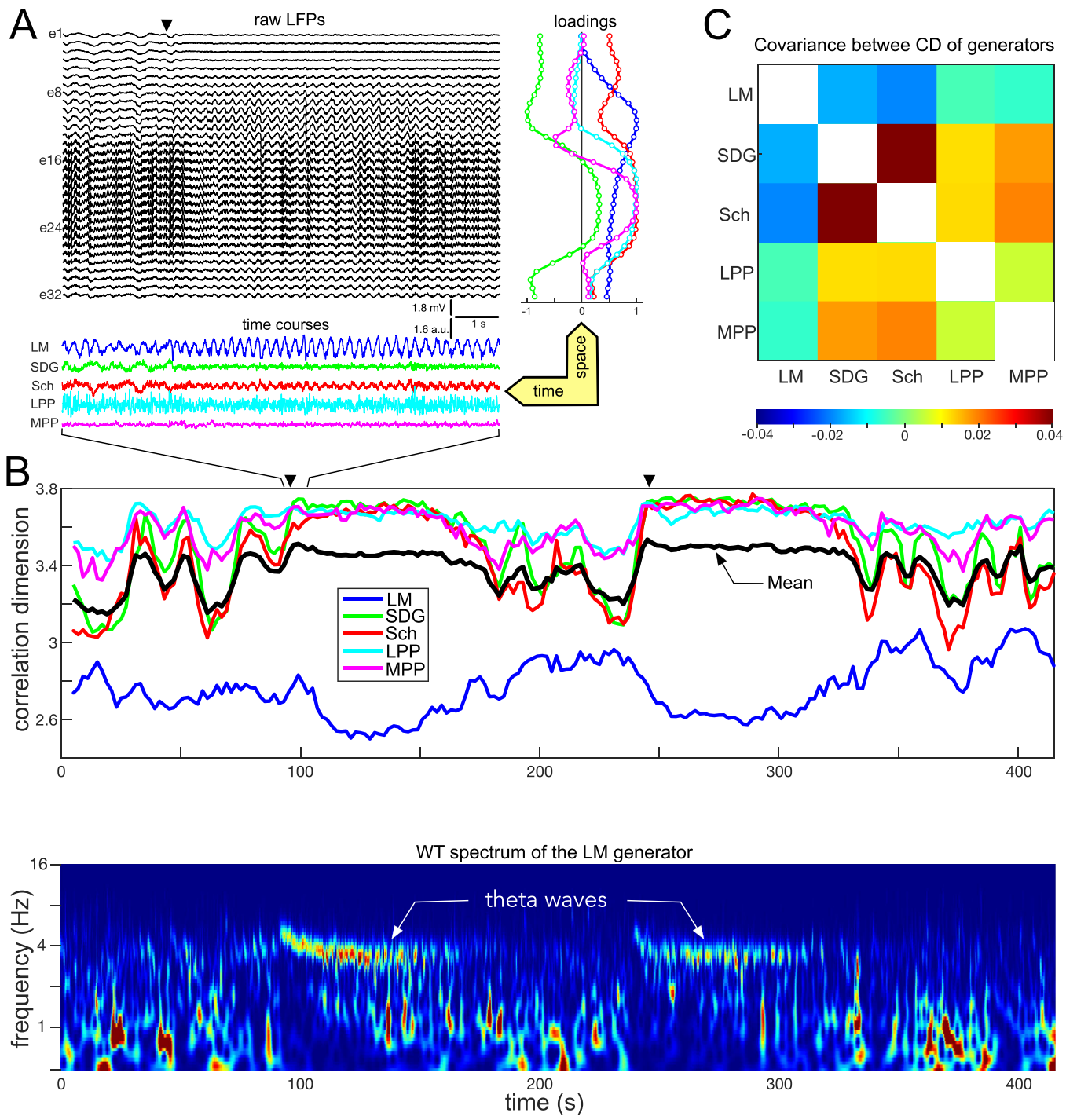

FIG. 4. The CD analysis of experimental data. (a) A short epoch of raw LFPs recorded by a 32-channel electrode in the rat hippocampus and five generators composing the LFPs (obtained by ICAofLFPs package): Lacunosum Moleculare (LM), Somatic Dentate Gyrus (SDG), Schaffer and recurrent CA3 (Sch), Lateral Perfornat Path (LPP), and Medial Perfornat Path (MPP). The black triangle marks the time instant of a tail pinch. (b) *Top:* Dynamics of the correlation dimension obtained by the method for each generator (colored curves) and the mean (black curve, see the main text). *Bottom:* The wavelet spectrum of the time course of the LM generator. It exhibits two theta waves corresponding to drops of the CD of the LM generator. (c) The covariance matrix of the CD dynamics of the five generators.

red curve) and facilitate better CD estimation. The former counterintuitive behavior is related to treating LFPs using the independent component analysis.

E. Analysis of experimental LFPs

1. LFP recordings

We recorded LFPs from the rat hippocampus with a 32-channel linear silicone probe with 100 μm intersite distance (Atlas Neuroengineering, Belgium). Adult Wistar rats were anesthetized with Urethane (1.2 g/kg, i.p.) and placed in a stereotaxic device. The surgical and stereotaxic procedures were described elsewhere.[32] The probes were stereotaxically placed in the CA1 area and Dentate Gyrus (DG) of the hippocampus (AP: 4.5; L: 2.6; V: 3.5, according to the atlas[33]). Signals were amplified and acquired using Multi-Channel Systems (Germany) at the sampling rate of 10 kHz for 450 s and then downsampled to 1 kHz for the CD analysis [Fig. 4(a), raw LFPs]. During the recording, the rat tail was pinched twice [small triangles in Fig. 4(a) at $t = 92$ and 242 s]. Such a stimulus elicits a sensory-driven electrographic state characterized by the appearance of strong Theta waves (3.5–6 Hz) in the CA1 area and DG of the hippocampus.[34]

2. Data analysis

The independent component analysis revealed five LFP generators [Fig. 4(a), loadings, and time courses], which have been identified by their biophysical characteristics as Lacunosum Moleculare (LM), Somatic Dentate Gyrus (SDG), Schaffer and recurrent CA3 (Sch), Lateral Perfornat Path (LPP), and Medial Perfornat Path (MPP) (for details on the procedure, see Ref. 35). Then, the correlation dimensions of the generators were evaluated by the proposed method over a sliding time window of 10 s (a sample of 10^4 points). Such a time interval is a compromise between a good enough CD estimation of the method [Fig. 1(b)] and a reasonable time resolution of the measure (see Sec. II E).

Figure 4(b) (top) shows the time evolution of the CD of the generators and the mean CD (showing the mean instead of the total CD of LFPs simplifies the data visualization). The total CD of the LFPs can be obtained by multiplying the mean by five, i.e., by the number of generators. It oscillates around 17.

The CD of the LM generator is significantly lower than that of the others, i.e., this generator has low complexity compared to other generators. Moreover, its CD drops at the tail pinches and slowly recovers the original value. Such an observation is supported by the wavelet spectrum [Fig. 4(b), bottom; for details, see Ref. 36] exhibiting pronounces theta waves in these time intervals. We also observe that the CDs of other generators strongly oscillate during irregular hippocampal activity (out of theta waves). The total CD of LFPs also oscillates during irregular activity, staying almost constant in periods of theta waves.

We note that the decomposition of the total CD into contributing parts not only allows better estimation but also provides a means for deeper analysis of the information processing. We can now study, e.g., the relation between the dynamic complexity of the generators. For example, we evaluated the covariance matrix for the CD of the generators [Fig. 4(c)]. In particular, there is high covariance between the Schaffer and Somatic Dentate Gyrus. Thus, these generators synchronously change the complexity of their oscillations (note that their time courses may not be synchronized). The LM generator exhibits antiphase complexity and low relation to the complexity of information received by the hippocampus from the Entorhinal cortex through the LPP and MPP generators.

IV. CONCLUSIONS

The estimation of the correlation dimension of 2HD data is a challenging problem. The main issues are the nonstationarity of biological data and their potentially high dimension. The nonstationarity requires reducing the sample size while high dimension prescribes its increase. Such a contradiction could explain the controversy observed in estimating the CD from experimental data.

Here, we have provided a novel method and its mathematical justification for an accurate estimate of the CD from 2HD experimental data. The method reduces the sample size by several orders of magnitude without compromising the accuracy of the CD estimation. It is based on decomposing the original time series into independent components and then estimating each component's CD individually. The CD estimation requires exponentially shorter samples since individual components have much lower dimensions. We have illustrated the method's performance on semi-synthetic data simulating hippocampal LFPs. The results indeed showed orders of magnitude improvement in the sample size and a significant error drop for increasing data complexity compared to the standard method.

Another advantage of this method is the information provided on the CDs of individual generators. It enables studying information processing and how the complexity of different components contributes to the total CD over time, e.g., the origin of abrupt changes in brain dynamics. In the case of LFPs, the contributing components can be physiologically identified and related to pathways conveying information to the region of interest. We have exemplified this approach on LFPs recorded from the rat hippocampus. In particular, we have shown that during theta waves elicited by a tail pinch, the information complexity of the Lacunosum–Moleculare generator drops and then slowly recovers. Other generators exhibit strong CD oscillations during irregular activity, while the LM generator keeps constant complexity. These results confirm the significant potential of the method and support its use together with other linear measures in data analysis.

ACKNOWLEDGMENTS

This work has been supported by the Spanish Ministry of Science and Innovation (Project No. PID2021-124047NB-I00), the Santander-UCM grant PR44/21-29927 to V.A.M., and the Next Generation EU grant PDC2021-121103-I00 to O.H.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Ethics approval

The experiments were performed in accordance with EU (2010/63/UE), Spanish (RD 53/2013), and local (Autonomous Community of Madrid, Order 4/8/1988) regulations regarding the use of laboratory animals, and the experimental protocols were approved by the Research Committee of the Cajal Institute.

Author Contributions

Valeri A. Makarov: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Software (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Ricardo Muñoz:** Data curation (equal); Validation (equal); Writing – review & editing (equal). **Oscar Herreras:** Conceptualization (equal); Funding acquisition (equal); Writing – review & editing (equal). **Julia Makarova:** Conceptualization (equal); Data curation (equal); Validation (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

REFERENCES

[1] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," Physica D **9**, 189–208 (1983).
[2] K. Lehnertz and C. E. Elger, "Can epileptic seizures be predicted? Evidence from nonlinear time series analysis of brain electrical activity," Phys. Rev. Lett. **80**, 5019–5023 (1998).
[3] F. Mormann, T. Kreuz, C. Rieke, R. G. Andrzejak, A. Kraskov, P. David, C. E. Elger, and K. Lehnertz, "On the predictability of epileptic seizures," Clinical Neurophysiol. **116**(3), 569–587 (2005).
[4] M. Y. Boon, B. I. Henry, C. M. Suttle, and S. J. Dain, "The correlation dimension: A useful objective measure of the transient visual evoked potential?," J. Vision **8**(1), 1–21 (2008).
[5] C. E. Elger and K. Lehnertz, "Seizure prediction by non-linear time series analysis of brain electrical activity," Eur. J. Neurosci. **10**, 786–789 (1998).
[6] T. Maiwald, M. Winterhalder, R. Aschenbrenner-Scheibe, H. U. Voss, A. Schulze-Bonhage, and J. Timmer, "Comparison of three nonlinear prediction methods by means of the seizure prediction characteristic," Physica D **194**, 357–368 (2004).
[7] R. Aschenbrenner-Scheibe, T. Maiwald, M. Winterhalder, H. U. Voss, J. Timmer, and A. Schulze-Bonhage, "How well can epileptic seizures be predicted? An evaluation of a nonlinear method," Brain **126**, 2616–2626 (2003).
[8] R. Bellman, "Dynamic programming," Science **153**(3731), 34–37 (1966).
[9] B. Mandelbrot, *The Fractal Geometry of Nature* (W. H. Freeman and Co., 1982).
[10] P. Grassberger and I. Procaccia, "Characterization of strange attractors," Phys. Rev. Lett. **50**, 346–349 (1983).
[11] J. Sprott and G. Rowlands, "Improved correlation dimension calculation," Int. J. Bifurcation Chaos **11**(7), 1865–1880 (2001).
[12] J. D. Farmer, E. Ott, and J. A. Yorke, "The dimension of chaotic attractors," Physica D **7**, 153–180 (1983).
[13] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," J. Stat. Phys. **65**, 579–616 (1991).
[14] J. Theiler, "Lacunarity in a best estimator of fractal dimension," Phys. Lett. A **133**(4), 195–200 (1998).
[15] V. A. Makarov, J. Makarova, and O. Herreras, "Disentanglement of local field potential sources by independent component analysis," J. Comp. Neurosci. **29**, 445–457 (2010).
[16] A. Korovaichuk, J. Makarova, V. A. Makarov, N. Benito, and O. Herreras, "Minor contribution of principal excitatory pathways to hippocampal LFPs in the anesthetized rat: A combined independent component and current source density study," J. Neurophysiol. **104**, 484–497 (2010).
[17] O. Herreras, D. Torres, G. Martin-Vazquez, S. Hernandez-Recio, V. J. Lopez-Madrona, N. Benito, V. A. Makarov, and J. Makarova, "Site-dependent shaping of field potential waveforms," Cereb. Cortex **33**(7), 3636–3650 (2023).
[18] R. Lorente de No, "Analysis of the distribution of action currents of nerves in volume conductors," in *A Study of Nerve Physiology* (The Rockefeller Institute, New York, 1947), Vol. 132, Part 2, pp. 384–477.
[19] P. L. Nunez and R. Srinivasan, *Electric Fields of the Brain: The Neurophysics of EEG*, 2nd ed. (New York, 2006).
[20] O. Herreras, D. Torres, V. A. Makarov, and J. Makarova, "Theoretical considerations and supporting evidence for the primary role of source geometry on field potential amplitude and spatial extent," Front. Cell. Neurosci. **17**, 1129097 (2023).
[21] H. Whitney, "Differentiable manifolds," Ann. Math. **37**(3), 645–680 (1936).
[22] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence*, edited by D. A. Rand and L. S. Young (Springer-Verlag, Berlin, 2002), pp. 366–381.
[23] A. N. Gorban, V. A. Makarov, and I. Y. Tyukin, "High-dimensional brain in a high-dimensional world: Blessing of dimensionality," Entropy **22**, 82 (2020).
[24] S. Wallot and S. Monster, "Calculation of average mutual information (AMI) and false-nearest neighbors (FNN) for the estimation of embedding parameters of multidimensional time series in Matlab," Front. Psychol. **9**, 1679 (2018).
[25] I. Vlachos and D. Kugiumtzis, "State space reconstruction from multiple time series," in *Topics on Chaotic Systems*, edited by C.H. Skiadas, I. Dimotikalis, and C. Skiadaspp (Hanover College, Indiana, USA, 2009), pp. 378–387.
[26] M. B. Kennel, R. Brown, and H. D. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," Phys. Rev. A **45**, 3403 (1992).
[27] A. Hyvarinen and E. Oja, "Independent component analysis: Algorithms and applications," Neur. Networks **13**(4–5), 411–430 (2000).
[28] A. Chen, "Fast kernel density independent component analysis," Lect. Notes Comput. Sci. **3889**, 24–31 (2006).
[29] O. Herreras, J. Makarova, and V. A. Makarov, "New uses of LFPs: Pathway-specific threads obtained through spatial discrimination," Neurosci. **310**, 486–503 (2015).
[30] M. W. Hirsch, S. Smale, and R. Devaney, *Differential Equations, Dynamical Systems. An Introduction to Chaos*, 2nd ed. (Academic Press, Boston, MA, 2003).
[31] L. Chua, M. Komuro, and T. Matsumoto, "The double scroll family," IEEE Trans. Circ. Syst. **33**(11), 1072–118 (1986).
[32] S. Canals, I. Makarova, L. López-Aguado, C. Largo, J. M. Ibarz, and O. Herreras, "Longitudinal depolarization gradients along the somatodendritic axis of CA1 pyramidal cells: A novel feature of spreading depression," J. Neurophysiol. **94**(2), 943–951 (2005).
[33] G. Paxinos and C. Watson, *The Rat Brain in Stereotaxic Coordinates*, 6th ed. (Academic Press, 2007).
[34] O. S. Vinogradova, "Expression, control, and probable functional significance of the neuronal theta-rhythm," Prog. Neurobiol. **45**(6), 523–583 (1995).
[35] N. Benito, A. Fernández-Ruiz, V. A. Makarov, J. Makarova, A. Korovaichuk, and O. Herreras, "Spatial modules of coherent activity in pathway-specific LFPs in the hippocampus reflect topology and different modes of presynaptic synchronization," Cereb. Cortex **24**(7), 1738–1752 (2014).
[36] A. E. Hramov, A. A. Koronovskii, V. A. Makarov, V. A. Maksimenko, A. N. Pavlov, and E. Sitnikova, *Wavelets in Neuroscience*, 2nd ed. (Springer, 2021).