

Manifold-adaptive dimension estimation revisited

Zsigmond Benkő^{1,2,*}, Marcell Stippinger¹, Roberta Rehus¹, Attila Bencze¹, Dániel Fabó⁴,
Boglárka Hajnal^{2,4}, Loránd Eröss^{5,6}, András Telcs^{1,3,7}, and Zoltán Somogyvári^{1,8}

¹Department of Computational Sciences, Wigner Research Centre for Physics, H-1121, Hungary

²János Szentágotthai Doctoral School of Neurosciences, Semmelweis University, H-1085, Hungary

³Department of Computer Science and Information Theory, Faculty of Electrical Engineering and Informatics,
Budapest University of Technology and Economics, Budapest, H-1111, Hungary

⁴Epilepsy Center, Department of Neurology, National Institute of Clinical Neurosciences, Budapest, H-1145,
Hungary

⁵Department of Functional Neurosurgery, National Institute of Clinical Neurosciences, Budapest, H-1145,
Hungary

⁶Faculty of Information Technology and Bionics, Péter Pázmány Catholic University, Budapest, H-1083,
Hungary

⁷Department of Quantitative Methods, University of Pannonia, Faculty of Business and Economics, H-8200,
Veszprém, Hungary

⁸Neuromicrosystems Ltd., Budapest, H-1113, Hungary

*benko.zsigmond@wigner.hu

August 11, 2020

Abstract

Data dimensionality informs us about data complexity and sets limit on the structure of successful signal processing pipelines. In this work we revisit and improve the manifold adaptive Farahmand-Szepesvári-Audibert (FSA) dimension estimator, making it one of the best nearest neighbor-based dimension estimators available. We compute the probability density function of local FSA estimates, if the local manifold density is uniform. Based on the probability density function, we propose to use the median of local estimates as a basic global measure of intrinsic dimensionality, and we demonstrate the advantages of this asymptotically unbiased estimator over the previously proposed statistics: the mode and the mean. Additionally, from the probability density function, we derive the maximum likelihood formula for global intrinsic dimensionality, if i.i.d. holds. We tackle edge and finite-sample effects with an exponential correction formula, calibrated on hypercube datasets. We compare the performance of the corrected-median-FSA estimator with kNN estimators: maximum likelihood (ML, Levina-Bickel) and two implementations of DANCo (R and matlab). We show that corrected-median-FSA estimator beats the ML estimator and it is on equal footing with DANCo for standard synthetic benchmarks according to mean percentage error and error rate metrics. With the median-FSA algorithm, we reveal diverse changes in the neural dynamics while resting state and during epileptic seizures. We identify brain areas with lower-dimensional dynamics that are possible causal sources and candidates for being seizure onset zones.

Introduction

Dimensionality sets profound limits on the stage where data takes place, therefore it is often crucial to know the intrinsic dimension of data to carry out meaningful analysis. Intrinsic dimension provides direct information about data complexity, as such, it was recognised as a useful measure to describe the dynamics of dynamical systems[1], to detect anomalies in time series[2], to diagnose patients with various conditions[3, 4, 5, 6] and to use it simply as plugin parameter for signal processing algorithms.

Most of the multivariate datasets lie on a lower dimensional manifold embedded in a potentially very high-dimensional embedding space. This is because the observed variables are far from independent, and this interdependence introduces redundancies resulting in a lower intrinsic dimension

(ID) of data compared with the number of observed variables. To capture this – possibly non-linear – interdependence, nonlinear dimension-estimation techniques can be applied[7, 8, 9, 10].

To estimate the ID of data various approaches have been proposed, for a full review of techniques see the work of Campadelli et al.[11]. Here we discuss the k -Nearest Neighbor (k NN) ID estimators, with some recent advancements in the focus.

A usually basic assumption of k NN ID estimators is that the fraction of points in a neighborhood is approximately determined by the intrinsic dimensionality (D) and distance (R) times a – locally almost constant – mostly density-dependent factor ($\eta(x, R)$, Eq. 1).

$$\frac{k}{n} \approx \eta(x, R) * R_k^D \quad (1)$$

where k is the number of samples in a neighborhood and n is the total number of samples on the manifold.

Assuming a Poisson sampling process on the manifold Levina and Bickel[12] derived a Maximum Likelihood estimator, which became a popular method and got several updates[13, 14]. These estimators are prone to underestimation of dimensionality because of finite sample effects and overestimations because of the curvature.

To address the challenges posed by curvature and finite sample, new estimators were proposed [15, 16, 17, 18]. To tackle the effect of curvature, a minimal neighborhood size can be taken on normalized neighborhood distances as in the case of MIND_{ML}[15]. To tackle the underestimation due to finite sample effects, empirical corrections were applied. A naive empirical correction approach was applied by Camastra and Vinciarelli[19]: a perceptron was trained on the estimates computed for randomly sampled hypercubes to learn a correction function. Motivated by the correction in the previous work, the IDEA method was created[15]; and a more principled approach was carried out, where the full distribution of estimates was compared to the distributions computed on test data sets using the Kullback-Leibner divergence (MIND_{KL}[15], DANCo[17]). In the case of DANCo, not just the nearest neighbor distances, but the angles are measured and taken into account in the estimation process resulting in more accurate estimates.

In the recent years, further estimators have been proposed, such as the estimator that uses minimal neighborhood information leveraging the empirical distribution of the ratio of the nearest neighbors to fit intrinsic dimension[18], or other approaches based on simplex skewness[20] and normalized distances [21, 22, 23, 24].

In the following section, we revisit the manifold adaptive dimension estimator proposed by Farahmand et al.[25] to measure intrinsic dimensionality of datasets. From Eq. 1 we can take the logarithm of both sides:

$$\begin{aligned} \ln\left(\frac{k}{n}\right) &\approx \ln \eta + D \ln R_k \\ \ln\left(\frac{2k}{n}\right) &\approx \ln \eta + D \ln R_{2k} \end{aligned} \quad (2)$$

If η is slowly varying and R is small, we can take it as a constant.

If we subtract the two equations from each other we get:

$$\ln(2) \approx D \ln\left(\frac{R_{2k}}{R_k}\right) \quad (3)$$

Thus, to compute the local estimates, we fit a line through the log-distance k th and $2k$ th nearest neighbor at a given location.

$$d(x) = \frac{\ln(2)}{\ln(R_{2k}/R_k)} \quad (4)$$

To compute a global ID estimate, the authors proposed the mean of local estimates at sample-points, or a vote for the winner global ID value (the mode), if the estimator is used in integer-mode.

They proved that the above global ID estimates are consistent for $k > 1$, if η is differentiable and the manifold is regular. They calculated the upper bound for the probability of error for the global estimate, however this bound contains unknown constants[25].

In this paper we propose an improved FSA estimator, based on the assumption that the density is locally uniform. We suggest to use the median of local values for a global intrinsic dimension estimate. We correct the underestimation effect by an exponential formula and test the new algorithm on benchmark datasets. We apply the proposed estimator to locate epileptic focus on field potential measurements.

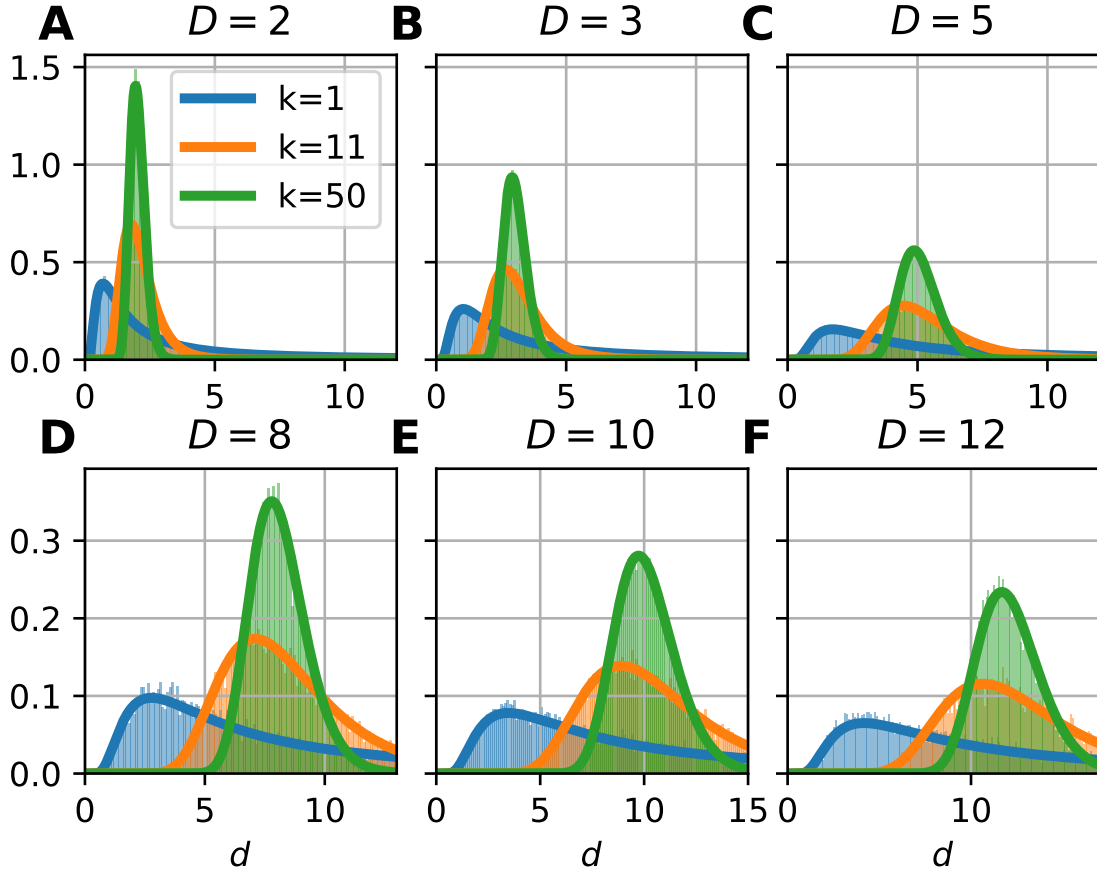


Figure 1: **Probability density function of the Farahmand-Szepesvári-Audibert estimator (d) for various dimensions (D) and neighborhood sizes (k).** **A-F** The sublots show that theoretical pdfs (continuous lines) fits to the histograms ($n = 10000$) of local estimates calculated on uniformly sampled hypercubes ($D = 2, 3, 5, 8, 10, 12$). The three colors denote three presented neighborhood sizes: $k = 1$ (blue), $k = 11$ (orange) and $k = 50$ (green).

Results

Manifold adaptive dimension estimator revisited

The probability density of Farahmand-Szepesvári-Audibert estimator

We compute the probability density function of Farahmand-Szepesvári-Audibert (FSA) intrinsic dimension estimator based on normalized distances.

The normalized distance density of the k NN can be computed in the context of a K -neighborhood, where the normalized distance of $K - 1$ points follows a specific form:

$$p(r|k, K-1, D) = \frac{D}{B(k, K-k)} r^{Dk-1} (1-r^D)^{K-k-1} \quad (5)$$

where r is the normalized distance of the k th neighbor by the distance of K th neighbor ($r_k = R_k/R_K$, $k < K$) and B is the Euler-beta function (see SI A.1 for a derivation). A maximum likelihood estimator based on Eq. 5 leads to the formula of the classical Levina-Bickel estimator ([12]). For a derivation of this probability density and the maximum likelihood solution see SI A.1 and SI A.2 respectively.

We realize that the inverse of normalized distance appears in the formula of FSA estimator, so we can express it as a function of r :

$$d_k = \frac{\log 2}{\log(R_{2k}/R_k)} = -\frac{\log 2}{\log(R_k/R_{2k})} = -\frac{\log 2}{\log r_k} \quad (6)$$

Where $r_k = R_k/R_{2k}$.

Thus, we can compute the pdf of the estimated values as plugging in $K = 2k$ into Eq. 5 followed by change of variables:

$$q(d_k) \equiv p(r|k, 2k-1, D) \left| \frac{dr}{dd_k} \right| = \frac{D \log(2)}{B(k, k)} \frac{2^{-\frac{Dk}{d_k}} \left(1 - 2^{-\frac{D}{d_k}}\right)^{k-1}}{d_k^2} \quad (7)$$

Theorem 1. *The median of $q(d_k)$ is at D .*

Proof. We apply the substitution $a = 2^{-D/d_k}$ in Eq. 7 (Eq. 10):

$$p(a) = q(d_k) \left| \frac{dd_k}{da} \right| = \quad (8)$$

$$= \frac{D \log(2)}{B(k, k)} \frac{a^k (1-a)^{k-1} \log^2 a}{D^2 \log^2 2} \frac{D \log 2}{a \log^2 a} \quad (9)$$

$$= \frac{1}{B(k, k)} a^{k-1} (1-a)^{k-1} \quad (10)$$

The pdf in Eq.10 belongs to a beta distribution. The cumulative distribution function of this density is the regularized incomplete Beta function with k as both parameters symmetrically.

$$P(a) = I_a(k, k) \quad (11)$$

The median of this distribution is at $a = \frac{1}{2}$, thus at $d_k = D$ since:

$$a = 2^{-\frac{D}{d_k}} = \frac{1}{2} \quad (12)$$

$$D = d_k \quad (13)$$

□

This means that the median of the FSA estimator is equal to the intrinsic dimension independent of neighborhood size, if the locally uniform point density assumption holds. The sample median is a robust statistic, therefore we propose to use the sample median of local estimates as a global dimension estimate. We will call this modified method the median Farahmand-Szepesvári-Audibert (mFSA) estimator.

Let's see the form for the smallest possible neighborhood size: $k = 1$ (Fig.1). The pdf for the estimator takes a simpler form (Eq. 14).

$$q(d|k=1, D) = D \log(2) \frac{2^{-\frac{D}{d_1}}}{d_1^2} \quad (14)$$

Also, we can calculate the cumulative distribution function analytically (Eq. 15).

$$Q(d|k=1, D) = \int_0^{d_1} q(t|k=1, D) dt = 2^{-D/d_1} \quad (15)$$

The expectation of d_k diverges for $k=1$ – but not for $k > 1$ – although the median exists. From Eq. 15 the median is at D (Eq. 16).

$$Q(d_1 = D) = 0.5 \quad (16)$$

Sampling distribution of the median

We can easily compute the pdf of the sample median if an odd sample size is given ($n = 2l + 1$) and if sample points are drawn independently according to Eq. 7. Roughly half of the points have to be smaller, half of the points have to be bigger and one point has to be exactly at m (Eq. 17).

$$p(m|k, D, n) = \frac{1}{B(l+1, l+1)} \left[P(a = 2^{-D/m}) \left(1 - P(a = 2^{-D/m}) \right) \right]^l q(m) \quad (17)$$

Where $p(a)$ and $P(a)$ are the pdf and cdf of a (Eq. 10, 11) and q is the pdf of the FSA estimator (Fig. 2).

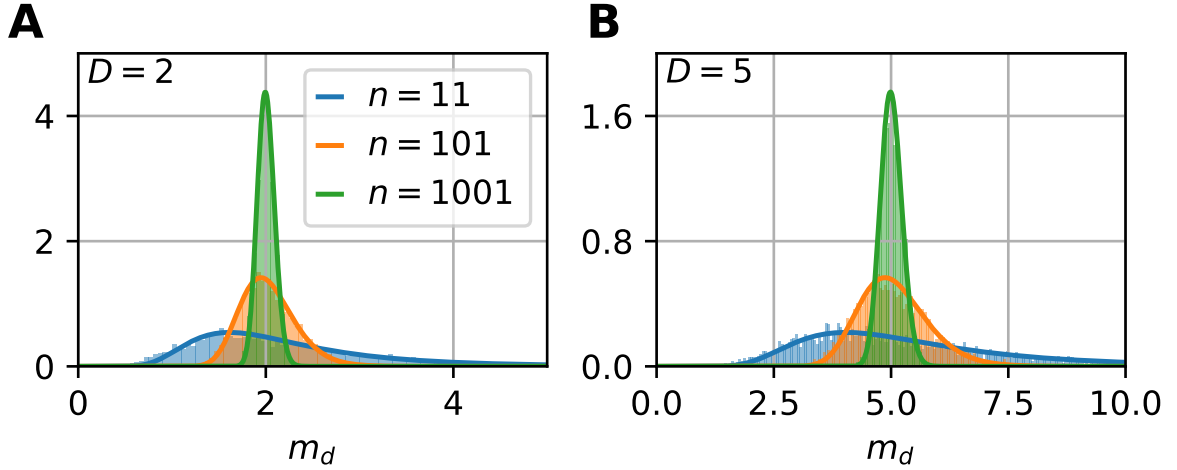


Figure 2: **The sampling distribution of the median for the FSA estimator ($k=1$) on uniformly sampled hypercubes.** The figure shows the pdf of median-FSA estimator of points uniformly sampled from a square (A) and from a 5D hypercube (B) for three sample sizes: $n=11$ (blue), $n=101$ (orange) and $n=1001$ (green) respectively. The solid lines represent the theoretical pdf-s of the median and the shaded histograms are the results of simulations ($N=5000$ realizations).

Maximum Likelihood solution for the manifold-adaptive estimator

If the samples are independent and identically distributed, we can formulate the likelihood function as the product of sample-likelihoods (Eq. 18). We seek for the maximum of the log likelihood function, but the derivative is transcendental for $k > 1$. Therefore, we can compute the place of the maximum numerically (Eq. 20).

$$\mathcal{L} = \prod_{i=1}^n \frac{D \log(2) 2^{-Dk/d_k^{(i)}} (1 - 2^{-D/d_k^{(i)}})^{k-1}}{B(k, k) (d_k^{(i)})^2} \quad (18)$$

$$\log \mathcal{L} = n \log \frac{\log(2)}{B(k, k)} + n \log D - Dk \log(2) \sum \frac{1}{d_k^{(i)}} + (k-1) \sum \log(1 - 2^{-D/d_k^{(i)}}) - 2 \sum \log(d_k^{(i)}) \quad (19)$$

$$\frac{\partial \log \mathcal{L}}{\partial D} = \frac{n}{D} - \log(2)k \sum \frac{1}{d_k^{(i)}} + \log(2)(k-1) \sum \frac{1}{d_k^{(i)}(2^{D/d_k^{(i)}} - 1)} \stackrel{!}{=} 0 \quad (20)$$

For $k = 1$, the ML formula is equal to the Levina-Bickel ($k = 1$) and MIND_{1ML} formulas.

Results on randomly sampled hypercube datasets

Theoretical probability density function of the local FSA estimator fits to empirical observations (Eq. 7, Fig. 1). We simulated hypercube datasets with fixed sample size ($n = 10000$) and of various intrinsic dimensions ($D = 2, 3, 5, 8, 10, 12$). We measured the local FSA estimator at each sample point with 3 different k parameter values ($k = 1, 11, 50$). We visually confirmed that the theoretical pdf fits perfectly to the empirical histograms.

The empirical sampling distribution of mFSA fits to the theoretical curves for small intrinsic dimension values (Fig. 2). To demonstrate the fit, we drew the density of mFSA on two hypersphere datasets $D = 2$ and $D = 5$ with the smallest possible neighborhood ($k = 1$), for different sample sizes ($n = 11, 101, 1001$). At big sample sizes the pdf is approximately a Gaussian[26], but for small samples the pdf is non-Gaussian and skewed.

The mFSA estimator underestimates intrinsic dimensionality in high dimensions. This phenomena is partially a finite sample effect (Fig. 3), but edge effects make this underestimation even more severe. We graphically showed that mFSA estimator asymptotically converged to the real dimension values for hypercube-datasets, when we applied periodic boundary conditions (Fig. 4). We found, that the convergence is much slower for hard boundary conditions, where edge effects make estimation errors higher.

We could estimate the logarithm of relative error with an s -order polynomial:

$$\log(E_{rel}) = \log\left(\frac{D}{d}\right) = \sum_{i=1}^s \alpha_i d^i \quad (21)$$

The order of the polynomial was different for the two types of boundary conditions. When we applied hard boundary, the order was $s = 1$, however in the periodic case higher order polynomials fit the data. Thus, in the case of hard-boundary, we could make the empirical correction formula:

$$D \approx C(\hat{d}) = de^{\alpha_n d} \quad (22)$$

where α_n is a sample size dependent coefficient that we could fit with the least squares method.

Results on synthetic benchmarks

We tested the mFSA estimator and its corrected version on synthetic benchmark datasets[27, 11]. We simulated $N = 100$ instances of 15 manifolds (M_i , $n = 2500$) with various intrinsic dimensions (see Table 1, 2, 4 in Campadelli et al.[11], <http://www.mL.uni-saarland.de/code/IntDim/IntDim.htm>).

We estimated the intrinsic dimensionality of each sample and computed the mean, the error rate and Mean Percentage Error (MPE) for the estimators. We compared the mFS, cmFS, the R and the matlab implementation of DANCo, and the Levina-Bickel estimator (Table 1). cmFSA and DANCo was evaluated in two modes, in a fractal-dimension mode and in an integer dimension mode.

The mFSA and the Levina-Bickel estimator underestimated intrinsic dimensionality, especially in the cases when the data had high dimensionality.

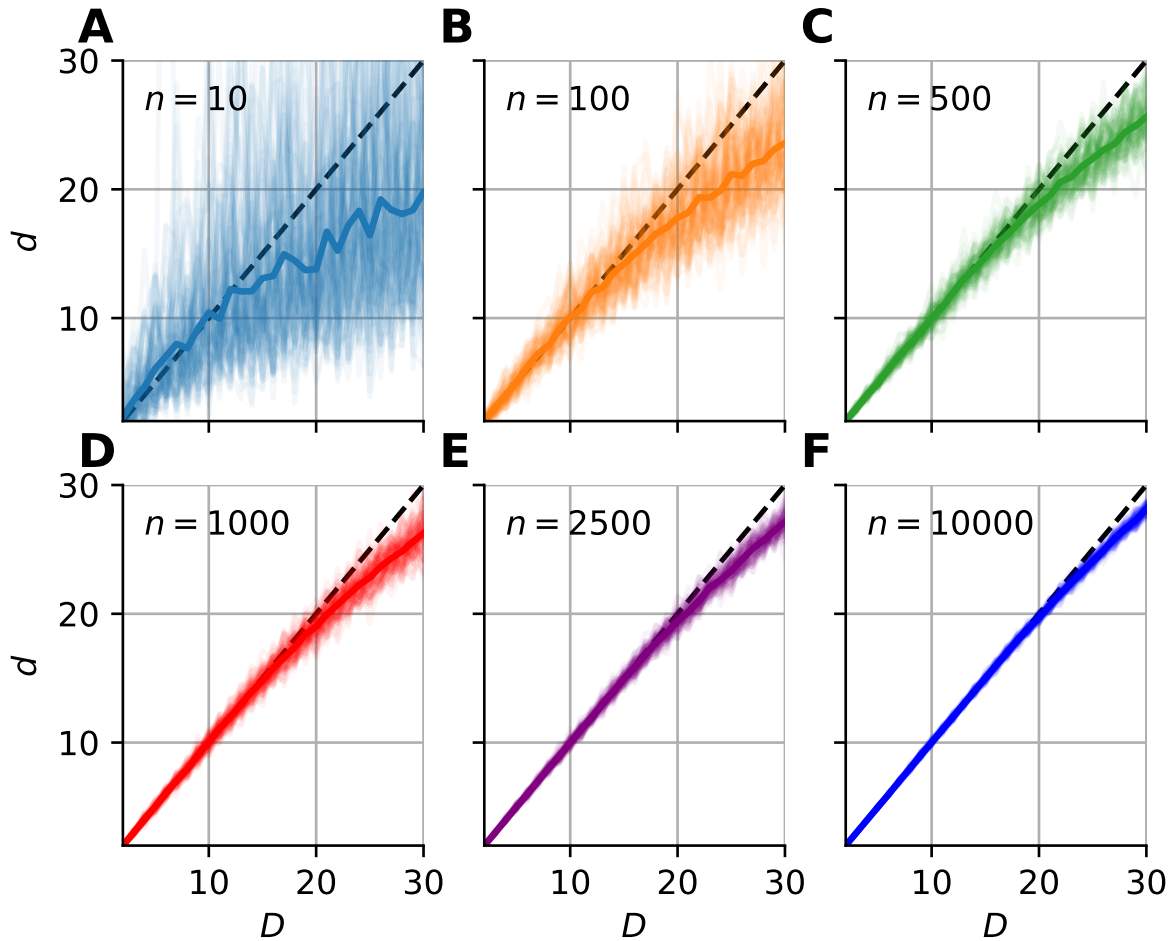


Figure 3: **Intrinsic dimension dependence of the median-FSA estimator for uniformly sampled unit hypercubes with various sample sizes ($k = 1$).** Subplots **A-F** show the mean of median-FSA estimator (thick line) values from $N = 100$ realizations (shading) of uniformly sampled unit hypercubes with periodic boundary.

In contrast, the cmFSA (cmFSA) estimator found the true intrinsic dimensionality of the datasets, it reached the best overall error rate (0.277) and 2nd best MPE (Fig. 6, Table 1). In some cases, it slightly over-estimated the dimension of test datasets. Interestingly, DANCo showed implementation-dependent performance, the matlab algorithm showed the 2nd best error rate (0.323) and the best MPE value (Table 1). The R version overestimated the dimensionality of datasets in most cases.

Analysing epileptic seizures

To show how mFSA works on real-world noisy data, we applied it to human neural recordings of epileptic seizures.

We acquired field potential measurements from a patient with drug-resistant epilepsy by 2 electrode grids and 3 electrode strips. We analyzed the neural recordings during interictal periods and during epileptic activity to map possible seizure onset zones (see Methods).

We found several characteristic differences in the dimension patterns between normal and control conditions. In interictal periods (Fig. 7 A), we found the lowest average dimension value at the FbB2 position on the fronto-basal grid. Also, we observed a diagonal gradient of intrinsic dimensions on the cortical grid (Gr). In contrast, we observed the lowest dimension values at the hippocampal electrode strip (JT), and the gradient on the cortical grid disappeared during seizures (Fig. 7 B). Curiously, the intrinsic dimensionality became higher at fronto-basal recording sites during seizure (Fig. 7 C).

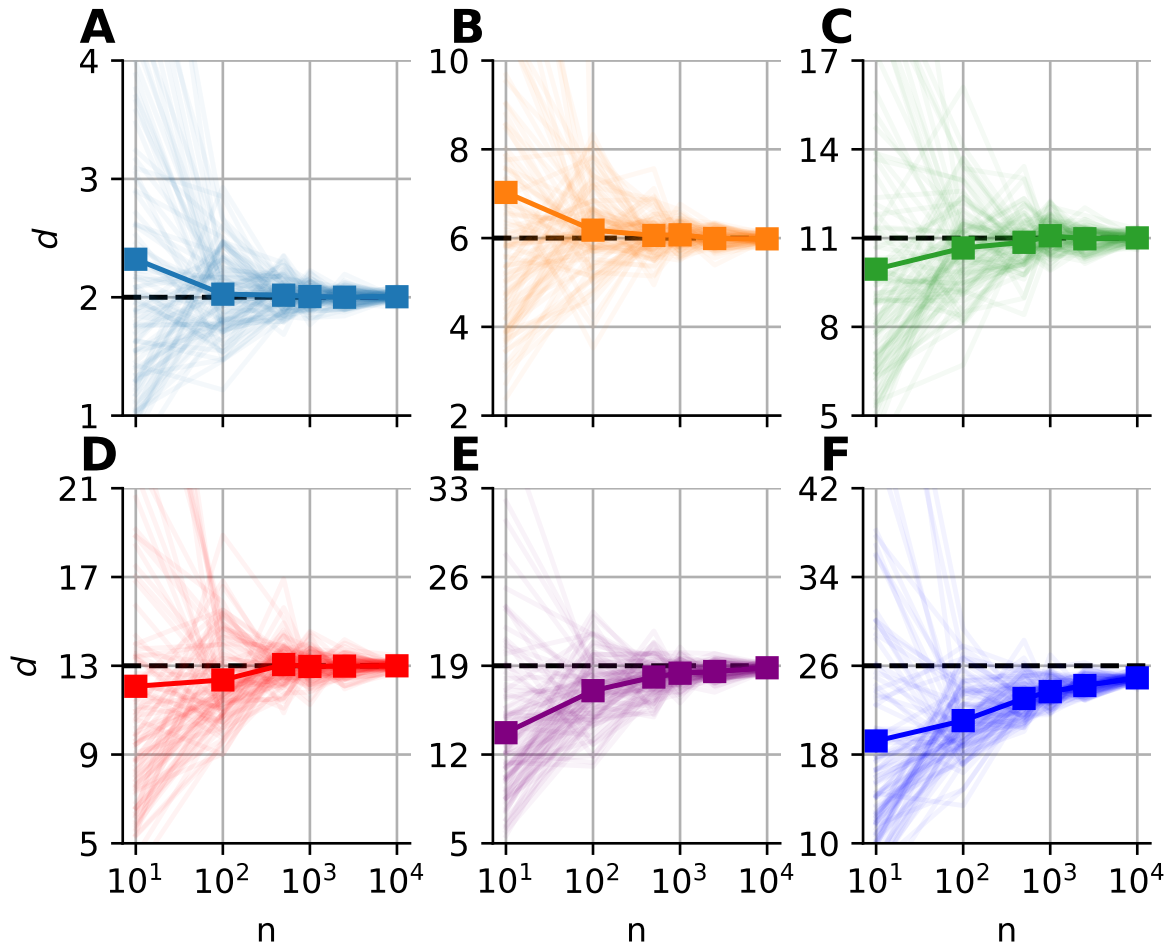


Figure 4: **Sample size dependence of the median-FSA estimator for uniformly sampled unit hypercubes with varied intrinsic dimension value ($k = 1$).** Subplots A-F show the mean of median-FSA estimator (thick line) values from $N = 100$ realizations (shading).

Discussion

In this work we revisited and improved the manifold adaptive FSA dimension estimator. We computed the probability density function of local estimates if the local density was uniform. From the pdf, we derive the maximum likelihood formula for intrinsic dimensionality.

We proposed to use the median of local estimates as a global measure of intrinsic dimensionality, and demonstrated that this measure is asymptotically unbiased.

We tackled edge effects with a correction formula calibrated on hypercube datasets. We showed that the coefficients are sample-size dependent. Camastra and Vinciarelli [19] took a resembling empirical approach, where they corrected correlation dimension estimates with a perceptron, calibrated on d -dimensional datasets. Our approach is different, because we tried to grasp the connection between underestimation and intrinsic dimensionality more directly, by showing that the dimension-dependence of the relative error is exponential. The calibration procedure of DANCo may generalize better, because it compares the full distribution of local estimates rather than just a centrality measure [17]. Also, we are aware that our simple correction formula overlooks the effect of curvature and noise. We tried to address the former with the choice of minimal neighborhood size ($k = 1$), thus the overestimation effect due to curvature is minimal. Additionally, the effect of noise on the estimates is yet to be investigated. There are several strategies to alleviate noise effects such as undersample the data while keeping the neighborhood fixed [18], or using a bigger neighborhood size, while keeping the sample size fixed. Both of these procedures make the effect of curvature more severe, which makes the dimension estimation of noisy curved data a challenging task.

We benchmarked the new mFSA and corrected-mFSA method against Levina-Bickel estimator and

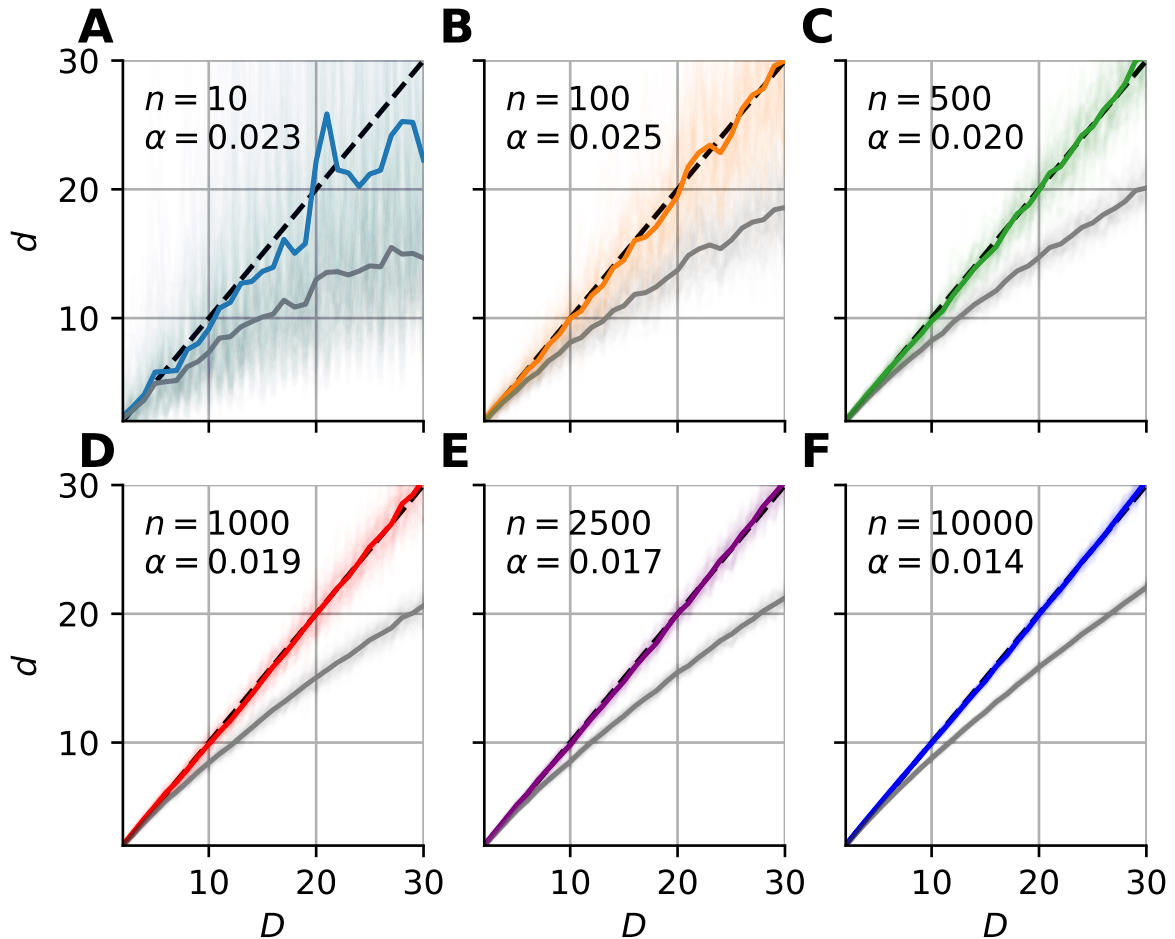


Figure 5: **Bias-correction of the median-FSA estimator for uniformly sampled unit hypercubes with various sample sizes ($k = 1$).** Subplots **A-F** show the mean of median-FSA estimator (grey line) values from $N = 100$ realizations (shading) of uniformly sampled unit hypercubes. The boundary condition is hard, so the edge effect makes under-estimation more severe. The colored lines show the corrected estimates according to the $\hat{w}_c = \hat{w} \exp(\alpha \hat{w})$.

DANCo on synthetic benchmark datasets and found that cmFSA showed comparable performance to DANCo. For many datasets, R-DANCo overestimated the intrinsic dimensionality, which is most probably due to rough default calibration[20]; the matlab implementation showed the best overall results in agreement with Campadelli et al[11]. This superiority was however dataset-specific: cmFSA performed genuinely the best in 4, DANCo in 2 out of the 15 benchmark datasets (with 7 ties, Table 1). Also, cmFSA showed better overall error rate than DANCo. Combining the performance measured by different metrics, we recognise that cmFSA found the true intrinsic dimension of the data in more cases, but when mistaken, it makes relatively bigger errors compared with DANCo.

The mFSA algorithm revealed diverse changes in the neural dynamics during epileptic seizures. In normal condition, the gradient of dimension values on the cortical grid reflects the hierarchical organization of neocortical information processing[28]. During seizures, this pattern becomes disrupted pointing to the breakdown of normal activation routes. Some channels showed lower dimensional dynamics during seizures; that behaviour is far from the exception: the decrease in dimensionality is due to widespread synchronization events between neural populations[29], a phenomenon reported by various authors [4, 30, 31]. These lower-dimensional areas are possible causal sources[7, 10, 9] and candidates for being the seizure onset zone. Interestingly, Esteller et al found, that the Higuchi fractal dimension values were higher at seizure onset and decreased to lower values as the seizures evolved over time[32]. We found, that most areas showed decreased dimensionality, but few areas also showed increased dimension values as seizure takes place. This may suggest that new - so far unused - neural circuits are activated at seizure onset; whether this circuitry contributes to or counteracts epileptic

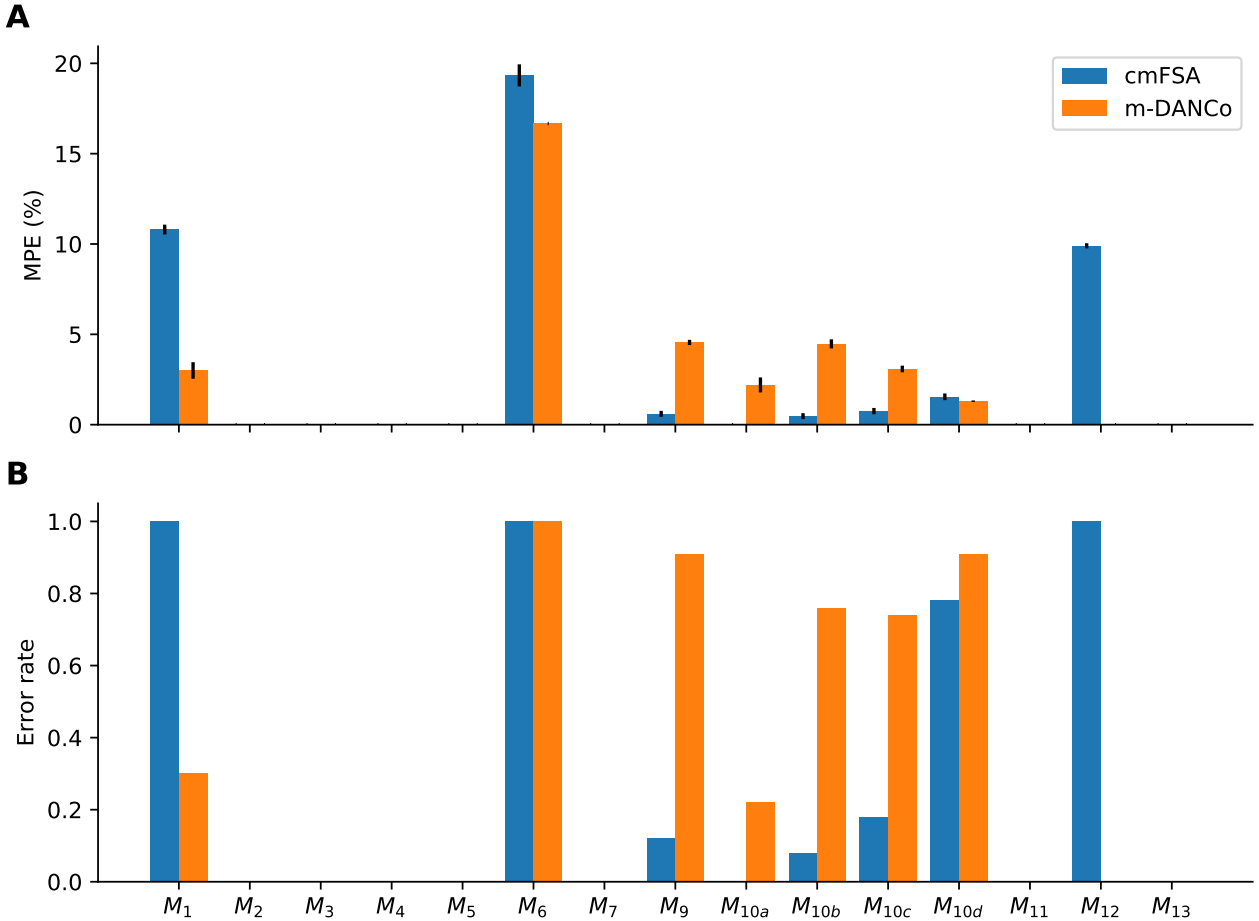


Figure 6: **Performance-comparison between cmFSA and DANCo on synthetic benchmark datasets.** **A** Dataset-wise Mean Percentage Error (MPE) on benchmark data. cmFSA (blue) shows smaller MPE in 4 cases and bigger MPE in 4 cases compared with DANCo (matlab). **B** Dataset-wise error rate for cmFSA and DANCo. cmFSA shows smaller error rates in 5 cases and bigger error rates in 2 cases compared with DANCo.

seizure is unclear.

Methods

The simulations and the FSA algorithms were implemented in python3[33] using the numpy[34], scipy[35] and matplotlib[36] packages, unless otherwise stated.

Simulations

We generated test-datasets by uniform random sampling from the unit D -cube to demonstrate, that theoretical derivations fit to data. We measured distances with a circular boundary condition to avoid edge effects, hence the data is as close to the theoretical assumptions as possible.

To illustrate the probability density function of the FSA estimator, we computed the local FSA intrinsic dimension values (Fig.1). We generated d -hypercubes ($n = 10000$, one realization) with dimensions of 2, 3, 5, 8, 10 and 12, then computed histograms of local FSA estimates for three neighborhood sizes: $k = 1, 11, 50$ respectively (Fig.1 A-F). We drew the theoretically computed pdf to illustrate the fit.

To show that the theoretically computed sampling distribution of the mFSA fits to the hypercube datasets, we varied the sample size ($n = 11, 101, 1001$) with $N = 5000$ realizations from each. We

Table 1:

Dimension estimates on synthetic benchmark datasets.

The table shows true dimension values, median-Farahmand-Szepesvári-Audibert, Maximum Likelihood, corrected median Farahmand-Szepesvári-Audibert and DANCo mean estimates from $N = 100$ realizations. The MPE values can be seen in the bottom line, the matlab version of DANCo estimator produced the smallest error followed by the cmFSA estimator.

	dataset	d	mFSA	cmFSA _{fr}	cmFSA	R-DANCo	M-DANCo _{fr}	M-DANCo	Levina
1	M_1	10	9.09	11.19	11.08	12.00	10.42	10.30	9.40
2	M_2	3	2.87	3.02	3.00	3.00	2.90	3.00	2.93
3	M_3	4	3.83	4.14	4.00	5.00	3.84	4.00	3.86
4	M_4	4	3.95	4.29	4.00	5.00	3.92	4.00	3.92
5	M_5	2	1.97	2.00	2.00	2.00	1.98	2.00	1.99
6	M_6	6	6.38	7.38	7.16	9.00	6.72	7.00	5.93
7	M_7	2	1.95	1.98	2.00	2.00	1.96	2.00	1.98
8	M_9	20	14.58	20.07	20.10	19.13	19.24	19.09	15.56
9	M_{10a}	10	8.21	9.90	10.00	10.00	9.56	9.78	8.64
10	M_{10b}	17	12.76	16.95	16.96	16.01	16.39	16.24	13.60
11	M_{10c}	24	16.80	24.10	24.06	23.15	23.39	23.26	18.05
12	M_{10d}	70	35.64	69.84	69.84	71.52	71.00	70.91	40.12
13	M_{11}	2	1.97	2.00	2.00	2.00	1.97	2.00	1.98
14	M_{12}	20	15.64	21.96	21.98	21.03	20.88	20.00	17.26
15	M_{13}	1	1.00	0.96	1.00	1.00	1.00	1.00	1.00
	MPE		13.58	4.73	2.89	10.07	3.39	2.35	10.81

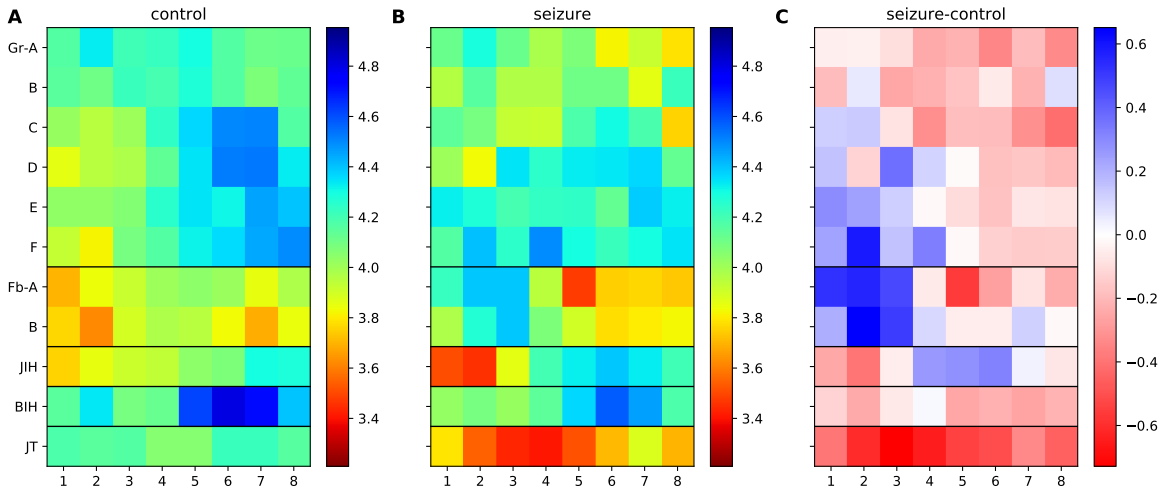


Figure 7: **mFSA Dimension estimates on intracranial Brain-LFP measurements during interictal activity and epileptic seizures.** The figure shows the dimension estimates on an intracranial cortical grid (Gr A-F), a smaller Frontobasal grid (Fb A, B) and 3 electrode strips with hippocampal and temporal localization (JIH, BIH, JT). The areas with lower-dimensional dynamics are marked by stronger colors. **A** Average of mFSA dimension values from interictal LFP activity ($N=16$, $k=10-20$). **B** Average of mFSA dimension values from seizure LFP activity ($N=18$, $k=10-20$). **C** Difference of dimension values. Stronger red color marks areas, where the dynamics during seizure was smaller-dimensional than its interictal counterpart. However, stronger blue indicates electrodes, where the during-seizure dynamics was higher dimensional than the interictal dynamics.

computed the mFSA for each realization and plotted the results for $d = 2$ (Fig. 2 A) and $d = 5$ (Fig. 2 B).

We investigated the dimensionality and sample-size effects on mFSA estimates (Fig. 3 A-F).

We simulated the hypercube data in the 2-30 dimension-range, and applied various sample sizes: $n = 10, 100, 1000, 2500, 10000$; one hundred realizations each ($N = 100$). We computed the mFSA values with minimal neighborhood size ($k = 1$), and observed finite-sample-effects, and asymptotic convergence. The finite sample effect was pronounced at low sample sizes and high dimensions, but we experienced convergence to the real dimension value as we increased sample size. We repeated the analysis with hard boundary conditions.

We fitted a correction formula on the logarithm of dimension values and estimates with the least squares method (Eq. 23), using all 100 realizations for each sample sizes separately.

$$\alpha = \frac{\sum (\ln E_i) d^{(i)}}{\sum (d^{(i)})^2} \quad (23)$$

Where $E_i = D_i/d^{(i)}$ is the relative error, D_i is the intrinsic dimension of the data, and $d^{(i)}$ are the corresponding mFSA estimates. This procedure fit well to data in the intrinsic dimension range 2-30 (Fig. 5 A-F).

Wider range of intrinsic dimension values (2-80) required more coefficients in the polynomial fit procedure (SFig. 1 A). Also, we used orthogonal distance regression to fit the mean over realizations of $\ln E_i$ with the same D_i value (SFig. 1 B). We utilized the mean and standard deviation of the regression error to compute the error rate of cmFSA estimator, if the error-distributions are normal (SFig. 1 C-D). We applied this calibration procedure ($n = 2500$) to compute cmFSA on the following benchmark datasets.

Comparison on synthetic benchmark datasets

We simulated $N = 100$ instances of 15 manifolds ($M_i, n = 2500$) with various intrinsic dimensions (see Table 1, 2, 4 in Campadelli et al.[11], <http://www.mL.uni-saarland.de/code/IntDim/IntDim.htm>).

We measured the performance of the mFSA and corrected-mFSA estimators on the benchmark datasets, and compared them with the performance of ML[12] and DANCo[17] estimators. We used the matlab[37, 38](<https://github.com/cran/intrinsicDimension>) and an R package[20] implementation of DANCo.

To quantify the performance we adopted the Mean Percentage Error (MPE, Eq.24) metric[11]:

$$\text{MPE} = \frac{100}{MN} \sum_{j=1}^M \sum_{i=1}^N \frac{|D_j - \hat{d}_{ij}|}{D_j} \quad (24)$$

Where there is N realizations of M types of manifolds, d_j are the true dimension values, \hat{d}_{ij} are the dimension estimates.

Also, we used the error rate – the fraction of cases, when the estimator did not find (missed) the true dimensionality – as an alternative metric.

We found that the corrected-mFSA estimator produced the second smallest MPE and the smallest error rate on the test datasets (Fig. 6).

Dimension estimation of interictal and epileptic dynamics

We used data of intracranial field potentials from two subdural grids positioned – parietofrontally and frontobasally – on the brain surface and from three strips located in the left and the right hippocampus and in the right temporal cortex as part of presurgical protocol for a subject with drug resistant epilepsy. This equipment recorded extracellular field potentials at 88 neural channels at a sampling rate of 2048 Hz. Moreover, we read in – using the neo package[39]– selected 10 second long chunks of the recordings from interictal periods ($N = 16$) and seizures ($N = 18$) to further analysis.

We standardised the data series and computed the Current Source Density (CSD) as the second spatial derivative of the recorded potential. We rescaled the 10 second long signal chunks by subtracting the mean and dividing by the standard deviation. Then, we computed the CSD of the signals by applying the graph Laplacian operator on the time-series. The Laplacian contains information about

the topology of the electrode grids, to encode this topology, we used von Neumann neighborhood in the adjacency matrix. After CSD computation, we bandpass-filtered the CSD signals[40] (1-30 Hz, fourth order Butterworth filter) to improve signal to noise ratio.

We embedded CSD signals and subsampled the embedded time series. We used an iterative manual procedure to optimize embedding parameters (SFig. 2). Since the fastest oscillation is (30 Hz) in the signals, a fixed value with one fourth period ($2048/120 \approx 17$ samples) were used as embedding delay. We inspected the average space-time separation plots of CSD signals to determine a proper subsampling, (with embedding dimension of $D=2$ (Fig. 7 A). We found, that the first local maximum of the space-time separation was at around 5 ms: 9 – 10, 10 – 11, 11 – 12 samples for the 1%, 25%, 50% percentile contour-curves respectively. Therefore, we divided the embedded time series into 10 subsets to ensure the required subsampling. Then, we embedded the CSD signal up to $D = 12$ and measured the intrinsic dimensionality for each embeddings (Fig. 7 B C). We found that intrinsic dimension estimates started to show saturation at $D \geq 3$, therefore we chose $D = 7$ as a sufficiently high embedding dimension (averaged over $k = 10 - 20$ neighborhood sizes).

We measured the intrinsic dimensionality of the embedded CSD signals using the mFSA method during interictal and epileptic episodes (Fig. 7). We selected the neighborhood size between $k = 10$ and $k = 20$ and averaged the resulting estimates over the neighborhoods and subsampling realizations. We investigated the dimension values (Fig. 7 A B) and differences (Fig. 7 C) in interictal and in epileptic periods.

We found characteristic changes in the pattern of intrinsic dimensions during seizures, which may help to localize seizure onset zone.

Acknowledgments

We are grateful for m Zlatniczki for his comments on the manuscript.

Author contributions

Zsigmond Benk performed the analytical and numerical calculations and wrote the manuscript.

Marcell Stippinger corrected analytical calculations, wrote python code for numerical calculations and corrected the manuscript.

Roberta Rehus carried out exploratory data analysis and proofreading.

Daniel Fab, Boglrka Hajnal, Lornd Erss recorded the EEG data, helped with data analysis and contributed to the manuscript text.

Attila Bencze and Andrs Telcs had profound effect on the mFSA derivations and contributed to the manuscript.

Zoltn Somogyvri led the research, helped to interpret the results of data analysis and contributed to the text.

Funding

The research reported in this paper was supported by the BME NC TKP2020 grant of NKFIH Hungary, by the BME-Artificial Intelligence FIKP grant of EMMI (BME FIKP-MI/SC), by the National Brain Research Program of Hungary (NAP-B, KTIA_NAP_12-2-201), by the National Brain Project II, NRDIO Hungary, PATTERN Group, and by 2017-1.2.1-NKP-2017-00002 of NKFIH.

References

- [1] Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 1983.

- [2] M E Houle, E Schubert, and A Zimek. On the Correlation Between Local Intrinsic Dimensionality and Outlierness. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11223 LNCS:177–191, 2018.
- [3] Martin Dlask and Jaromir Kukal. Correlation Dimension Estimation from EEG Time Series for Alzheimer Disease Diagnostics. In *Proceedings of the International Conference on Bioinformatics Research and Applications 2017 - ICBRA 2017*, pages 62–65. ACM Press, 2017.
- [4] G. E. Polychronaki, P. Y. Ktonas, S. Gatzonis, A. Siatouni, P. A. Asvestas, H. Tsekou, D. Sakas, and K. S. Nikita. Comparison of fractal dimension estimation algorithms for epileptic seizure onset detection. *Journal of Neural Engineering*, 7(4), 2010.
- [5] Manish Sharma, Ram Bilas Pachori, and U. Rajendra Acharya. A new approach to characterize epileptic seizures using analytic time-frequency flexible wavelet transform and fractal dimension. *Pattern Recognition Letters*, 94:172–179, jul 2017.
- [6] U. Rajendra Acharya, S. Vinitha Sree, G. Swapna, Roshan Joy Martis, and Jasjit S. Suri. Automated EEG analysis of epilepsy: A review. *Knowledge-Based Systems*, 45:147–165, jun 2013.
- [7] Mahito Sugiyama and Karsten M. Borgwardt. Measuring statistical dependence via the mutual information dimension. *IJCAI International Joint Conference on Artificial Intelligence*, pages 1692–1698, 2013.
- [8] Simone Romano, Oussama Chelly, Vinh Nguyen, James Bailey, and Michael E. Houle. Measuring dependency via intrinsic dimensionality. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, number 4, pages 1207–1212. IEEE, dec 2016.
- [9] Zsigmond Benkő, Ádám Zlatniczki, Marcell Stippinger, Dániel Fabó, András Sólyom, Loránd Erőss, András Telcs, and Zoltán Somogyvári. Complete Inference of Causal Relations between Dynamical Systems. *arXiv*, aug 2018.
- [10] Anna Krakovská. Correlation dimension detects causal links in coupled dynamical systems. *Entropy*, 21(9), 2019.
- [11] P Campadelli, E Casiraghi, C Ceruti, and A Rozza. Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework. *Mathematical Problems in Engineering*, 2015:1–21, 2015.
- [12] Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*. Neural information processing systems foundation, 2005.
- [13] Zoubin Ghahramani and David Mckay. Comments on ‘Maximum likelihood estimation of intrinsic dimension’, 2005.
- [14] Mithun Das Gupta and Thomas S. Huang. Regularized maximum likelihood for intrinsic dimension estimation. *Uai*, 1(1), 2010.
- [15] A Rozza, G Lombardi, C Ceruti, E Casiraghi, and P Campadelli. Novel high intrinsic dimensionality estimators. *Machine Learning*, 89(1-2):37–65, 2012.
- [16] S Bassis, A Rozza, C Ceruti, G Lombardi, E Casiraghi, and P Campadelli. A Novel Intrinsic Dimensionality Estimator Based on Rank-Order Statistics. In Francesco Masulli, Alfredo Petrosino, and Stefano Rovetta, editors, *Clustering High-Dimensional Data*, pages 102–117, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.
- [17] Claudio Ceruti, Simone Bassis, Alessandro Rozza, Gabriele Lombardi, Elena Casiraghi, and Paola Campadelli. DANCo: An intrinsic dimensionality estimator exploiting angle and norm concentration. *Pattern Recognition*, 47(8):2569–2581, 2014.

- [18] Elena Facco, Maria D 'errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, (August):1–8, 2017.
- [19] Francesco Camastra and Alessandro Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(10):1404–1407, oct 2002.
- [20] Kerstin Johnsson, Charlotte Soneson, and Magnus Fontes. Low Bias Local Intrinsic Dimension Estimation from Expected Simplex Skewness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):196–202, jan 2015.
- [21] Oussama Chelly, Michael E. Houle, and Ken Ichi Kawarabayashi. Enhanced estimation of local Intrinsic Dimensionality using auxiliary distances. *NII Technical Reports*, (7), 2016.
- [22] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E Houle, Ken-ichi Kawarabayashi, and Michael Nett. Estimating Local Intrinsic Dimensionality. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, number Cd, pages 29–38, New York, New York, USA, 2015. ACM Press.
- [23] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E. Houle, Ken-ichi Kawarabayashi, and Michael Nett. Extreme-value-theoretic estimation of local intrinsic dimensionality. *Data Mining and Knowledge Discovery*, 32(6):1768–1805, nov 2018.
- [24] Laurent Amsaleg, Oussama Chelly, Michael E. Houle, Ken-ichi Kawarabayashi, Miloš Radovanović, and Weeris Treeratanajaru. Intrinsic Dimensionality Estimation within Tight Localities. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 181–189. Society for Industrial and Applied Mathematics, may 2019.
- [25] Amir Massoud Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. Manifold-adaptive dimension estimation. In *Proceedings of the 24th international conference on Machine learning - ICML '07*, pages 265–272. ACM Press, 2007.
- [26] Pierre Simon Laplace. Memoir on the Probability of the Causes of Events. *Statistical Science*, 1(3):364–378, aug 1986.
- [27] Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in \mathbb{R}^d . In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pages 289–296. ACM Press, 2005.
- [28] Satoshi Tajima, Toru Yanagawa, Naotaka Fujii, and Taro Toyozumi. Untangling Brain-Wide Dynamics in Consciousness by Cross-Embedding. *PLOS Computational Biology*, 11(11):e1004537, nov 2015.
- [29] Florian Mormann, Klaus Lehnertz, Peter David, and Christian E. Elger. Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients. *Physica D: Nonlinear Phenomena*, 144(3):358–369, 2000.
- [30] E.T. Bullmore, M.J. Brammer, P. Bourlon, G. Alarcon, C.E. Polkey, R. Elwes, and C.D. Binnie. Fractal analysis of electroencephalographic signals intracerebrally recorded during 35 epileptic seizures: evaluation of a new method for synoptic visualisation of ictal events. *Electroencephalography and Clinical Neurophysiology*, 91(5):337–345, nov 1994.
- [31] Niina Päivinen, Seppo Lammi, Asla Pitkänen, Jari Nissinen, Markku Penttonen, and Tapio Grönfors. Epileptic seizure detection: A nonlinear viewpoint. *Computer Methods and Programs in Biomedicine*, 79(2):151–159, aug 2005.

- [32] R. Esteller, G. Vachtsevanos, J. Echauz, T. Henry, P. Pennell, C. Epstein, R. Bakay, C. Bowen, and B. Litt. Fractal dimension characterizes seizure onset in epileptic patients. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, pages 2343–2346 vol.4. IEEE, 1999.
- [33] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.
- [34] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.
- [35] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.
- [36] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95, 2007.
- [37] MATLAB. *MATLAB version 9.8.0.1396136 (R2020a)*. The Mathworks, Inc., Natick, Massachusetts, 2020.
- [38] Gabriele Lombardi. Intrinsic dimensionality estimation techniques. *MATLAB Central File Exchange*, Retrieved July 16, 2020.
- [39] S. Garcia, D. Guarino, F. Jaillet, T.R. Jennings, R. Pröpper, P.L. Rautenberg, C. Rodgers, A. Sobolev, T. Wachtler, P. Yger, and A.P. Davison. Neo: an object model for handling electrophysiology data in multiple formats. *Frontiers in Neuroinformatics*, 8:10, February 2014.
- [40] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7:267, 2013.

Supplemental information

A Calculations for normalized distances

A.1 Distance density of the k -th nearest neighbor

Let's take $K - 1$ points in the unit D -sphere randomly, and we chose one with r distance from the center. This situation simulates a K -neighborhood, with normalized distances of $K - 1$ points from the center. The next formula tells us the probability that a selected point at r was the k th from the center.

$$P(k|r, K, D) = \binom{K-2}{k-1} r^{D(k-1)} (1-r^D)^{K-k-1} \quad (\text{S.1})$$

here r can take values from the $[0, 1]$ interval.

Moreover the probability density that there is a point at r radius is given by the following derivation formula:

$$p(r|D) = Dr^{D-1} \quad (\text{S.2})$$

If sampling process is independent, the pdf that a point is on the radius r from $K - 1$ points is the same and independent of sample size:

$$\begin{aligned} p(r|K-1, D) &= \sum_{j=1}^n \frac{1}{n} \underbrace{\int dr_1 \cdots \int_{i \neq j} dr_i \cdots \int dr_n}_{i \neq j} p(r_1, r_2, \dots, r_j = r, \dots, r_n | D) \\ &= \sum_{j=1}^n \frac{1}{n} \underbrace{\int \cdots \int}_{n-1} \prod_{i=1}^n Dr_i^{D-1} \underbrace{dr_i}_{i \neq j} = \frac{1}{n} \sum_{j=1}^n Dr_j^{D-1} = Dr^{D-1} \end{aligned} \quad (\text{S.3})$$

This is the prior pdf of distance, we assume uniform density in the n -sphere. This prior can be any density, we chose this specific form with respect to the maximum entropy principle and also for practical reasons.

From the previous two formulas, we can write up the joint mixed probability function:

$$p(k, r|K-1, D) = D \binom{K-2}{k-1} r^{Dk-1} (1-r^D)^{n-k} \quad (\text{S.4})$$

Also:

$$p(k|K-1, D) = \frac{1}{K-1} \quad (\text{S.5})$$

Using Bayes theorem, we derive the distance distribution of the k th neighbor:

$$p(r|k, K-1, D) = \frac{P(k|r, K-1, D)p(r|K-1, D)}{p(k|K-1, D)} \quad (\text{S.6})$$

$$= (K-1)D \binom{K-2}{k-1} r^{Dk-1} (1-r^D)^{K-k-1} \quad (\text{S.7})$$

$$= \frac{D}{B(k, K-k)} r^{Dk-1} (1-r^D)^{K-k-1} \quad (\text{S.8})$$

Where B is the beta function.

A.2 Maximum Likelihood estimation of intrinsic dimension

Given a dataset, we can use the Maximum Likelihood principle to estimate intrinsic dimensionality by using Eq. S.8. The dataset is $K - 1$ randomly sampled points inside a d -dimensional sphere. But first we have to express the likelihood function:

$$\mathcal{L}(D|X) = p(r_1, \dots, r_{K-1}|D) \quad (\text{S.9})$$

This expression can be factorized into a chain because $p(r_k|r_{k+1}, r_{k+2}, \dots, r_{K-1}) = p(r_k|r_{k+1})$ which is a Markov property of neighbor distances.

$$\mathcal{L}(D|X) = p(r_1, \dots, r_{K-1}|D) = \prod_1^{K-1} p(r_k|r_{k+1}, D) \quad (\text{S.10})$$

where $r_K = 1$.

$$p(r_k|r_{k+1}, D) = kD \left(\frac{r_k}{r_{k+1}} \right)^{kD-1} \frac{1}{r_{k+1}} \quad (\text{S.11})$$

So if we substitute back into the previous expression:

$$\begin{aligned} \mathcal{L}(D|X) &= p(r_1, \dots, r_n|D) = \prod_1^{K-1} p(r_k|r_{k+1}, d) \\ &= (K-1)! D^{K-1} \frac{r_1^{D-1}}{r_2^D} \frac{r_2^{2D-1}}{r_3^{2D}} \frac{r_3^{3D-1}}{r_4^{3D}} \cdots \frac{r_{K-1}^{(K-1)D-1}}{r_K^{(K-1)D}} \\ &= (K-1)! D^{K-1} \left(\prod_1^{K-1} r_k \right)^{D-1} \end{aligned} \quad (\text{S.12})$$

The log likelihood:

$$\log \mathcal{L}(D|X) = \left(\sum_1^{K-1} \log k \right) + (K-1) \log D + (D-1) \sum_1^{K-1} \log r_k \quad (\text{S.13})$$

We seek for extrema of the likelihood function:

$$\begin{aligned} \frac{\partial \log \mathcal{L}(D|X)}{\partial D} &\stackrel{!}{=} 0 \\ \frac{K-1}{D} + \sum_1^{K-1} \log r_k &\stackrel{!}{=} 0 \end{aligned} \quad (\text{S.14})$$

$$\boxed{d_{\text{ML}} = \frac{K-1}{-\sum_1^{K-1} \log r_k}} \quad (\text{S.15})$$

This latter formula is basically equivalent to the local Levina-Bickel ML intrinsic dimension estimator if $r_k = \frac{R_k}{R_K}$.

B Derivation of the pdf of the FSA estimator

The starting point of our derivation is the posterior density of r , computed in Section 1:

$$p(r|k, K-1, D) = \frac{D}{B(k, K-k)} r^{Dk-1} (1-r^D)^{K-k-1} \quad (\text{S.16})$$

We fill in $K = 2k$ to the previous expression:

$$p(r|k, 2k - 1, D) = \frac{D}{B(k, k)} r^{Dk-1} (1 - r^D)^{k-1} \quad (\text{S.17})$$

The pdf of w can be expressed from the pdf of r with simple intergal transform:

$$p(r|k, 2k - 1, D) dr = q(d) dd \quad (\text{S.18})$$

so

$$q(d) = p(r|k, 2k - 1, D) \left| \frac{dr}{dd} \right| \quad (\text{S.19})$$

To compute the above expression, we first express r as a function of d , then we compute the derivative. Afterwards we put the things together.

$$d = -\frac{\log 2}{\log r} \implies r = \exp\left(-\frac{\log 2}{d}\right) \implies \frac{dr}{dd} = \exp\left(-\frac{\log 2}{d}\right) \frac{\log 2}{d^2} \quad (\text{S.20})$$

And finally, we put together these parts to get the pdf of the FSA estimator (Fig. 1):

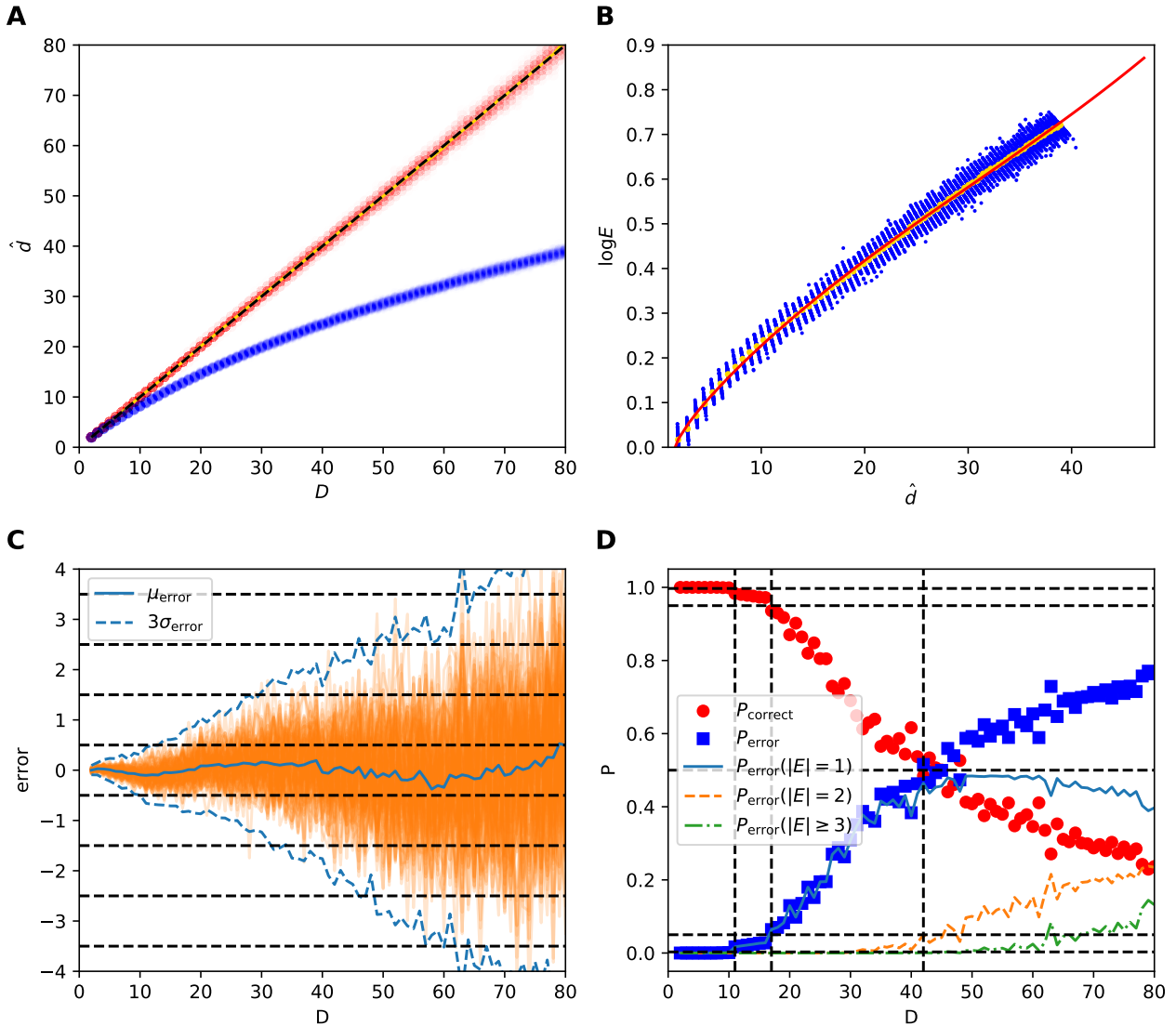
$$\begin{aligned} q(d|k, D) &= \frac{D}{B(k, k)} e^{(-\frac{\log 2}{d}(Dk-1))} \left(1 - e^{(-\frac{\log 2}{d}D)}\right)^{k-1} e^{(-\frac{\log 2}{d})} \frac{\log 2}{d^2} = \\ &= \boxed{\frac{D \log(2)}{B(k, k)} \frac{2^{-\frac{Dk}{d}} \left(1 - 2^{-\frac{D}{d}}\right)^{k-1}}{d^2}} \end{aligned} \quad (\text{S.21})$$

where $B(k, k)$ is the Euler beta function.

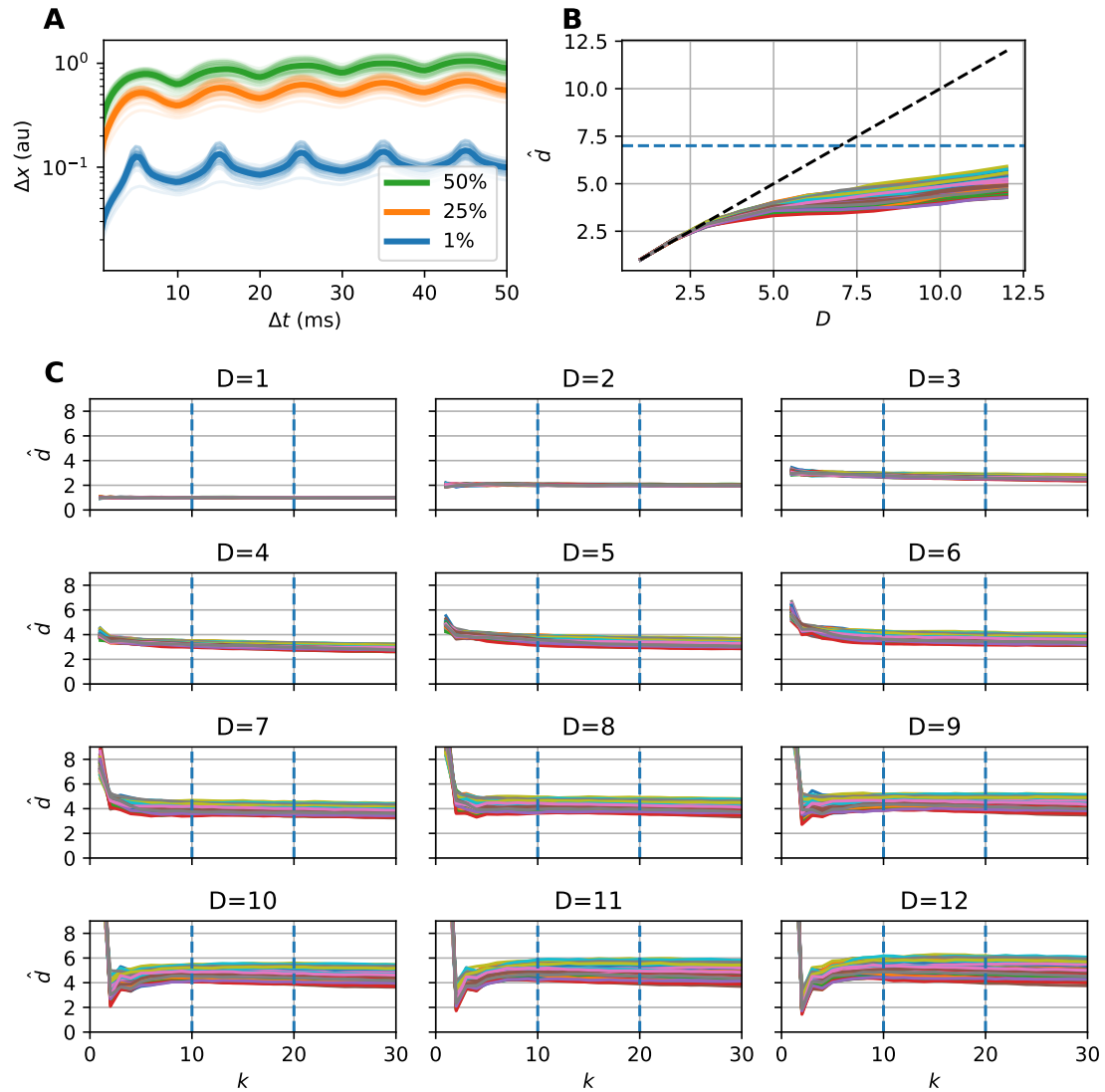
C Supplemental Figures and tables

STable 1: **Used symbols with interpretation.**

k	-	the order of the neighbor (increasing order as the distance from the center rises)
$K - 1$	-	number of points in the neighborhood
R	-	distance from center
r	-	normalized distance from center $r = R/R_K$ ($r \in [0, 1]$)
η	-	local density-dependent factor, approximately independent of R
D	-	intrinsic dimensionality of the space where the points are.
d	-	estimated intrinsic dimension value
P	-	Probability, probability mass function
p or q	-	probability density function (pdf)
n	-	sample size
N	-	number of realizations
B	-	Euler beta function



SFig. 1: **Calibration procedure for the $n = 2500$ datasets up to $D = 80$ ($k = 5$).** The figure shows the calibration procedure on 100 instances of uniformly sampled hypercubes. **A** Dimension estimates in the function of intrinsic dimensionality for the calibration hypercubes. The diagonal (dashed) is the ideal value, however the mFSA estimates (blue) show saturation because of finite sample and edge effects. cmFSA estimates (red) are also shown, with the mean (yellow) almost aligned with the diagonal. **B** The relative error (E) in the function of uncorrected mFSA dimension on semilogarithmic scale. The error-mFSA pairs (blue) lie on a short stripe for each intrinsic dimension value. The subplot also shows id-wise average points (yellow) and the polynomial fitting curve (red). **C** The error of cmFSA estimates in the function of intrinsic dimension on the calibration datasets. The mean error (blue line) oscillates around zero and the 99.7% confidence interval (blue dashed) widens as ID grows. The rounding switchpoints are also shown. **D** The probability that cmFSA hits the real ID of data, or misses by one, two or more as a function of ID on the calibration dataset.



SFig. 2: **Subsampling and embedding of the CSD signals.** **A** Mean Space-time separation plot of the CSD recordings, the lines show the contours of the 1% (blue), 25% (orange), and 50% (green) percentiles for the 34 - 16 interictal and 18 seizures - recordings (thin lines) and their average (thick line, $D = 2$). The first local maximum is at around 5 ms (10 time steps), which appoints the proper subsampling to avoid the effect of temporal correlations during the dimension estimation. **B** Intrinsic dimension in the function of the embedding dimension for the 88 recording-channels (averaged between $k = 5 - 10$, for the first seizure). Dimension-estimates deviate from the diagonal above $D = 3$, thus we chose $D = 2*3+1 = 7$ as embedding dimension. **C** Intrinsic dimension in the function of neighborhood size for various embedding dimensions (88 channels, for the first seizure). The dimension estimates are settled at the neighborhood size between $k=10 - 20$ (dashed blue). The knee because of the autocorrelation becomes pronounced for $D \geq 8$.