HAL open science

Analytical properties of UMAP dimensionality reductions

Nicolas Levernier, Hervé Rouault

To cite this version:

Nicolas Levernier, Hervé Rouault. Analytical properties of UMAP dimensionality reductions. 2023. hal-04046849

HAL Id: hal-04046849

https://hal.science/hal-04046849v1

Preprint submitted on 26 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analytical properties of UMAP dimensionality reductions

N. Levernier[1,2] and H. Rouault[1,2]

[1]*Aix-Marseille Université, CNRS, Inserm, INMED UMR 1249, Turing Centre for Living systems, Marseille, France*

[2]*Aix-Marseille Université, Université de Toulon, CNRS, CPT UMR 7332, Turing Centre for Living systems, Marseille, France*

Abstract

Dimensionality reduction techniques are essential tools for simplifying and interpreting high-dimensional datasets by mapping them to a lower-dimensional space. Among existing methods, Uniform Manifold Approximation and Projection (UMAP) has recently gained a huge popularity, being applied in contexts as diverse as transcriptomics, machine learning, image processing or thermodynamics. However, understanding of this method is still sparse. Here we gives analytical prediction for the UMAP projection of well defined datasets including pure noise, binary data and one-dimensional signal. For the latter, we uncover a phase transition in the lateral spreading of the projected points. We hope that our results could help improving data analysis using non-linear dimensionality reduction in various fields.

PACS numbers:

I. INTRODUCTION

Dimensionality reduction techniques are essential tools for simplifying and interpreting high-dimensional datasets by mapping them to a lower-dimensional space. Several popular algorithms exist, including principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) [1], isometric mapping (ISOMAP) [2], and Uniform Manifold Approximation and Projection (UMAP) [3]. Among them, UMAP (Uniform Manifold Approximation and Projection) has gained popularity due to its ability to process high-dimensional data faster and more accurately than other techniques.

UMAP is a newer algorithm that has gained popularity due to its ability to process high-dimensional data faster and more accurately than other non-linear dimensionality reduction techniques. It uses a combination of graph theory and manifold learning to identify the underlying structure in the data and project it into a lower-dimensional space. This makes it an ideal tool for visualizing complex data and identifying patterns or clusters that may be difficult to detect otherwise.

The power and limitations of UMAP have been studied in several research papers, and it has been shown to be highly effective in a variety of applications, us diverse as transcriptomic profiles [4, 5], evolution of microglial morphology [6] or pan-viral interactome of Sars-COV2 [7] in biology, and free-energy landscapes [8] and phase transitions in physics [9]. However, there is still much to learn about the properties of UMAP, especially in well-defined conditions. Therefore, in this article, we propose to bridge this gap by studying UMAP's properties analytically under three well-defined conditions: pure Gaussian noise, discrete clusters, and one-dimensional datasets.

Through this study, we aim to provide a deeper understanding of the properties of UMAP's dimensionality reduction, and how it can be used to improve data analysis in various fields. By investigating the systematic properties of the dimensionality reduction obtained through UMAP, we can better understand its strengths and limitations and how to optimize its performance in different applications. Among our analytical results, we uncover phase transitions in the UMAP projection of one-dimensional signal, of different nature from transitions observed when using PCA [10].

A. Optimization function

As stated in the article presenting UMAP [3], the algorithm can be seen as an optimization problem with the following function:

$$\mathcal{L}_{\text{UMAP}} = -\sum_{i\neq j} v_{ij}\log(w_{ij}) + (1-v_{ij})\log(1-w_{ij}) \tag{1}$$

In the latter, the weights v_{ij} measures the proximity of the points in the initial space, whereas w_{ij} measures the proximity in the space of embedding. In [3], the authors have chosen

$$v_{ij} = 2\exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j|| - \rho}{\sigma}\right) - \exp\left(-2\,\frac{||\mathbf{x}_i - \mathbf{x}_j|| - \rho}{\sigma}\right) \tag{2}$$

$$w_{ij} = \frac{1}{1 + a||\mathbf{y}_i - \mathbf{y}_j||^{2b}} \tag{3}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote respectively the position of point i in the initial and embedding space. The optimisation is then performed on the embedded positions $\mathbf{y}$.

Later, it has been seen that the actual optimization problem UMAP solves is slightly different [11]. Here however, we will rely on the optimization function 1.

II. DIMENSIONALITY REDUCTION OF PURE NOISE DATA AND CONNECTION TO RANDOM MATRIX THEORY

A. No disorder in the coupling constants

First, we consider a dataset in which $v_{ij} = v$ is the same for all pairs of points (all pairs are at the same distance relative to each other, which is asymptotically true for a gaussian cloud in high dimension). In such case, the function to minimize is:

$$\mathcal{L}_{\text{UMAP}} = -\sum_{i\neq j} v\log(w_{ij}) + (1-v)\log(1-w_{ij}) \tag{4}$$

$$= \sum_{i\neq j} v\log\left(\frac{1-w_{ij}}{w_{ij}}\right) - \log(1-w_{ij}) \tag{5}$$

We focus on the large N limit, so that we can replace the sum by an integral with the introduction of the density ρ of points in the low dimensional space. With such a manipulation, one should add a self-interaction energy which turns out to be negligible in the thermodynamic limit in the presence of long-range interactions [12, 13].

Let assume a projection in dimension $d = 2$. Using complex notation $z = \mathbf{y}^x + i\mathbf{y}^y$, one has to minimize over ρ:

$$\mathcal{L}_{\text{UMAP}} = \iint dz dz' \rho(z)\rho(z') \left[(v-1)\log(a|z-z'|^{2b}) + \log(1+a|z-z'|^{2b})\right] + \lambda \left(\int dz \rho(z) - 1\right) \tag{6}$$

where the last term is a Lagrange mupltiplier enforcing the normalization of the density ρ. We assume that the integral runs on a compact support, that should be a disk by symmetry. Cancelling the functional derivative leads to:

$$0 = \int dz' \rho(z') \left[(v-1)\log(a|z-z'|^{2b}) + \log(1+a|z-z'|^{2b})\right] + \lambda \tag{7}$$

Let now apply the Laplacien $\partial_z \partial_{\bar{z}}$:

$$0 = (v-1)b\pi \int dz' \rho(z')\delta(z-z') + \int dz' \rho(z') \frac{ab|z-z'|^{2b}}{1+a|z-z'|^{2b-2}} \tag{8}$$

$$= (v-1)b\pi\rho(z) + \int dz' \rho(z') \frac{ab|z-z'|^{2b-2}}{1+a|z-z'|^{2b}} \tag{9}$$

We now assume that we can neglect the $a|z-z'|^{2b}$ expression in the denominator in the second term (we explain the validity afterwards) and we choose $b = 1$:

$$0 = (v-1)\pi\rho(z) + a \tag{10}$$

So, on the compact support, ρ is uniform with density $\rho(z) = \frac{a}{\pi(1-v)}$. We thus get the radius of the disk $R = \sqrt{(1-v)/a}$ (Fig. 1A-B). Interestingly, this problem turns to be the same than the distribution of eigenvalues of complex matrices with independent and identically distributed entries (the Girko's circular law) [14]. Surprisingly enough, the way we have recovered the law above is, to the best of our knowledge, new. Neglecting the term in the denominator is thus valid for v close to 1. If not, the attracting force is not strong enough and the cut-off in the repulsion plays a role. For $b \neq 1$, we expect the density to be not perfectly flat, but the size of the disk should scale the same way.

B. Disordered coupling constants

We now turn to the case of disordered coupling v_{ij}. This is the usual regime of use of UMAP, which substract the typical large distance between pairs of points -coming from

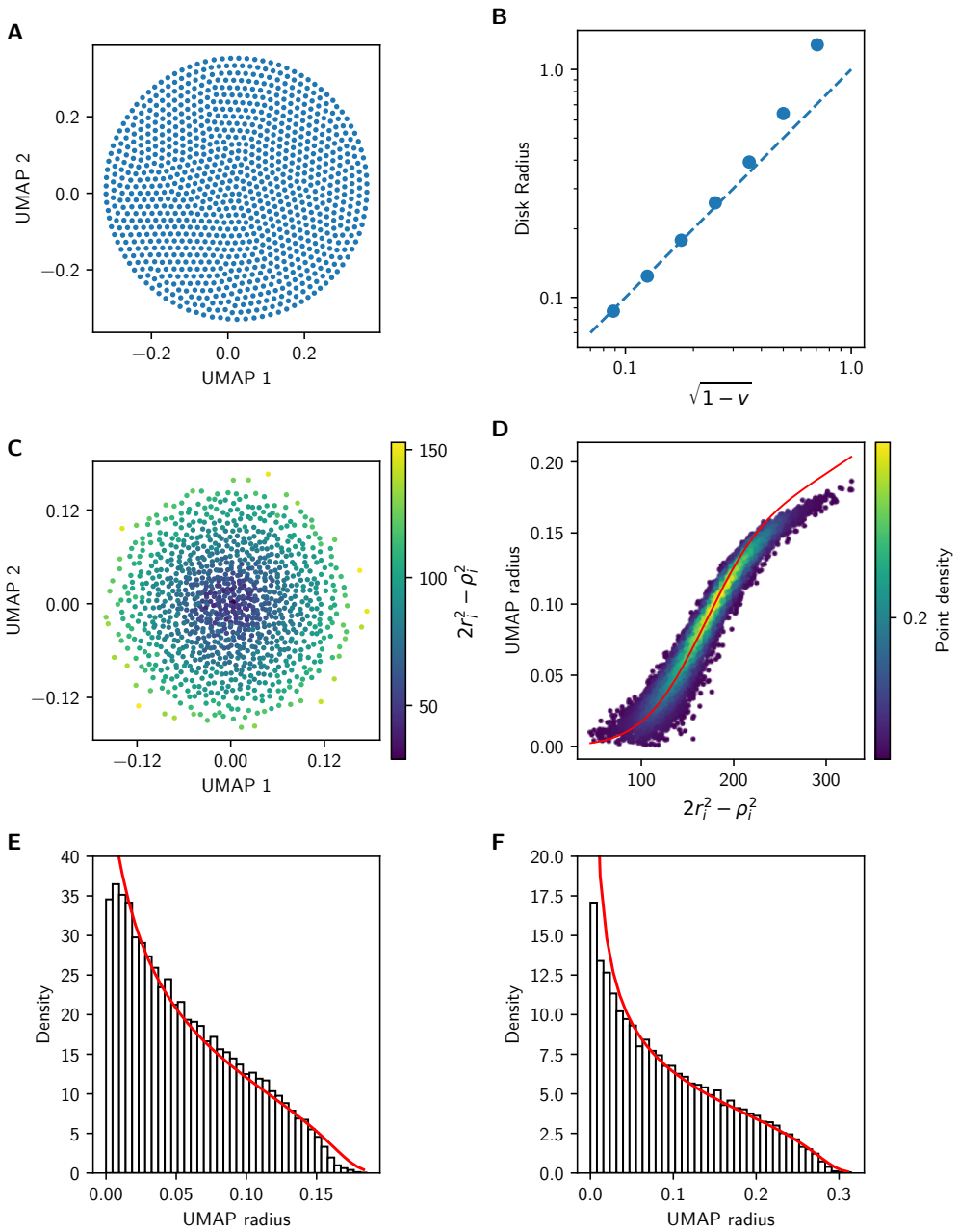

FIG. 1: UMAP dimensionality reduction for pure noise datasets. A. Example of a two dimensional UMAP dimensionality reduction for constant v. B. Radius of the dimensionality reduction as a function of $\sqrt{1-v}$, with simulation (dots) and our prediction (line). C. Same as A but for a gaussian input in D=512 dimensions with non-uniform ρ. D. Prediction of the mapping function $\mathcal{R}$ in the case of non uniform ρ. E. Comparison of our prediction of the radial density of the cloud (line) with simulations (histogram). F. Same as E but for uniform ρ.

dimensionality curse- to each distance. More precisely, we assume that points of the initial space are i.i.d, with any component being drawn from a normal distribution (up to a rescaling, we can choose the variance to be 1). In the large D limit, the distribution of radial squared distances is then normal with mean D and standard variance $3D$. We note r the *squared* radial distance. We also take ρ *uniform*. To compute $v(\mathrm{x}, \mathrm{x}')$, we first compute:

$$||\mathrm{x} - \mathrm{x}'||^2 = \sum_{i=1}^{D} x_i^2 + {x'}_i^2 + 2x_i{x'}_i \simeq r + r' \tag{11}$$

where the crossed term is negligible in the high D limit. Then, we get

$$v(r, r') = 2\exp\left(-\frac{\sqrt{r+r'} - \rho}{\sigma}\right) - \exp\left(-2\,\frac{\sqrt{r+r'} - \rho}{\sigma}\right) \tag{12}$$

$$= 1 - \frac{(\sqrt{r+r'} - \rho)^2}{\sigma^2} \tag{13}$$

after expanding for large σ sot that weights are close to 1. Noting $u = r - m$, $u' = r - m'$ and $\rho^2 = 2m - \delta$ we get (see Appendix):

$$1 - v(u, u') = \frac{4m + \delta^2 - 2\delta(u + u') + (u + u')^2}{8m\sigma^2} = A + B(u + u') + C(u + u')^2 \tag{14}$$

Indexing each point by its initial radial distance r, one has to minimize (after expansion of the repulsive term and $a = b = 1$):

$$\mathcal{L}_{\mathrm{UMAP}} = \iiiint dz dz' dr dr' \rho(z, r)\rho(z', r') \left[(v(r, r') - 1)\log(|z - z'|^2) + |z - z'|^2\right] \tag{15}$$

where ρ stands for the joint distribution in the initial and final space (we forget the Lagrange multiplier because we will directly enforce the normalisation). Note that the polar integration on radial distances r has been replaced by a cartesian integral because the distribution of r is peaked very far from $r = 0$ due to the large-D hypothesis.

In this disordered case, points are not equivalent anymore: points with large initial r will tend to be further of all the other points, and thus less attracted than average, whereas small r will tend to be attracted by more points (Fig 1C). Hence, it is quite understandable that after optimization in the low dimensional space, points with small initial r will be put in the center of the cloud and points with large initial r on the periphery. We thus expect a strong coupling between initial and final radial distance. To make the calculation feasible, we make the following ansatz:

$$\rho(z, r) = \rho_0(r)\delta(|z| - \mathcal{R}(r))(2\pi\mathcal{R}(r))^{-1} \tag{16}$$

meaning that any point with initial radial distance r will be placed at a radial distance $\mathcal{R}(r)$ in the space of embedding. $\rho_0(r)$ denotes the distribution of the initial distances r. The integration on z is now straightforward using $\int_0^{2\pi} \log(x^2 + y^2 - 2xy\cos\theta)d\theta = 2\pi \log(\max(x,y)^2)$:

$$\begin{aligned}\mathcal{L}_{\text{UMAP}} &= \iint drdr' 2\rho_0(r)\rho_0(r') \left[(v(r,r') - 1)\log(\max((\mathcal{R}(r)), \mathcal{R}(r'))) + \mathcal{R}(r)^2 + R(r')^2\right] \\ &= 4\int_{-\infty}^{\infty}\int_{-\infty}^{r} drdr'\rho_0(r)\rho_0(r')(v(r,r') - 1)\log(\mathcal{R}(r)) + 2\int_{-\infty}^{\infty} dr\rho_0(r)\mathcal{R}(r)^2 \qquad (17)\end{aligned}$$

We can now differentiate with respect to the unknown function $\mathcal{R}$, so that the optimal function satisfies:

$$\int_{-\infty}^{r} dr'\rho_0(r)\rho_0(r')(v(r,r') - 1)\frac{1}{\mathcal{R}(r)} + \rho_0(r)\mathcal{R}(r) = 0 \qquad (18)$$

The solution is thus

$$\mathcal{R}(r) = \sqrt{\int_{-\infty}^{r} (1 - v(r,r'))\rho_0(r')dr'} \qquad (19)$$

It is straightforward to check that this expression is perfectly consistent with the previous case of uniform $v(r,r') = v$, that is a uniform distribution of points stacked on a disk of radius $\sqrt{1-v}$. Let now compute this for the case of $v(r,r')$ given by (14). We get:

$$\begin{aligned}\mathcal{R}(u)^2 &= \int_{-\infty}^{u} du' \frac{e^{-\frac{u^2}{2s^2}}}{\sqrt{2\pi}s}(A + B(u+u') + C(u+u')) \\ &= (A + Bu + C(u^2 + s^2))\Phi(u/s) - \frac{s(B + 3Cu)}{\sqrt{2\pi}}e^{-\frac{u^2}{2s^2}} \qquad (20)\end{aligned}$$

where Φ is the cumulative distribution function of standard gaussian variable. This result is in striking contrast with the previous one: the fact that v's are disordered lead to a cloud which is not anymore of compact support (even though most points are densely packed inside a circle of radius $B + Cs$).

We can finally compute the local density of the embedded cloud as a function of the radial distance d:

$$f(d) = \int dr \int_{|z|=d} dz\rho(r,z)/(2\pi d) = \frac{\rho_0(u)}{2\pi\mathcal{R}(u)\mathcal{R}'(u)} \qquad (21)$$

with $u = \mathcal{R}^{-1}(d)$ (see Fig. 1E). On the other hand, the cumulative function (the fraction of points at radial distance smaller than d) is simply:

$$C(d) = \int_0^{\mathcal{R}^{-1}(d)} \rho_0(r)\, dr \qquad (22)$$

Let now turn to the case where ρ is not uniform, but depends on the closest neighbour of each point. ρ_i^2 is correlated to r_i^2 and we write $\rho_i^2 = r_i^2 + m - \alpha + \delta_i = 2m - \alpha + u_i + \delta_i$, where δ_i is approximately normally distributed of mean zero and with zero correlation with r_i^2. The calculation is mostly similar to the previous one and one obtain:

$$1 - v(u_i, u_j) = \frac{4m + \alpha^2 - \alpha(u_i + u_j - \delta_i - \delta_j) + 3/2(u_i + u_j - \delta_i - \delta_j)^2}{8m\sigma^2} \tag{23}$$

Hence, the problem is now turned into the mapping of the variable $v_i = 2u_i - \delta_i$ from high to low dimension (See Fig. 1D and F).

III. DIMENSION REDUCTION OF BINARY NOISY DATA

We now turn to the use of UMAP as a clustering tool. We focus on the case of data without any noise. Hence, let consider the case where there are two populations of points, with attractive weights $v_{11} = v_{22} = v$ and $v_{12} = v - \delta v$. For $\delta v = 0$, all points are equivalent and we are back to the situation of the previous part without noise: the embedding lead to a disk of radius $R = \sqrt{1 - v}$. For non-zero δv, this disk will break in two hemispheres H_1 and H_2, one for each community, separated by a distance Δ (see Fig. 2A-C). We expect that at first order, the diameter of both hemispheres is the same than for $\delta v = 0$. We want to estimate this distance as a function of δv. One has to minimize the following quantity over Δ:

$$\mathcal{L} = \int_{H_1}\int_{H_2} dzdz' \left[(v_{12} - 1)\log|z - z'|^2 + |z - z'|^2\right] \tag{24}$$

Let note H_2^* the hemisphere shifted by $-\Delta$ so that the reunion of H_1 and H_2^* is a disk. We can now write:

$$\begin{aligned}\mathcal{L} &= \int_{H_1}\int_{H_2^*} dzdz' \left[(v_{12} - 1)\log(|z - z'|^2 - 2\Delta(x - x') + \Delta^2) + |z - z'|^2 - 2\Delta(x - x') + \Delta^2\right] \\ &= (v - \delta v - 1)\int_{H_1}\int_{H_2^*} dzdz' \left[\log|z - z'|^2 - \frac{2(x - x')}{|z - z'|^2}\Delta + \left(\frac{1}{|z - z'|^2} + \frac{2(x - x')^2}{|z - z'|^4}\right)\Delta^2\right] \\ &\quad + \int_{H_1}\int_{H_2^*} dzdz' \left[|z - z'|^2 - 2\Delta(x - x') + \Delta^2\right]\end{aligned} \tag{25}$$

with $z = (x, y)$, $z' = (x', y')$ and x chosen along the axis of shift. We know that for $\delta v = 0$, $\Delta = 0$ is a minimum. Thus:

$$\int_{H_1}\int_{H_2^*} dzdz' \left[(1 - v)\frac{(x - x')}{|z - z'|^2} - (x - x')\right] = 0 \tag{26}$$

Finally, at second order in Δ:

$$\mathcal{L} = C\Delta^2 + 2\delta v \int_{H_1}\int_{H_2^*} dzdz' \frac{(x-x')}{|z-z'|^2}\Delta \tag{27}$$

with $C = \int_{H_1}\int_{H_2^*} dzdz' \left[\frac{v-1}{|z-z'|^2}\left(1+2\frac{(x-x')^2}{|z-z'|^2}\right)+1\right] > 0$. This leads to a trancritical transition:

$$\Delta = \frac{4\pi R^3}{3C}\delta v = \frac{4\pi R}{3C'}\delta v \tag{28}$$

where we have used (26) to simplify the numerator and $C' = C/R^2$ is a pure number (see Fig 2D). Thus, as soon as δv is of order of unity, both clouds are very well separated, with distance between them comparable to their size. For the sake of completeness, we also give the scaling behaviour for very dissimilar communities ($\delta v \ll 1$): $\Delta \sim 1/\sqrt{\delta v}$.

IV. DIMENSION REDUCTION OF 1D CONTINOUS SIGNAL AND PHASE TRANSITION

A. No noise

Let first investigate the case of perfect signal, with no noise. Let $\{\mathrm{x}_N\}$ be the set of points in the ambient space of dimension D, and let note $\{\mathrm{y}_N\}$ their UMAP-projection in the space of dimension $d < D$. Here, we focus on the case where the points $\{\mathrm{x}_N\}$ lie perfectly on a segment of length 1, and are equally spaced. We can also assume $d = 1$, and we note $\mathrm{y}_i = g(\mathrm{x}_i) = g(i/N)$ for $0 \leq i < N$.

To compute the length of the obtained projection, we look for a solution $g(x) = \alpha x$ that minimizes $\mathcal{L}_{\mathrm{UMAP}}$, in the limit of small σ (i.e. for large α). Performing a change of variable and using the symmetry of the integrand, we have to compute:

$$\mathcal{L}_{\mathrm{UMAP}} = -2\int_0^1 dv \int_0^v duK(u)\log(a\alpha^{2b}u^{2b}) + 2\int_0^1 dv\int_0^v du \log\left(1+\frac{1}{a\alpha^{2b}u^{2b}}\right) \tag{29}$$

The first integral is easy to evaluate:

$$\begin{aligned} I_1 &= -2\int_0^1 dv\int_0^v duK(u)\log(a\alpha^{2b}u^{2b}) \\ &= -4b\log\alpha\int_0^1 dv\int_0^v duK(u) + C \\ &= (6b\sigma + O(\sigma^2))\log\alpha + C \end{aligned} \tag{30}$$

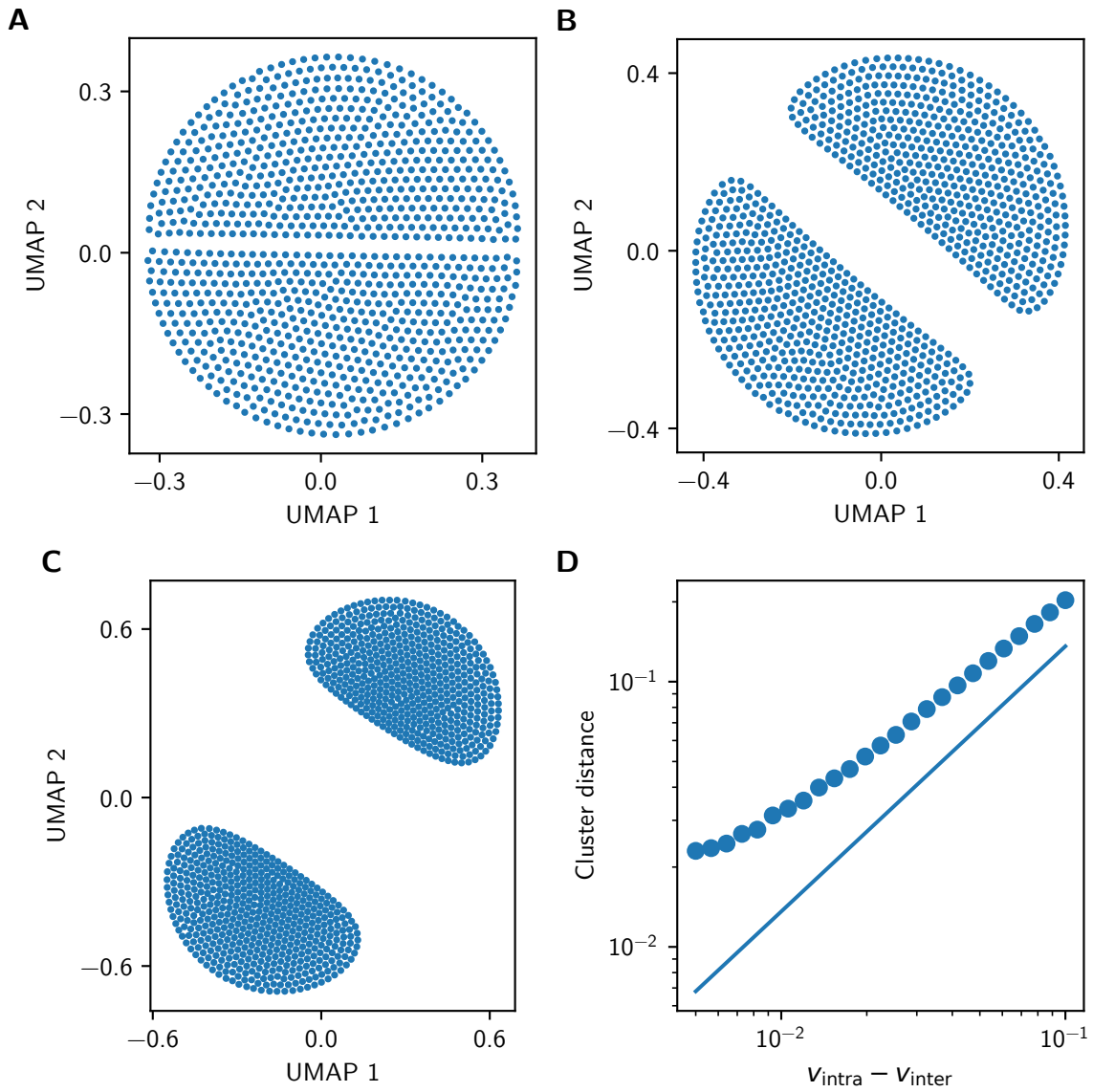

FIG. 2: UMAP dimensionality reduction for binary datasets. A, B and C. UMAP dimensionality reduction for binary signal with low, moderate and strong dissimilarity ($\delta_{\text{intra}} = 0.9$ and $\delta_{\text{inter}} = 0.89,\ 0.8,\ 0.5$ respectively. D. Distance between hemispheres in the low dissimilarity regime, with simulation (dots) and prediction (line).

where C does not depend on α.

The second part gives, for large α:

$$\begin{aligned} I_2 &= \frac{2}{\alpha a^{\frac{1}{2b}}} \int_0^{+\infty} du \log(1 + 1/u^{2b}) \\ &= \frac{2K}{\alpha a^{\frac{1}{2b}}} \end{aligned} \tag{31}$$

For UMAP default parameters, this integral is numerically evaluated to $K = 3.196$. Putting all together, deriving with respect to α to find the extremum, we obtain the length of the UMAP projected line:

$$\alpha = \frac{K}{3ba^{1/(2b)}\sigma} \tag{32}$$

B. With noise

We now assume that uncorrelated gaussian noise of amplitude $\epsilon/\sqrt{2}$ is added on each point (the $\sqrt{2}$ is only here to make the variance of the difference between two realisations of the noise to be ϵ). We are in interested in the regime $D \gg 1$ and $D\epsilon^2 \ll 1$ (so that the 1-d signal can be extracted from the noise). The Umap projection typically creates an elongated cloud of length α and thickness δ. Let first assume general σ and ρ. We focus in the limit of large α (which corresponds to small sigma, similarly to the no noise case).

To get an analytical estimate of α and δ, we will make some assumptions and approximation. The first one is that we can minimize the annealed functional, which is expected to be exact in the thermodynamic limit. The second one is that the density of the cloud along the transverse axis is gaussian. This is not exact but should give the same scaling with the problem parameters. Let note $m = D\epsilon^2$, $s = \sqrt{3D}\epsilon^2$.

First, after introducing η, the radial noise in the initial space $\eta = \sum_{i=1\dots D-1} \eta_i^2$, which is a gaussian v.a. of mean m and variance s in the large D limit, we get:

$$\begin{aligned}\mathcal{L} = &-\int_0^1 dx \int_0^x dy \int_{-\infty}^{\infty} \frac{e^{-\frac{(\eta-m)^2}{2s^2}}}{N} d\eta \int_{-\infty}^{\infty} \frac{e^{-\frac{\xi^2}{4\delta^2}}}{N'} d\xi \\ &[(-2e^{-\frac{\sqrt{y^2+\eta}-\rho}{\sigma}} + e^{-2\frac{\sqrt{y^2+\eta}-\rho}{\sigma}}) \log(a(\alpha^2 y^2 + \xi^2)^b) - \log\left(\frac{a(\alpha^2 y^2 + \xi^2)^b}{1 + a(\alpha^2 y^2 + \xi^2)^b}\right)\end{aligned} \tag{33}$$

For large α, we can neglect side effects and obtain a translation-invariant problem:

$$\begin{aligned}\mathcal{L} = &-\int_0^{\infty} dy \int_{-\infty}^{\infty} \frac{e^{-\frac{(\eta-m)^2}{2s^2}}}{N} d\eta \int_{-\infty}^{\infty} \frac{e^{-\frac{\xi^2}{4\delta^2}}}{N'} d\xi \\ &\left[(-2e^{-\frac{\sqrt{y^2+\eta}-\rho}{\sigma}} + e^{-2\frac{\sqrt{y^2+\eta}-\rho}{\sigma}}) \log(a(\alpha^2 y^2 + \xi^2)^b) - \log(\frac{a(\alpha^2 y^2 + \xi^2)^b}{1 + a(\alpha^2 y^2 + \xi^2)^b})\right]\end{aligned} \tag{34}$$

Now, we develop the expression in the limit of large noise. We have to compute $I = \int_{-\infty}^{\infty} d\eta \frac{e^{-\frac{(\eta-m)^2}{2s^2}}}{N} e^{-\frac{\sqrt{y^2+\eta}-\rho}{\sigma}}$. Let note $\rho = \sqrt{m} - \Delta\rho$. Within UMAP choice of parameters, we have $\Delta\rho = k\epsilon \log(ND^{1/4}\epsilon)$, but the following calculations is more general. In the limit we

are interested in, σ is always much smaller than m, so that:

$$I = \int_{-\infty}^{\infty} d\eta' \frac{e^{-\frac{\eta'^2}{2s^2}}}{N} e^{-\frac{y^2+\eta'+m-\rho^2}{\sigma(\sqrt{y^2+\eta'}+\rho)}} \tag{35}$$

$$\simeq e^{-\frac{y^2}{2\sigma\sqrt{m}}} e^{-\frac{\Delta\rho}{\sigma}} \int_{-\infty}^{\infty} d\eta' \frac{e^{-\frac{\eta'^2}{2s^2}}}{N} e^{-\frac{\eta'}{2\sigma\sqrt{m}}}$$

$$\simeq e^{-\frac{y^2}{2\sigma\sqrt{m}}} e^{-\frac{\Delta\rho}{\sigma}} e^{\frac{s^2}{8m\sigma^2}}$$

$$\simeq c e^{-\frac{y^2}{2\tilde{\sigma}^2}} \tag{36}$$

For the second exponential we get the same value by replacing $\tilde{\sigma}^2$ to $\tilde{\sigma}^2/2$ and c by c^2. We see that after averaging over the initial noise, we get an expression close to the one without noise : simply a change from σ to $\tilde{\sigma} = \sqrt{\sqrt{m}\sigma}$ and a gaussian kernel in place of an exponential one with prefactors c in front of exponential terms. Finally, one has to minimize:

$$\mathcal{L} = \int_0^{\infty} dy \int_{-\infty}^{\infty} d\xi \frac{e^{-\frac{\xi^2}{4\delta^2}}}{N'} \left[b(2ce^{-\frac{y^2}{2\tilde{\sigma}^2}} - c^2 e^{-\frac{y^2}{\tilde{\sigma}^2}}) \log(\alpha^2 y^2 + \xi^2) - \log\left(\frac{a(\alpha^2 y^2 + \xi^2)^b}{1 + a(\alpha^2 y^2 + \xi^2)^b} \right) \right] \tag{37}$$

Unfortunately, we cannot get explicit expressions with a gaussian kernel. We simplify a bit by replacing by a square of size $\tilde{\sigma}$ and amplitude $v = 2c - c^2$. We also restrict to $b = 1$ and $a = 1$ (which amounts to rescale length in low dimenion space). Thus:

$$\mathcal{L} = \frac{1}{\alpha} \int_0^{\infty} dy \int_0^{\delta} \frac{d\xi}{\delta} \left[vH(y/\alpha\tilde{\sigma}) \log(y^2 + \xi^2) - \log\left(\frac{y^2 + \xi^2}{1 + y^2 + \xi^2} \right) \right] \tag{38}$$

Let pose $S = \alpha\tilde{\sigma}$ and divide the former by $\tilde{\sigma}$. We now optimize on S and δ the function:

$$\mathcal{L} = \frac{1}{S} \int_0^{\infty} dy \int_0^{\delta} \frac{d\xi}{\delta} \left[vH(y/S) \log(y^2 + \xi^2) - \log\left(\frac{y^2 + \xi^2}{1 + y^2 + \xi^2} \right) \right] \tag{39}$$

Now we can compute everything. The first term with the Heaviside gives:

$$\mathcal{L}_{att} = v \left[\log(S^2 + \delta^2) + \frac{\delta}{S} \arctan \frac{S}{\delta} + \frac{S}{\delta} \arctan \frac{\delta}{S} \right] \tag{40}$$

The second one gives

$$\mathcal{L}_{rep} = -\frac{\pi}{2} \frac{\delta}{S} \left[1 - \sqrt{1 + 1/(2\delta^2)} - \mathrm{argsh}(\sqrt{2}\delta)/(2\delta^2) \right] \tag{41}$$

Now, it is not hard to see that if $v > 1$, then $\delta_{opt} = 0$. This is due to the fact that at short distance points always attract more then they repulse so that points with same signal

collapse. Then, we can easily compute $S = \frac{\pi}{2\sqrt{2}v}$ Then, we can get the critical behaviour close to $v = 1$. Let introduce $e = 1 - v$. At first order we get $\delta = \frac{3e}{2\sqrt{2}-2/(\pi S)}$. At this order, the value of S is unchanged. We have thus obtained a phase transition in the UMAP projection, with a transcritical bifurcation (due to the fact that there is no $\delta \leftrightarrow -\delta$ symmetry): for strong attraction v, one gets a line of strictly zero thickness, whereas for large attraction one obtains a thick elongated cloud.

We can estimate the aspect ratio of the cloud. Our results above are for general ρ and σ, but we will focus only to their values with UMAP default settings. Then, c is only controlled by $\Delta\rho$, so that:

$$\begin{aligned} a &= \delta/\alpha \\ &= \text{cst.}\tilde{\sigma}e \\ &= \text{cst.}\frac{\epsilon^{3/2}D^{1/4}}{\sigma^{1/2}} \end{aligned} \tag{42}$$

As for σ defined as the distance to the k-th neighbour minus ρ, we have two regimes:

$$\begin{aligned} \sigma &= d_k - \rho \\ &= \sqrt{(k/N)^2 + D\epsilon^2} - \sqrt{D}\epsilon + \Delta\rho \\ &= \Delta\rho + \frac{(k/N)^2}{\sqrt{D}\epsilon} \quad \text{for} \quad \sqrt{D}\epsilon \gg k/N \qquad (43) \\ &= k/N - \sqrt{D}\epsilon \quad \text{for} \quad \sqrt{D}\epsilon \ll k/N \qquad (44) \end{aligned}$$

The large noise regime can be split in two: either $\Delta\rho \ll \frac{(k/N)^2}{\sqrt{D}\epsilon}$ and $\sigma = \frac{(k/N)^2}{\sqrt{D}\epsilon}$ or $\Delta\rho \gg \frac{(k/N)^2}{\sqrt{D}\epsilon}$ and $\sigma = \Delta\rho \simeq k\epsilon$. Finally, we get:

$$\begin{aligned} a &= \text{cst.}\frac{\epsilon D^{1/4}}{\sigma^{1/2}} \quad \text{for} \quad \sqrt{D}\epsilon \gg k/ND^{1/4} \qquad (45) \\ &= \text{cst.}\frac{\epsilon^2 D^{1/2}}{k/N} \quad \text{for} \quad k/N \ll \sqrt{D}\epsilon \ll k/ND^{1/4} \qquad (46) \\ &= \text{cst.}\frac{\epsilon^{3/2}D^{1/4}}{k/N} \quad \text{for} \quad \sqrt{D}\epsilon \ll k/N \qquad (47) \end{aligned}$$

V. DISCUSSION

Our analytical study of UMAP dimensionality reductions brings insight into the structure of the obtained patterns. By considering the limit of a large number of points, we were able

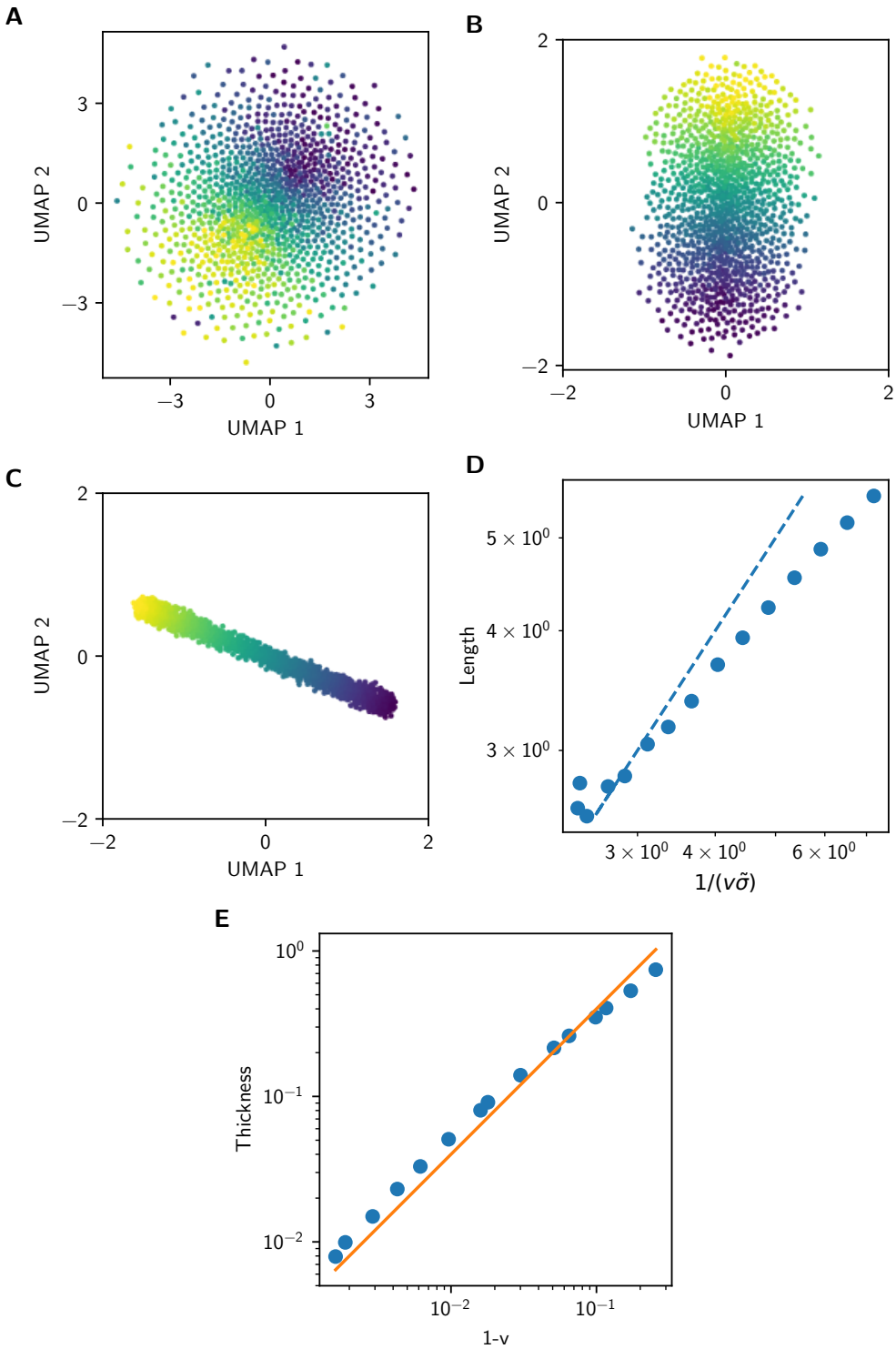

FIG. 3: UMAP dimensionality reduction for 1D continuous signal datasets. A, B, C. Example of a two dimensionality reductions for a noise amplitude equal to 0.05, 0.02, 0.005 respectively. B. Radius of the dimensionality reduction as a function of $\sqrt{1-v}$, with simulation (dots) and our prediction (line). D. Length of the dimensionality reduction (points: simulation, line: prediction). E. Thickness of the dimensionality reduction (points: simulation, line: prediction).

to use mean-field approaches and explore the continuous density of points. For instance, a pure Gaussian noisy input (without any structure) produces a cloud of size of order one with nearly uniform density. This should prompt users of this technique to pay attention to the size of the obtained reductions.

At the same time, the discrete cluster analysis shows that the technique is highly sensitive to minor differences between clusters. This effect results from the long-range interactions between dimension reduction points, which are specific to UMAP. While the entropic effect has a negligible effect on the average density, it remains present and has a strong impact on cluster segregation.

Finally, our mean-field approximation can predict the shape (length and thickness) of clouds resulting from the dimensionality reduction of high-dimensional signals containing only a one-dimensional manifold structure. We were able to characterize a transcritical phase transition as a function of the interaction strength, v, that explains the effectiveness of UMAP in revealing such structures.

We believe that our analysis provides a strong foundation for assessing the significance of UMAP dimensionality reduction in the presence of high levels of noise, which is often the case in realistic data analysis scenarios. Furthermore, this mean-field approach can extend beyond the simple scenarios that we have considered here.

Acknowledgments

The project leading to this publication has received funding from the "Investissements d'Avenir" French Government program managed by the French National Research Agency (ANR-16-CONV-0001) and from Excellence Initiative of Aix-Marseille University — A*MIDEX.

Appendix A: Calculation of $v(u, u')$ in the pure noise case

In the case of uniform ρ, the calculation is straightforward. Let first note $d_{ij}^2 = 2m + \epsilon_i j$. By expanding the exponential terms for large σ with the assumption that δ and ϵ_{ij} are small

compared to m, we obtain:

$$\begin{aligned}
&\sigma^2(1-v) = (d_{ij}-\rho)^2 \\
&= \frac{1}{8m}(\epsilon_{ij}^2 + \delta^2 - 2\epsilon_{ij}\delta) \\
&= \frac{1}{8m}(d_{ij}^4 + \delta^2 - (4m+2\delta)d_{ij}^2 + (2m+\delta)^2) \\
&= \frac{4m + \delta^2 - 2\delta(u+u') + (u+u')^2}{8m}
\end{aligned} \tag{A1}$$

This is the equation (14) in the main text.

The calculation for non-uniform is similar, except that one has to introduce a δ depending on the considered point δ_i.

Appendix B: Code availability

As noted previously, our implementation of UMAP stictly follows the optimization function 1, which is different from the initial implementation [3]. Our implementation has been coded in the Rust programming language and is available at `https://gitlab.com/rouault-team-public/analysis/umaprs`.

[1] L. Van der Maaten and G. Hinton, Journal of machine learning research **9** (2008).

[2] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, science **290**, 2319 (2000).

[3] L. McInnes, J. Healy, and J. Melville, 1802.03426, URL `http://arxiv.org/abs/1802.03426`.

[4] M. W. Dorrity, L. M. Saunders, C. Queitsch, S. Fields, and C. Trapnell, Nature Communications **11**, 1537 (2020).

[5] Y. Yang, H. Sun, Y. Zhang, T. Zhang, J. Gong, Y. Wei, Y.-G. Duan, M. Shu, Y. Yang, D. Wu, et al., Cell Reports **36** (2021).

[6] G. Colombo, R. J. A. Cubero, L. Kanari, A. Venturino, R. Schulz, M. Scolamiero, J. Agerberg, H. Mathys, L.-H. Tsai, W. Chachólski, et al., Nature Neuroscience **25**, 1379 (2022).

[7] A. Ghavasieh, S. Bontorin, O. Artime, N. Verstraete, and M. De Domenico, Communications Physics **4**, 83 (2021).

[8] B. W. B. Shires and C. J. Pickard, Phys. Rev. X **11**, 041026 (2021), URL `https://link.aps.org/doi/10.1103/PhysRevX.11.041026`.

[9] A. Tirelli, D. O. Carvalho, L. A. Oliveira, J. de Lima, N. C. Costa, and R. R. dos Santos, The European Physical Journal B **95**, 189 (2022).

[10] T. Lesieur, F. Krzakala, and L. Zdeborová, in *2015 IEEE International Symposium on Information Theory (ISIT)* (2015), pp. 1635–1639.

[11] S. Damrich and F. A. Hamprecht, in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, edited by M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan (2021), pp. 5798–5809, URL `https://proceedings.neurips.cc/paper/2021/hash/2de5d16682c3c35007e4e92982f1a2ba-Abstract.html`.

[12] F. J. Dyson, Journal of Mathematical Physics **3**, 1199 (1962).

[13] D. S. Dean and S. N. Majumdar, Phys. Rev. E **77**, 041108 (2008), URL `https://link.aps.org/doi/10.1103/PhysRevE.77.041108`.

[14] M. L. Mehta, *Random matrices* (Elsevier, 2004).