Bayesian dimensionality reduction via identifications of data intrinsic dimensions

Antonietta Mira

Università della Svizzera italiana$^{\nabla}$ and University of Insubria

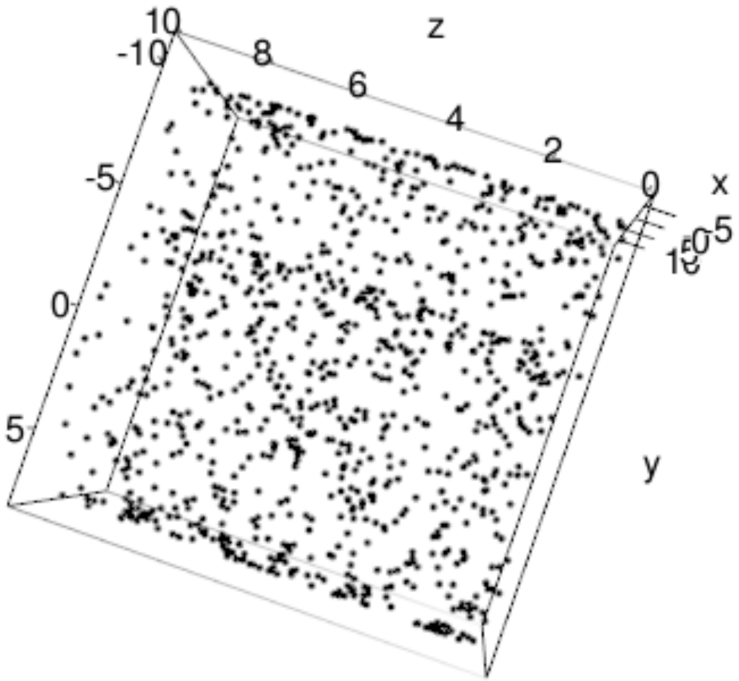
joint with

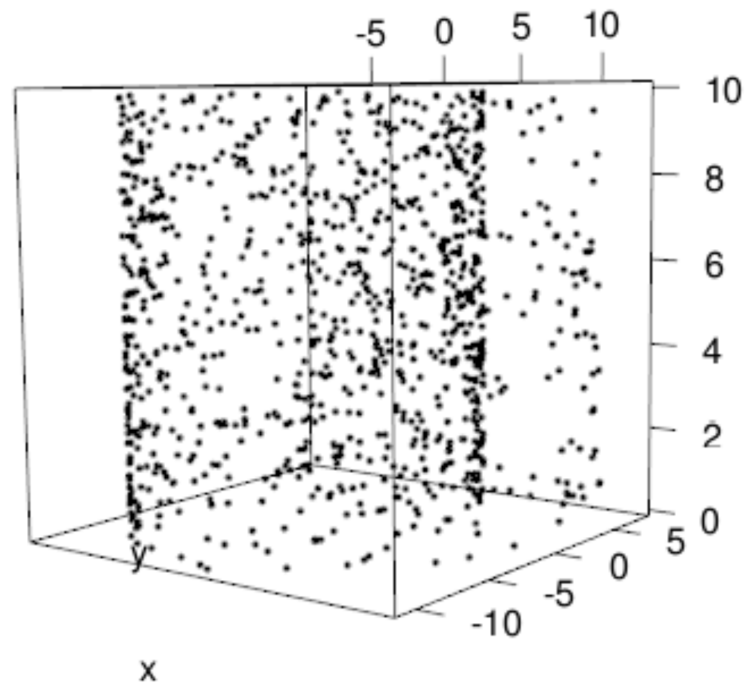
M. Allegra*, E. Facco*, A. Laio*, F. Denti$^{\nabla}$, M. Guindani$^{\diamond}$
SISSA*, Trieste, Italy and University of California$^{\diamond}$, Irvine
Basketball application joint with: E. Santos-Fernandez and K. Mergensen

March 8, 2019, WU Wirtschaftsuniversität Wien,
Institute for Statistics and Mathematics

A matter of perspective

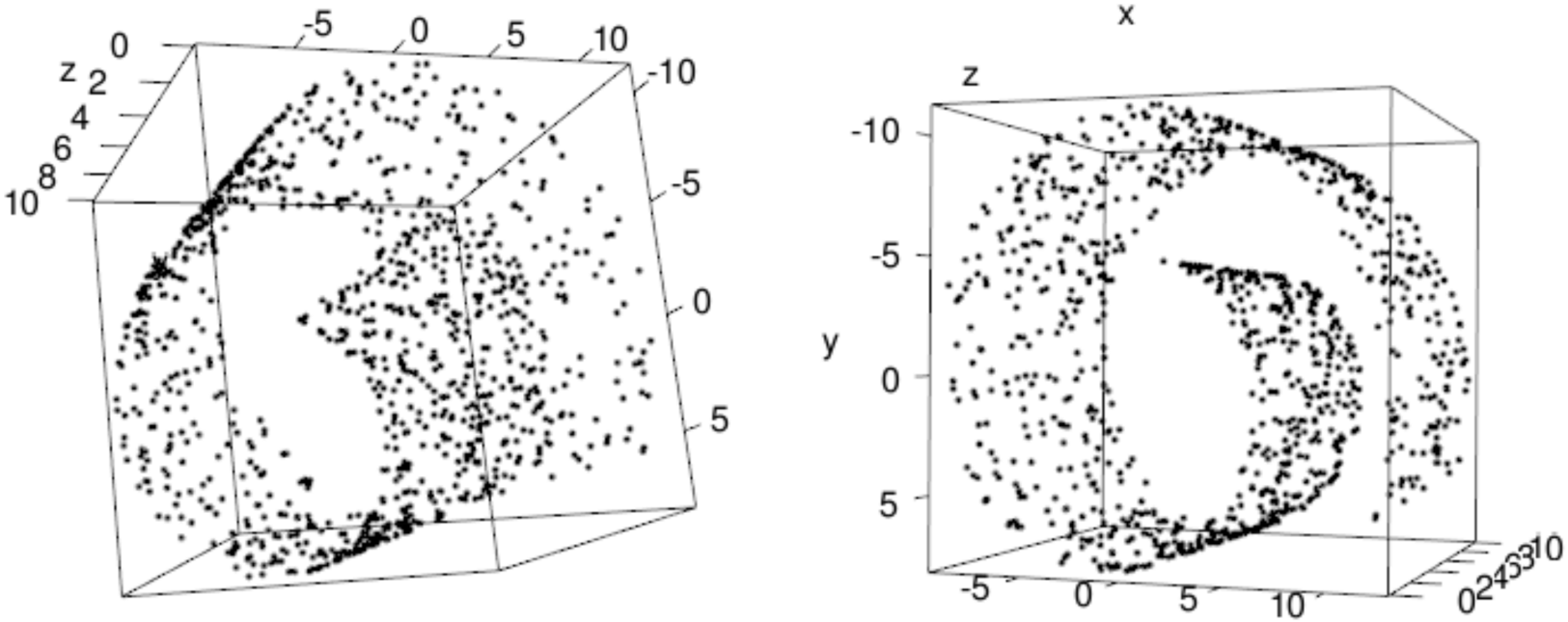
Can you guess the data generating mechanism?





A matter of perspective

And now?



Swissroll mapping: $(x, y) \to (x \cos x, y, x \sin x)$

Motivation

A small number of variable is often sufficient to effectively describe high-dimensional data

This number is called the **intrinsic dimension (ID)** of the data

The ID can vary within the same dataset

We exploit this fact to gain insight in the data structure by developing an approach to cluster regions with the same local ID

Regions with the same ID host points differing in core properties:

- firms with different financial risk in balance sheets data
- identified vs unidentified models in MCMC simulation
- winning vs loosing teams in NBA basketball
- folded vs unfolded state in protein configurations
- active vs non-active regions in brain imaging data

A simple topological feature uncovers a rich data structure

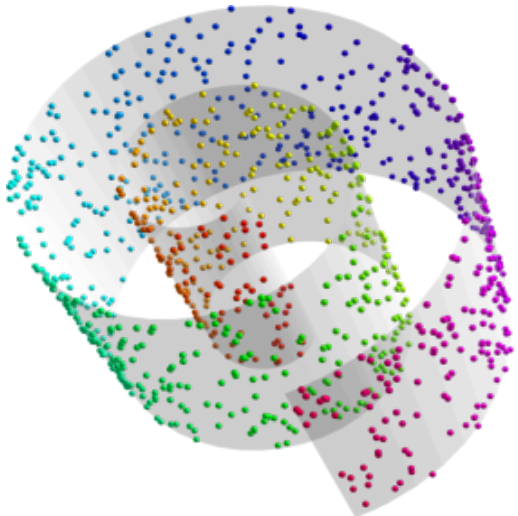
The Intrinsic Dimension (ID) of the Data

Given data points with D coordinates, the **ID** $= d$ is **the minimum number of dimensions required to describe the data, while minimizing the information loss**

Many methods [e.g. 1-3] for estimating the ID are based on the scaling of the number of neighbors of point x_i within distance r

$$N_i(r) \approx r^d \rho(x_i)$$

where $\rho(x_i)$ is the density of the data evaluated in x_i



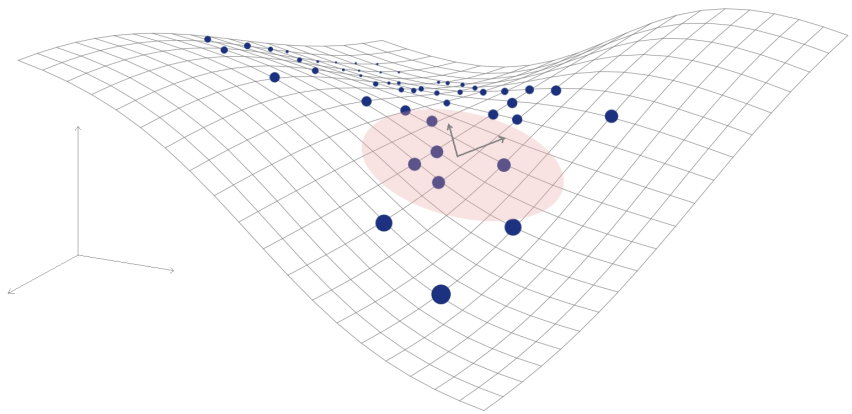
1 Physics Rew. Letters, 1983

2 Proc. Machine Vision, 2003

3 Scientific Report, 2016

Statistical inference based on Local Intrinsic Dimension

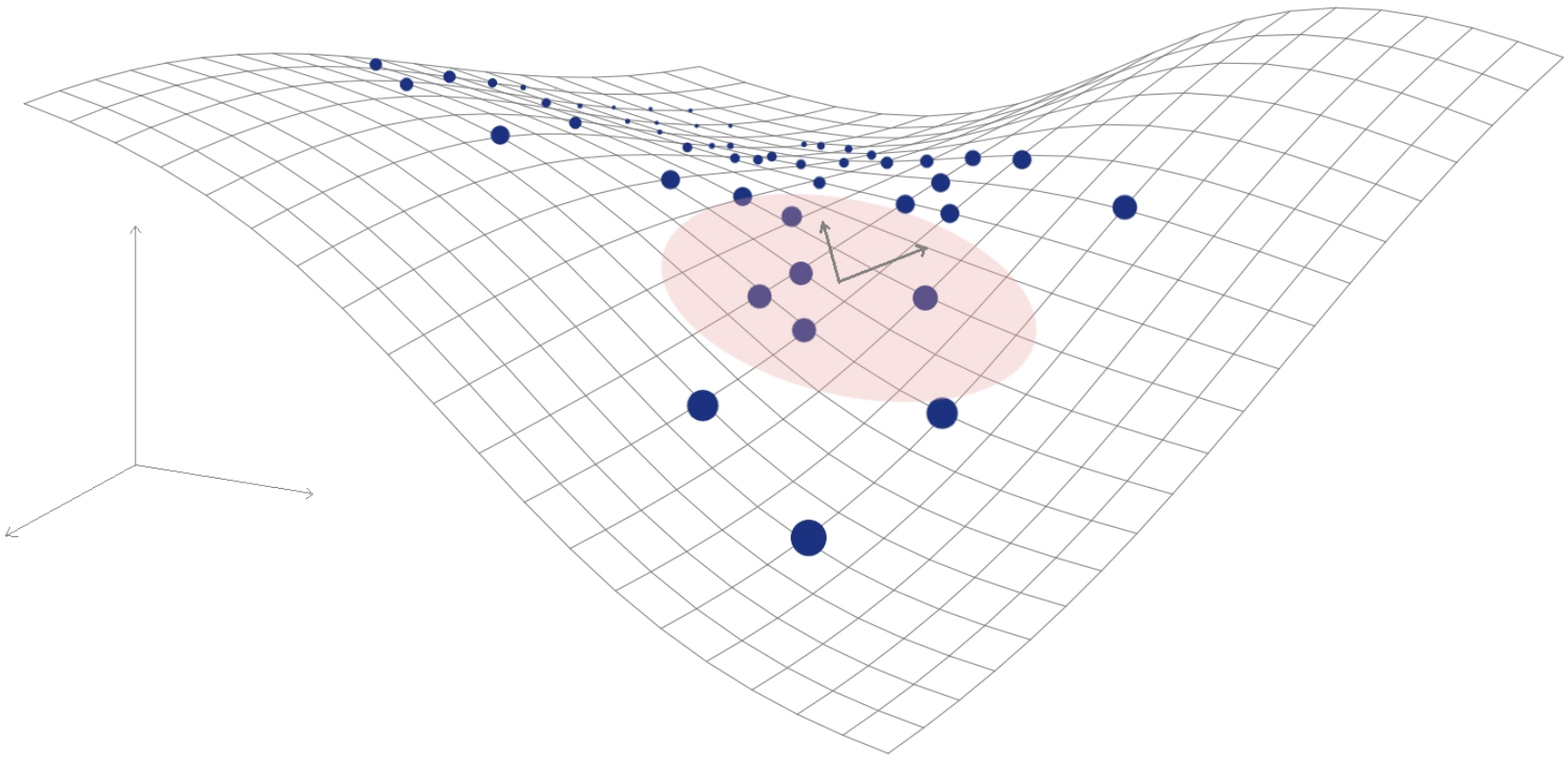
Referring to the ID, we assume that the number of **independent directions** of variation of the data can be lower, equal to $d < D$



Accounting for the ID can improve statistical analysis such as identification of patterns and classification schemes which are computationally hard in high dimension D (many variables)

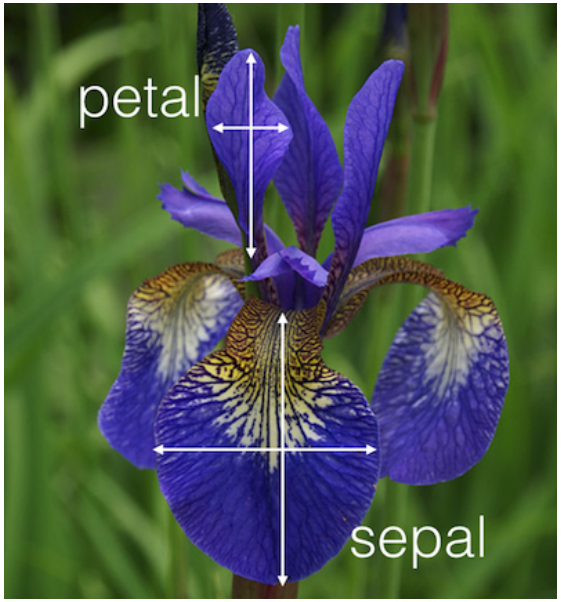
Intrinsic dimension: what is the right d?

If the data actually lie on hypersurface of lower dimension than D the density should be evaluated on this hypersurface

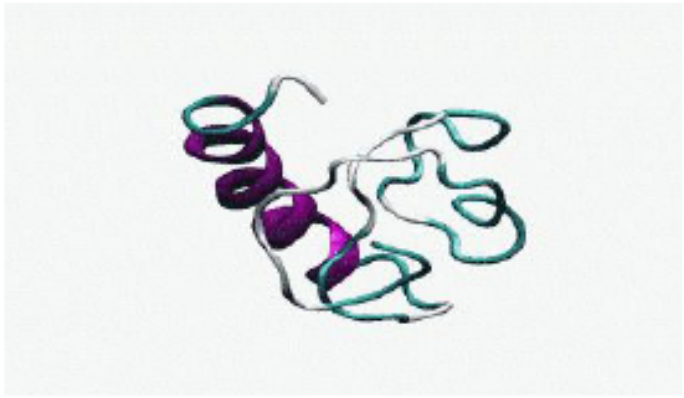


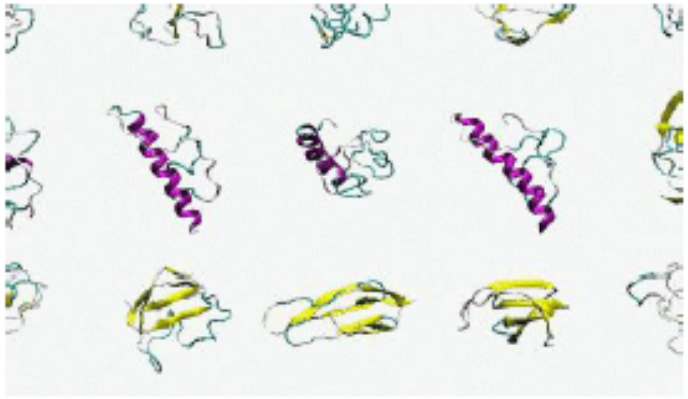
ID Toy Example: Iris Data

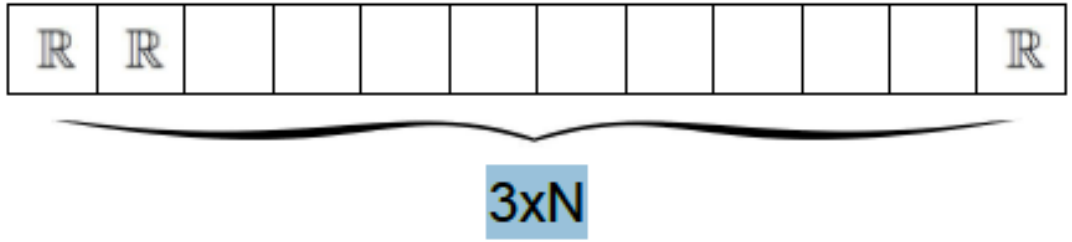
Three types of Iris flowers,
$D = 4$ recorded variables



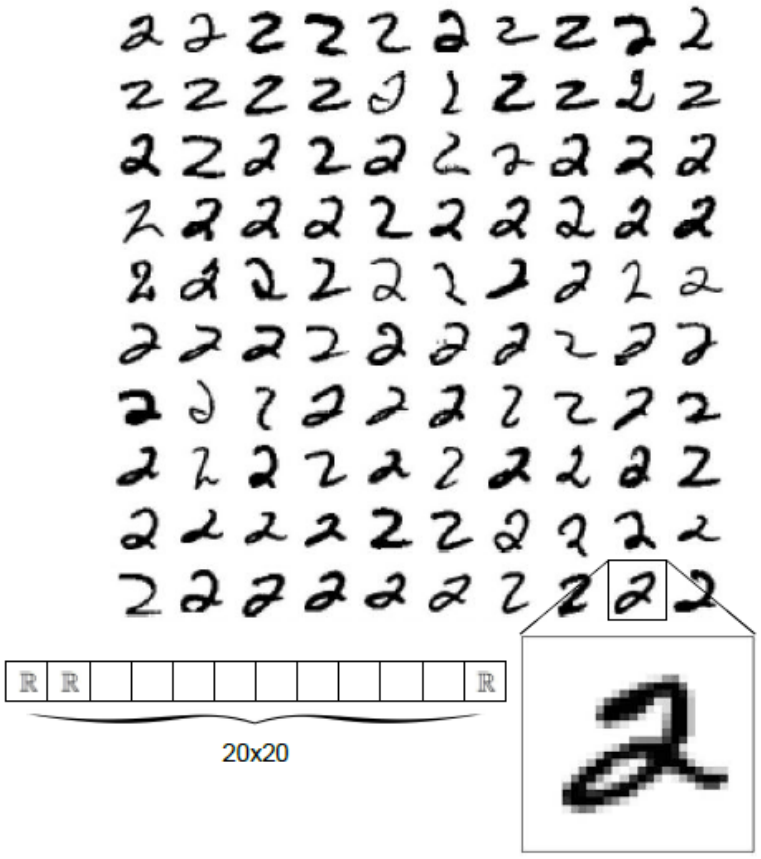
ID Example I: Molecular Dynamics

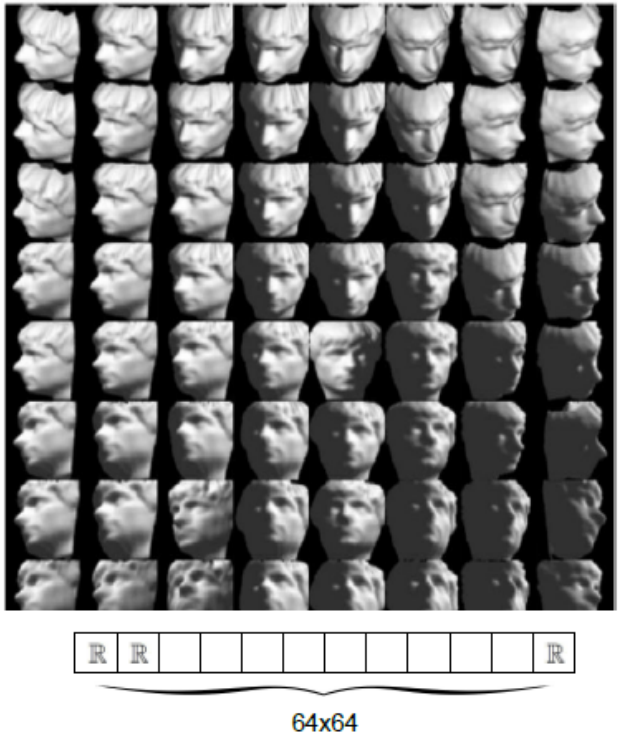




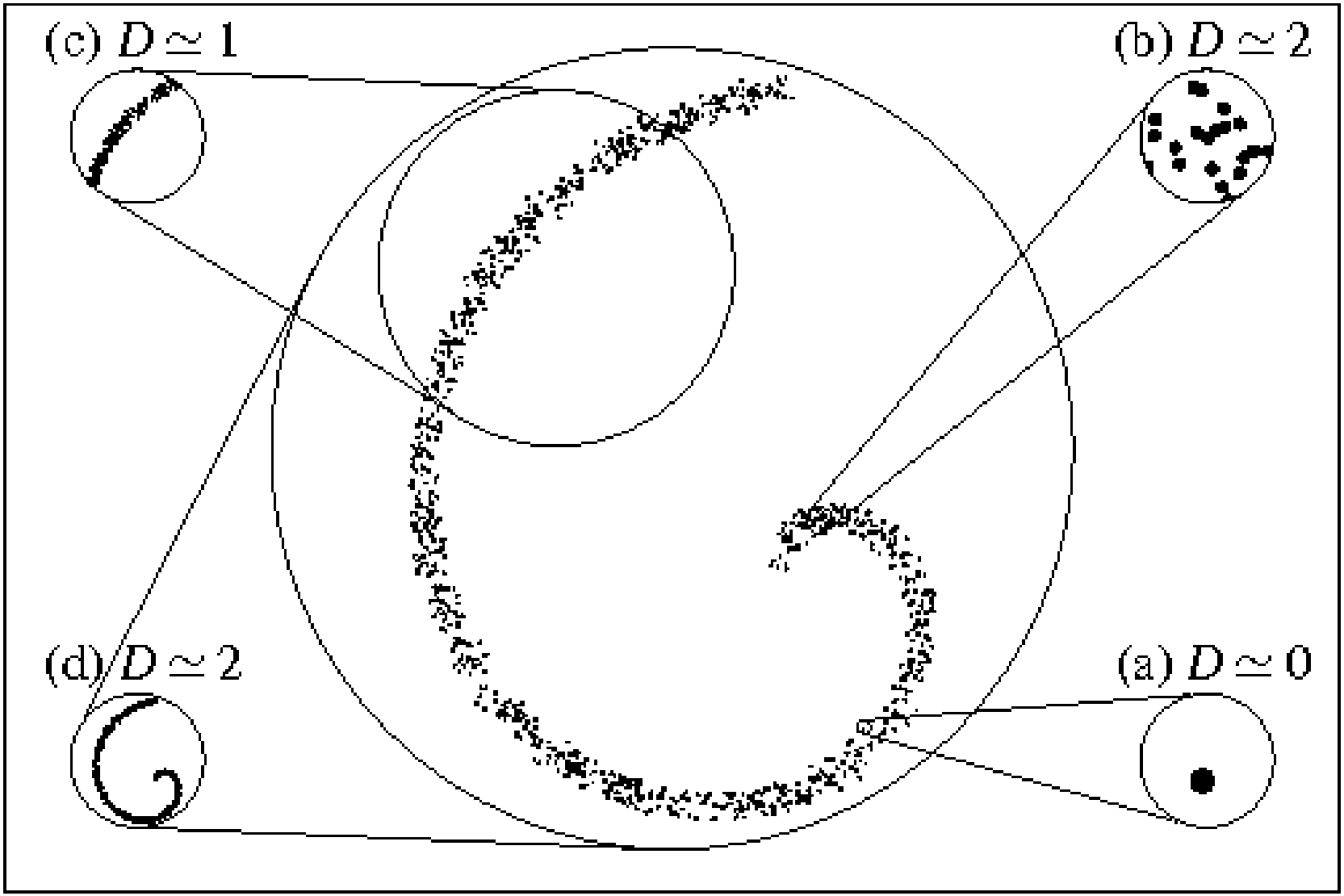


ID Example II: Image Processing



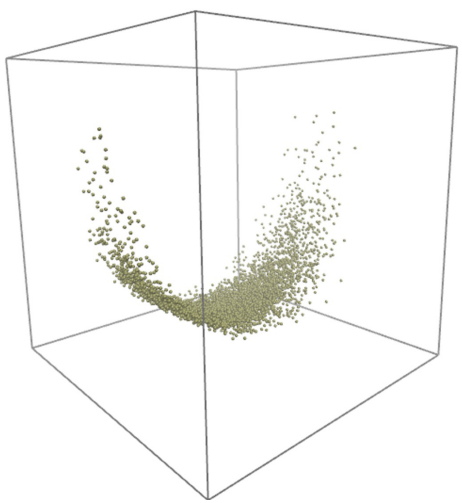


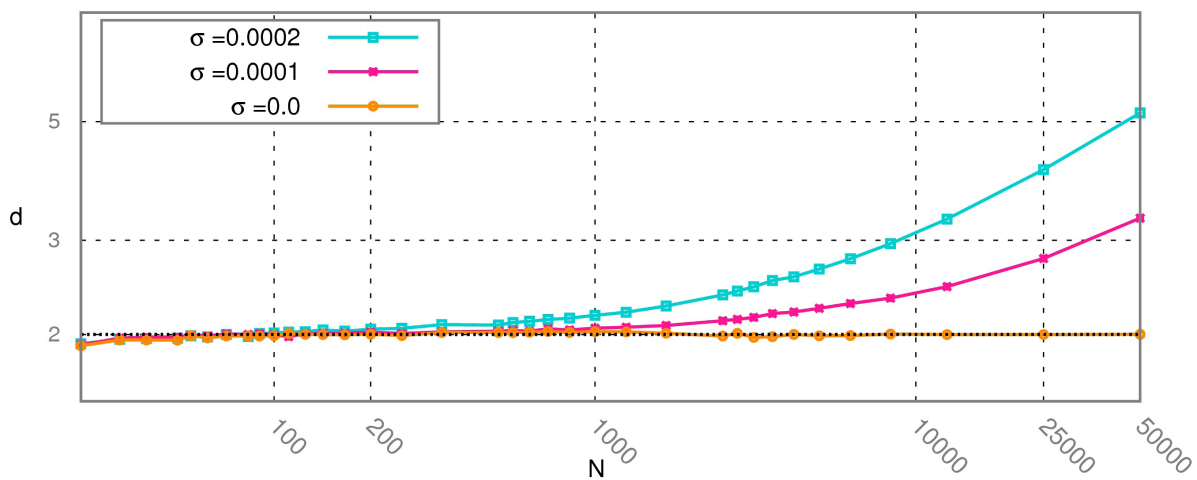
ID Example III: Different Manifolds



The Intrinsic Dimension as a function of the scale

We randomly extract samples from the dataset
The smaller is the sample size, the larger is the typical nearest neighbor distance
We compute the ID as a function of the size of the subsample N
Example: $d = 2$ dimensional gaussian wrapped around a swissroll and embedded in a $D = 30$ dimensional space + 30 dimensional noise
A plateau in the plot of d vs N indicates the ID (d = 2)





How to estimate the ID? Projective approach

Project D-dimensional data into lower dimension d:

$$\Pi^d : \mathbf{x}_i \in \mathbb{R}^D \mapsto \mathbf{y}_i \in \mathbb{R}^d$$

We should try different d and evaluate for each a Loss function: $\mathcal{L}\left(\Pi^d\right)$, where $\mathcal{L}\left(\Pi^d\right)$ measures the data loss occurring in the projection

Examples:

$\mathcal{L}\left(\Pi^d\right) = \sum_i \|\mathbf{x}_i - \mathbf{y}_i\|^2$ Preserves the original distance relations

$\mathcal{L}\left(\Pi^d\right) = \sum_i \mathbf{x}_i\mathbf{x}_i^T - \mathbf{y}_i\mathbf{y}_i^T$ Preserves the original covariance matrix

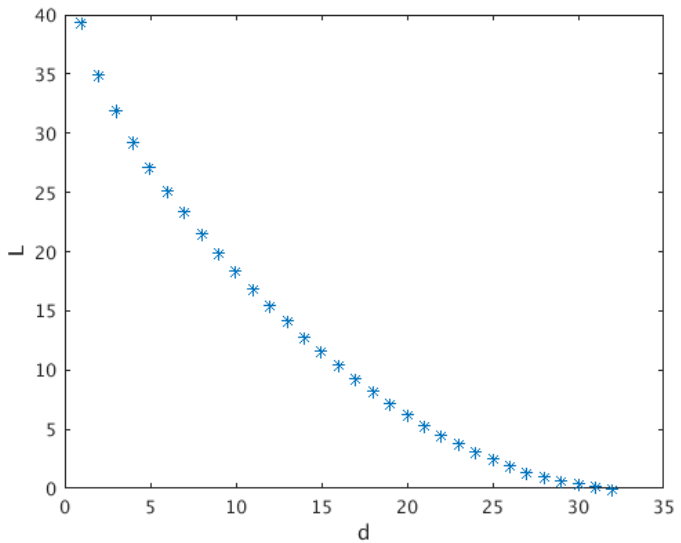
We try to balance the tradeoff between dimension reduction and data loss

How to estimate ID? Projective approach

Problem 1: Computationally burdensome (search for optimal projection for each d)
Problem 2: Robust ID estimates only if $\mathcal{L}\left(\Pi^d\right)$ has large gap as a function of d. If there is no gap, the estimation can be rather arbitrary

Example: Principal Component Analysis (PCA)

Projects data, X, onto linear subspace spanned by first d eigenvalues of covariance matrix $X^T X$
Loss:
$\mathcal{L}\left(\Pi^d\right) = \left\| \sum_i \mathbf{x}_i \mathbf{x}_i^T - \mathbf{y}_i \mathbf{y}_i^T \right\|$
on the villin headpiece MD simulation (example I):



How can one select an appropriate d?

How to estimate ID? A (first) Statistical Approach

Assume that data are sampled from a distribution with density $\rho(x)$
The distances between points in the dataset follow a scaling law that depends on $\rho(x)$ and d
If the dependence on $\rho(x)$ can be removed, then d can be estimated from the scaling relation

Example: Correlation Dimension The number of points at distance $< r$ from point i scales as

$$N_i(r) = \sum_j \mathbb{I}(d_{ij} < r) \approx r^d \rho(x_i)$$

If $\rho(x)$ is constant,

$$N(r) = \sum_{ij} \mathbb{I}(d_{ij} < r) \sim r^d \rho$$

d can be estimated with simple linear fit
However, when $\rho(x)$ is variable the estimation **fails dramatically**

ID: TWO-NN Estimator

TWO-NN idea: **decouple the estimation problem by finding suitable function of the distances that depends only on d**
Two assumptions:

- the data points x_i are independent samples from a density $\rho(x)$
- for all x_i, $\rho(x)$ is approximately constant in the region containing the first 2 neighbors of x_i

Then, consider $\mu_i = r_{i2}/r_{i1}$ where r_{ij} is the distance between i and its j-th nearest neighbor
Under the assumption of local uniformity, the distribution of μ_i depends only on d and follows a Pareto law:

$$\mathcal{L}(\mu_i) = d\mu_i^{-(d+1)}$$

Presented in: *Facco, Errico, Rodriguez, Laio, Scientific Reports (2017)*

ID: TWO-NN Estimator

The ID can be inferred from the μ_i of all points collectively (fit a Pareto distribution)
This is independent on the estimates of ρ (assuming ρ is constant over scale of first 2 neighbors)

There are several ways of fitting:

- Fit the empirical cumulative distribution of μ: $F(\mu) = 1 - \mu^{-d}$
- Equivalently, linear fit on $-\log(1 - F(\mu_i)) = d \cdot \log \mu_i$

If the assumptions are satisfied, then the distribution of μ_i is well fitted

ID: TWO-NN Estimator

The ID can be inferred from the μ_i of all points collectively (fit a Pareto distribution)
This is independent on the estimates of ρ (assuming ρ is constant over scale of first 2 neighbors)

There are several ways of fitting:

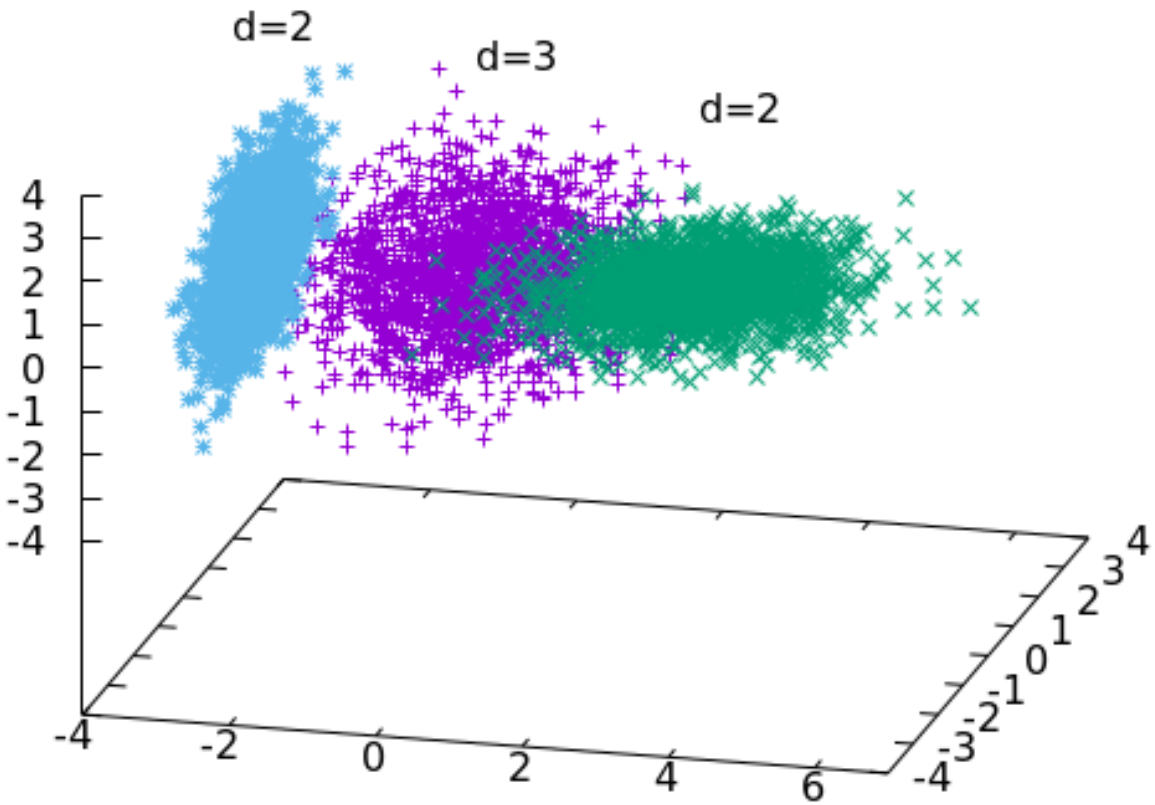
- Fit the empirical cumulative distribution of μ: $F(\mu) = 1 - \mu^{-d}$
- Equivalently, linear fit on $-\log(1 - F(\mu_i)) = d \cdot \log \mu_i$

If the assumptions are satisfied, then the distribution of μ_i is well fitted
If the fit is not good, it means the model fails because:

1. the density is strongly varying even on the scale of the first two neighbors
2. **the intrinsic dimension is not uniform in the dataset**

The problem of multiple IDs

The data may lie on several manifolds $\mathcal{M}_1, \ldots, \mathcal{M}_K$, each with different ID: $d_1 \ldots d_k$. Example with $D = 3$ and $K = 3$:



How to deal with this heterogeneous ID case? HIDALGO!

Heterogeneous ID algorithm - Hidalgo model
allows for the possibility that the ID may not be uniform in the dataset. Assumptions of the model:
H1) $\rho(x)$ **is constant (uniform) on scale of the first two neighbors**
H2) $\rho(x)$ **has support on the union of a finite number** K **of manifolds** $\mathcal{M}_1, \ldots, \mathcal{M}_K$ **with intrinsic dimensions** $d_1 \ldots d_k$

We postulate the density as a mixture

$$\rho(x) = \sum_{k=1}^{K} p_k \rho_k(x)$$

Under the previous assumptions one can show that the distribution of μ_i is a mixture of Pareto distributions

$$f(\mu_i) = \sum_{k=1}^{K} p_k d_k \mu_i^{-d_k - 1}$$

The likelihood of the data is

$$\mathcal{L}(\boldsymbol{\mu}|\mathbf{d},\mathbf{p}) = \prod_{i=1}^{N}\sum_{k=1}^{K} p_k d_k \mu_i^{-d_k-1}$$

where $\boldsymbol{\mu} = (\mu_1 \dots \mu_N)$
Then we can again estimate

$$\mathbf{d} = (d_1 \dots d_K), \quad \mathbf{p} = (p_1 \dots p_K)$$

The likelihood of the data is

$$\mathcal{L}(\boldsymbol{\mu}|\mathbf{d},\mathbf{p}) = \prod_{i=1}^{N}\sum_{k=1}^{K} p_k d_k \mu_i^{-d_k-1}$$

where $\boldsymbol{\mu} = (\mu_1 \dots \mu_N)$
Then we can again estimate

$$\mathbf{d} = (d_1 \dots d_K), \quad \mathbf{p} = (p_1 \dots p_K)$$

To estimate parameters, fix inferential approach

(A) Frequentist:

$$\mathbf{d}^e, \mathbf{p}^e = \text{argmax}(\mathcal{L}(\boldsymbol{\mu}|\mathbf{d},\mathbf{p}))$$

The likelihood of the data is

$$\mathcal{L}(\boldsymbol{\mu}|\mathbf{d},\mathbf{p}) = \prod_{i=1}^{N}\sum_{k=1}^{K} p_k d_k \mu_i^{-d_k-1}$$

where $\boldsymbol{\mu} = (\mu_1 \dots \mu_N)$
Then we can again estimate

$$\mathbf{d} = (d_1 \dots d_K)\,, \quad \mathbf{p} = (p_1 \dots p_K)$$

To estimate parameters, fix inferential approach

(A) Frequentist:

$$\mathbf{d}^e, \mathbf{p}^e = \text{argmax}(\mathcal{L}(\boldsymbol{\mu}|\mathbf{d},\mathbf{p}))$$

(B) Bayesian:
Fix $P_{prior}(\mathbf{d},\mathbf{p})$ and compute the posterior means
$\mathbf{d}^e, \mathbf{p}^e = \langle \mathbf{d},\mathbf{p}\rangle_{post} \quad P_{post}(\mathbf{d},\mathbf{p}) \propto \mathcal{L}(\mu|\mathbf{d},\mathbf{p}) P_{prior}(\mathbf{d},\mathbf{p})$

Because of the sum over K, hard to work with

$$\mathcal{L}(\boldsymbol{\mu}|\mathbf{d},\mathbf{p}) = \prod_{i=1}^{N}\sum_{k=1}^{K} p_k d_k \mu_i^{-d_k-1}$$

Solution: Introduce **Latent Variables** $\mathbf{Z} = Z_1, \ldots, Z_N$ which record the manifold membership of each point
Likelihood is seen as marginal

$$\mathcal{L}(\boldsymbol{\mu}|\mathbf{d},\mathbf{p},\mathbf{Z}) = \prod_{i=1}^{N} p_{Z_i} d_{Z_i} \mu_i^{-dz_i-1}$$

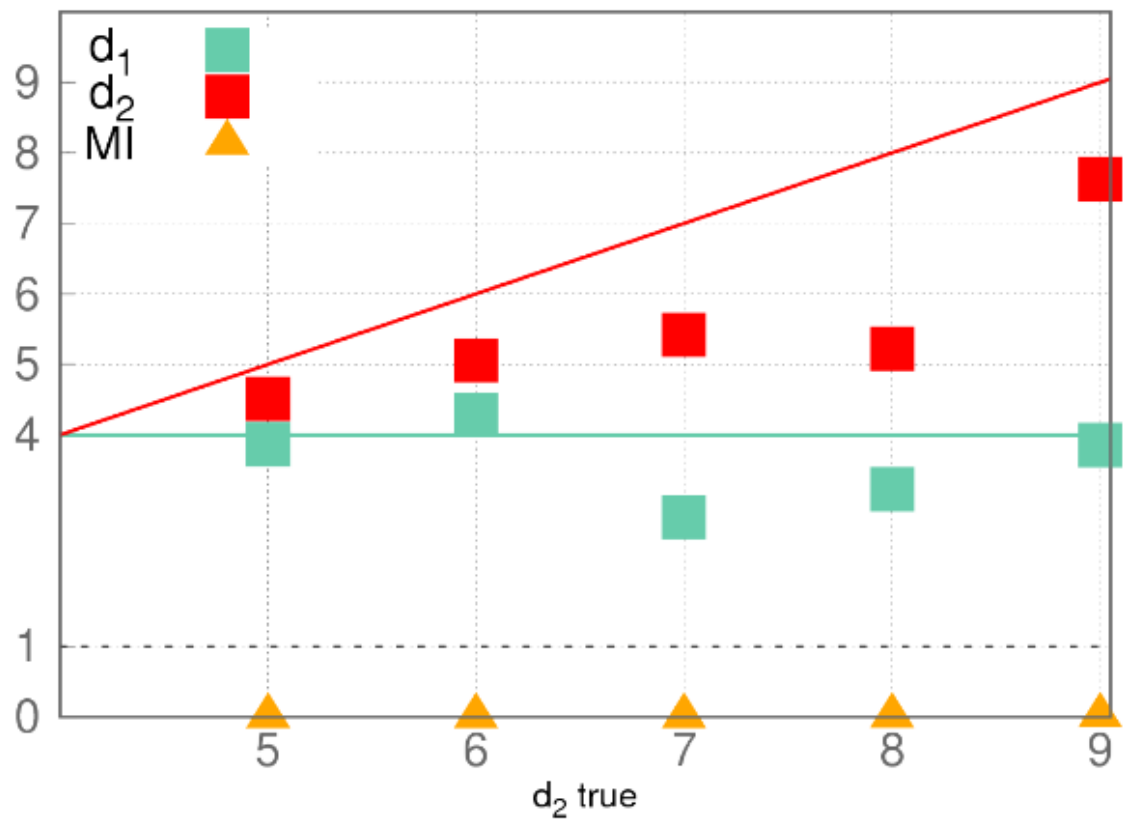
Jointly estimate $(\mathbf{d},\mathbf{p},\mathbf{Z})$

The number of components K is inferred by trying increasing values in $[1, K_{max}]$ and performing model selection with BIC

Simulation Study: $\rho(x) =$ Gaussian

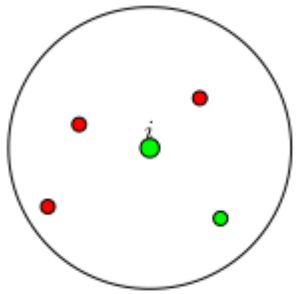
Comparison between two manifolds of dimension $d_1 = 4$ and $d_2 = 5, \ldots, 9$

In every case, the estimation of d_1 and d_2 is inaccurate and the estimation of Z is wrong



This approach does not work! Why?

Pareto distributions with different d are highly overlapping
Difficult to assign a point i based only on its μ_i
Neighboring points have different Z



We must assume that the manifolds are separated, with at most a (small) intersection

One more necessary hypothesis

We then add the following

Hypothesis: the first q neighbors of a point mostly belong to the same manifold

This implies that the neighborhoods of each point must be approximately homogeneous (H1)

We enforce this through **additional term in the likelihood**:

Let the neighborhood of point i be defined by its first q neighbors

$$n_i^{in} = \text{number of neighbors with same Z as } i,$$
$$n_i^{out} = \text{number of neighbors with different Z from } i$$

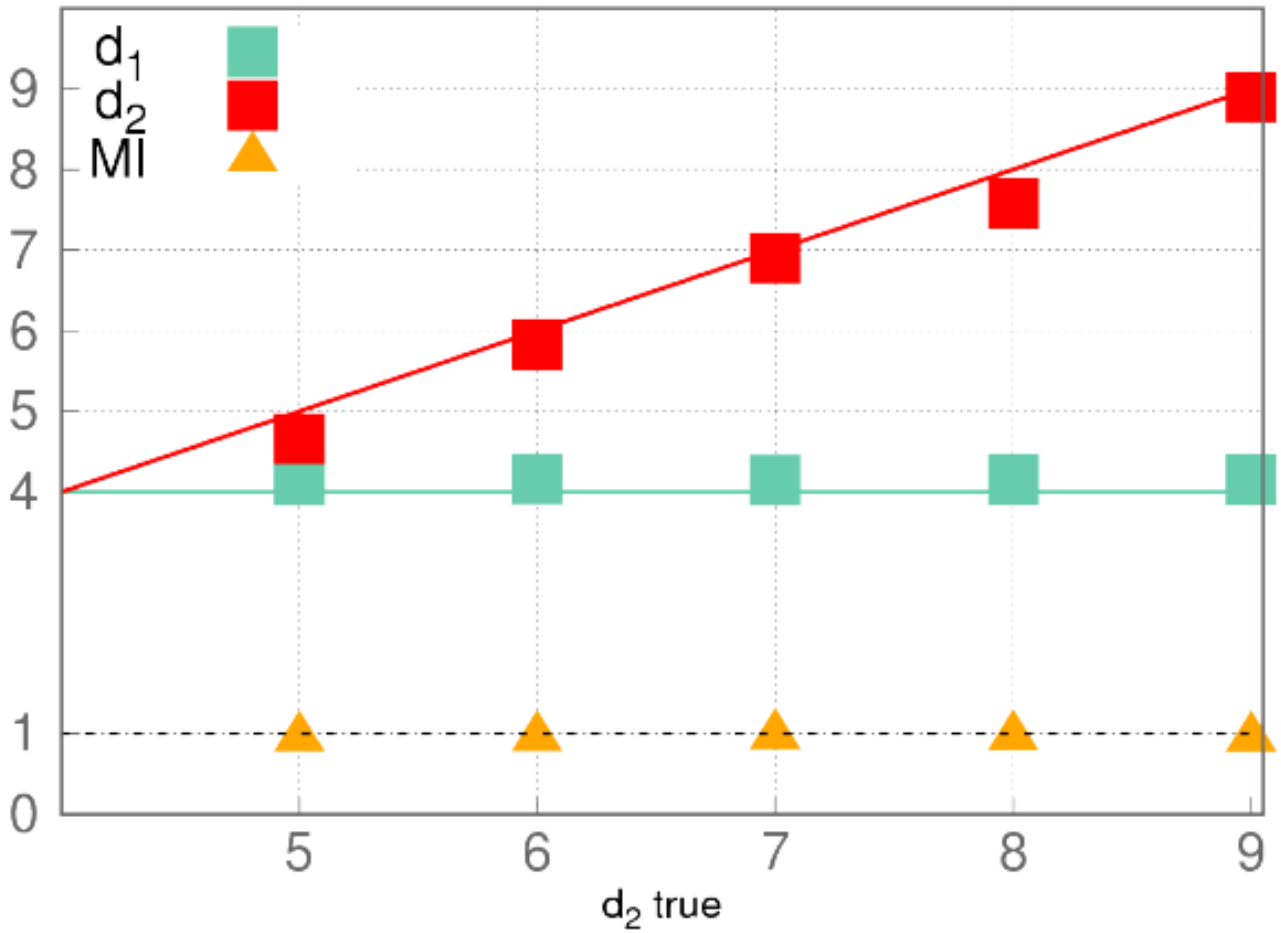
New term in the Likelihood:

$$\mathcal{L}\left(n_i^{in}|\mathbf{Z}\right) = \frac{\zeta^{n_i^{in}}(1-\zeta)^{n_i^{out}}}{\mathcal{Z}}$$

$\zeta > \frac{1}{2}$: Parameter that controls the degree of uniformity

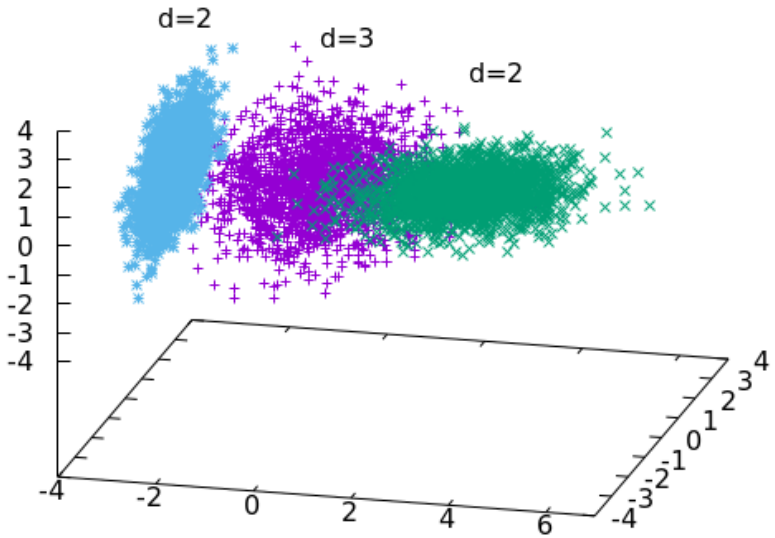
Enforcing uniform neighborhoods

Thanks to $\mathcal{L}\left(n^{in}|\mathbf{Z}\right) = \prod_i \frac{\zeta^{n_i^{in}}(1-\zeta)^{n_i^{out}}}{\mathcal{Z}}$, we get correct estimates of both $\mathbf{d}, \mathbf{p}$ and $\mathbf{Z}$:



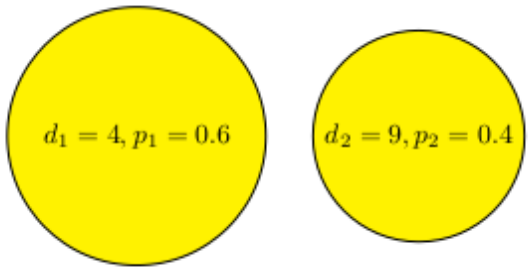
A Global Topological Description

We achieve a global topological description of the data space, dividing the space into regions of uniform intrinsic dimension



A Global Topological Description

We achieve a global topological description of the data space, dividing the space into regions of uniform intrinsic dimension.
We can also represent the size and dimension of the manifolds with a diagram:

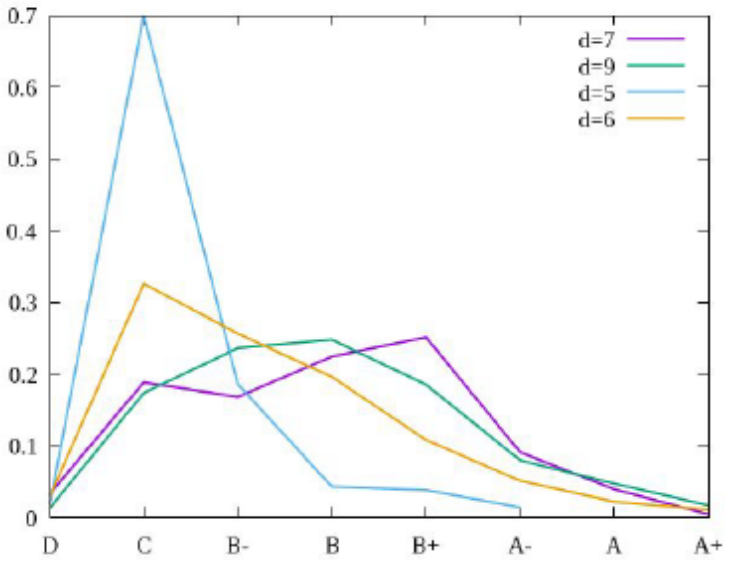


Example: firms from Compustat

- consider ~8000 firms in the Compustat Database
- for each of the firms, D=31 balance sheet variables

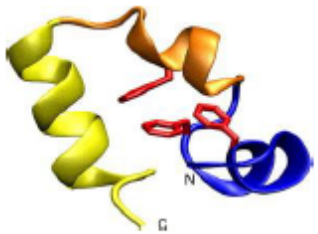
We find four manifolds: d=5, d=6, d=7, d=9

We compute S&P ratings for the different manifolds



Lower dimension tends to have lower ratings!

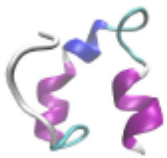
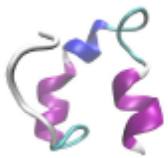
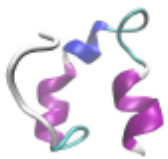
Example: molecular dynamics



- consider a MD of unfolding/refolding villing headpiece
- for each of the N ~ 32000 configurations, D=32 dihedral angles.

We find four manifolds

- d=12 d=13 d=13 d=23
- Q=0.53 Q=0.58 Q=0.64 Q=0.89 Fraction of native contacts

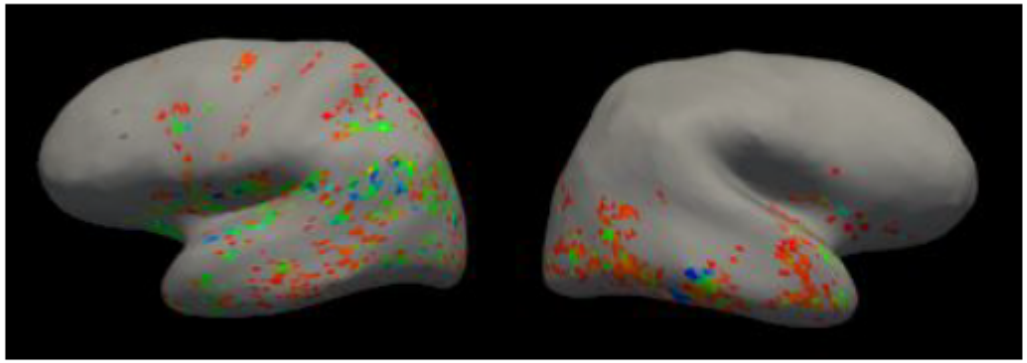





The folded state is recognized from its higher ID!

Example: fMRI time series

- consider ~30000 time series corresponding to BOLD signal of each voxel in an fMRI experiment
- for each of the N ~ 30000 time sieries, D=202 values

We find two manifolds: d=16, d=32



Red: high-ID voxels

Blue: “task-relevant” voxels

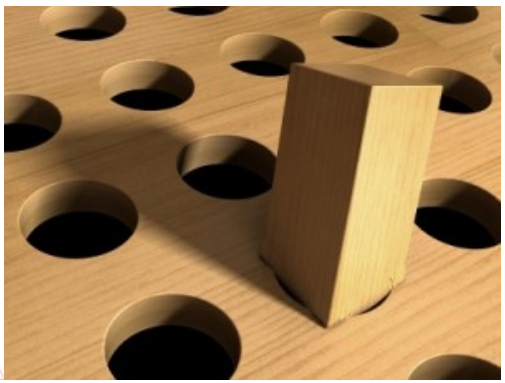
Green: intersection

Task-relevant voxels are in the manifold with higher ID

The low-dimensional manifold nostly includes “noise” voxels

Confirmatory Data Analysis vs Exploratory Data Analysis

CDA: Starts from assumed model for the data, given a priori, and uses statistics to verify whether the data fit the assumed model. Can be rigid: fail to exploit richness of the data



Confirmatory Data Analysis vs Exploratory Data Analysis

EDA: Set of procedures (algorithms) to find structure in the data
Often, no formal evaluation of the results: danger of falling into magical thinking (seeing structures that are not there)

Towards statistical validation of EDA

CDA techniques are rigid, EDA techniques usually lack statistical validation.
Can we have the flexibility of EDA and the reliability of CDA?
A tentative solution is to embed EDA within a statistical framework based on mild assumptions and extremely flexible models.

A possible compromise consists in what we did:
We started with EDA method with no statistical validation of results.
For statistical validation, some assumptions on the data were introduced.
As a result, we developed procedure to reconstruct the data intrinsic dimension.

- To adopt a full Bayesian approach, we need to address the uncertainty on the number of mixture components K
- Instead of making K stochastic, we adopt a Bayesian nonparametric approach, letting $K \to \infty$

Let us denote the *Pareto* $(1, d)$ distribution, with $\mathcal{P}(\cdot|d)$
We now model the ratios of the two smallest distances for every point μ_i as a infinite mixture of Pareto distributions:

$$\sum_{i=1}^{+\infty} p_i \cdot \mathcal{P}(\mu_i|d_i)$$

We can adopt a Dirichlet process prior for the parameters that model the ID

In this way, we formulate a Dirichlet Process Mixture Model:

$$\begin{aligned} \mu_i &\sim \mathcal{P}(\mu_i | d_i) \\ d_i &\sim G \\ G &\sim DP(\alpha, G_0) \end{aligned}$$

where the base measure $G_0 = Gamma(\alpha, \beta)$, to exploit conjugacy

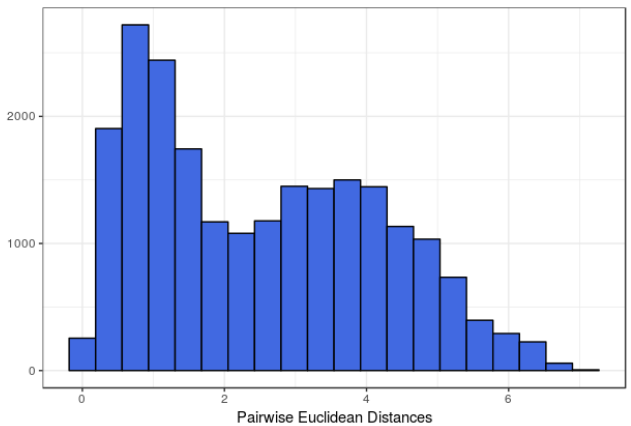
If we introduce a latent variable $\mathbf{Z}$ which denotes, for every observation, the assigned component of the mixture, we can rewrite the model as:

$$
\begin{aligned}
\mu_i|\mathbf{Z}, \mathbf{d}^* &\sim f\left(\mu|d^*_{z_i}\right) \\
Z_i|\mathbf{p} &\overset{ind}{\sim} \sum_{k=0}^{+\infty} p_k \delta_k \qquad \iff \mathbf{P}\left(Z_i = k\right) = p_k \\
n^{in}|\mathbf{Z} &\sim Q \\
\mathbf{p} &\sim SB(\alpha) \\
d^*_k &\sim G_0
\end{aligned}
$$

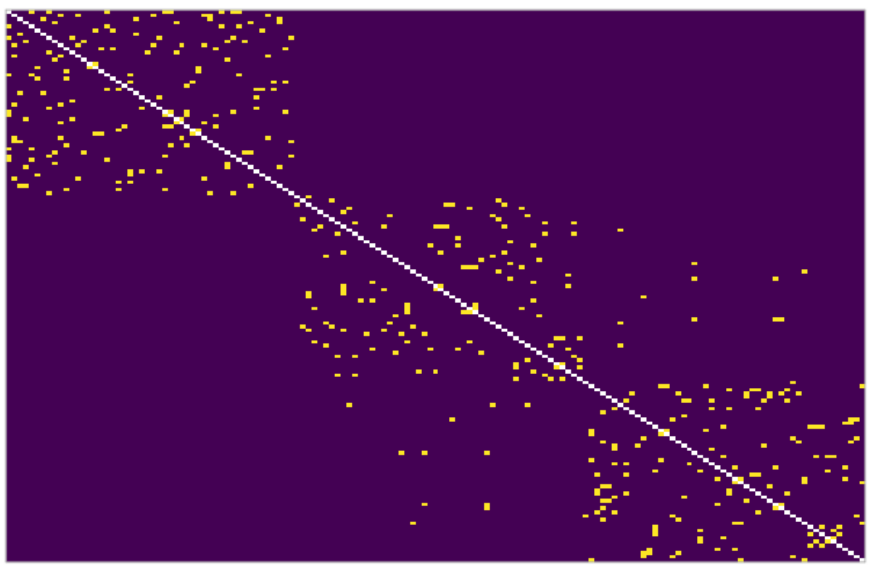
where with SB we denote the usual stick breaking prior, G_0 is a $Gamma(\alpha, \beta)$ and Q is a distribution with a density defined as $\mathcal{L}\left(n^{in}|\mathbf{Z}\right) = \prod_i \frac{\zeta^{n_i^{in}}(1-\zeta)^{n_i^{out}}}{\mathcal{Z}}$

BNP Hildago: Toy Example - Iris

The two plots contain an histogram of all the pairwise distances between the observations. The bottom one shows the histogram for all μ_i's. Our methodology proposes a stronger dimensionality reduction than other distance-base clustering models [1]

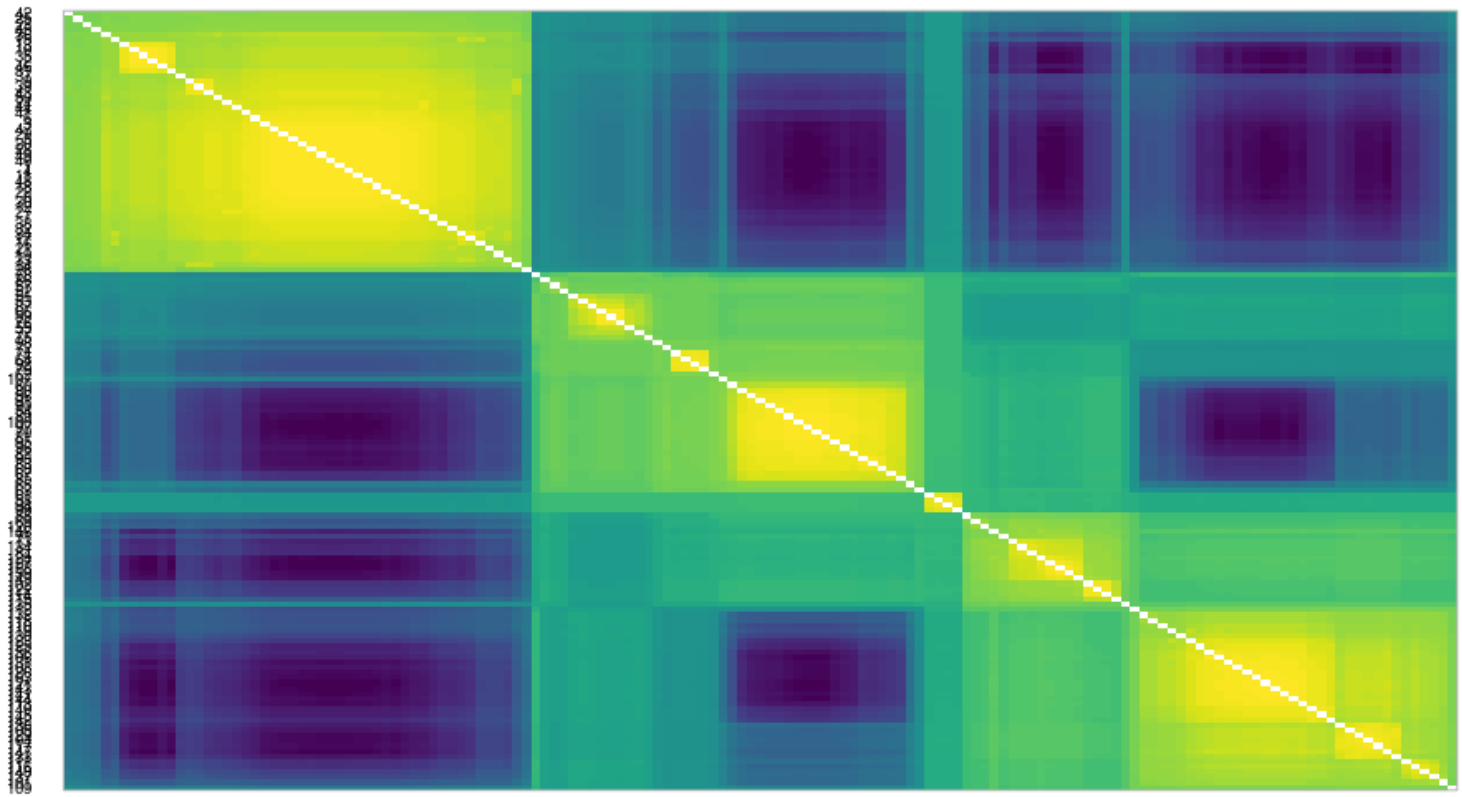


It seems that the data contains some sort of information: look at the matrix $\mathcal{N}^q$, with $q = 3$

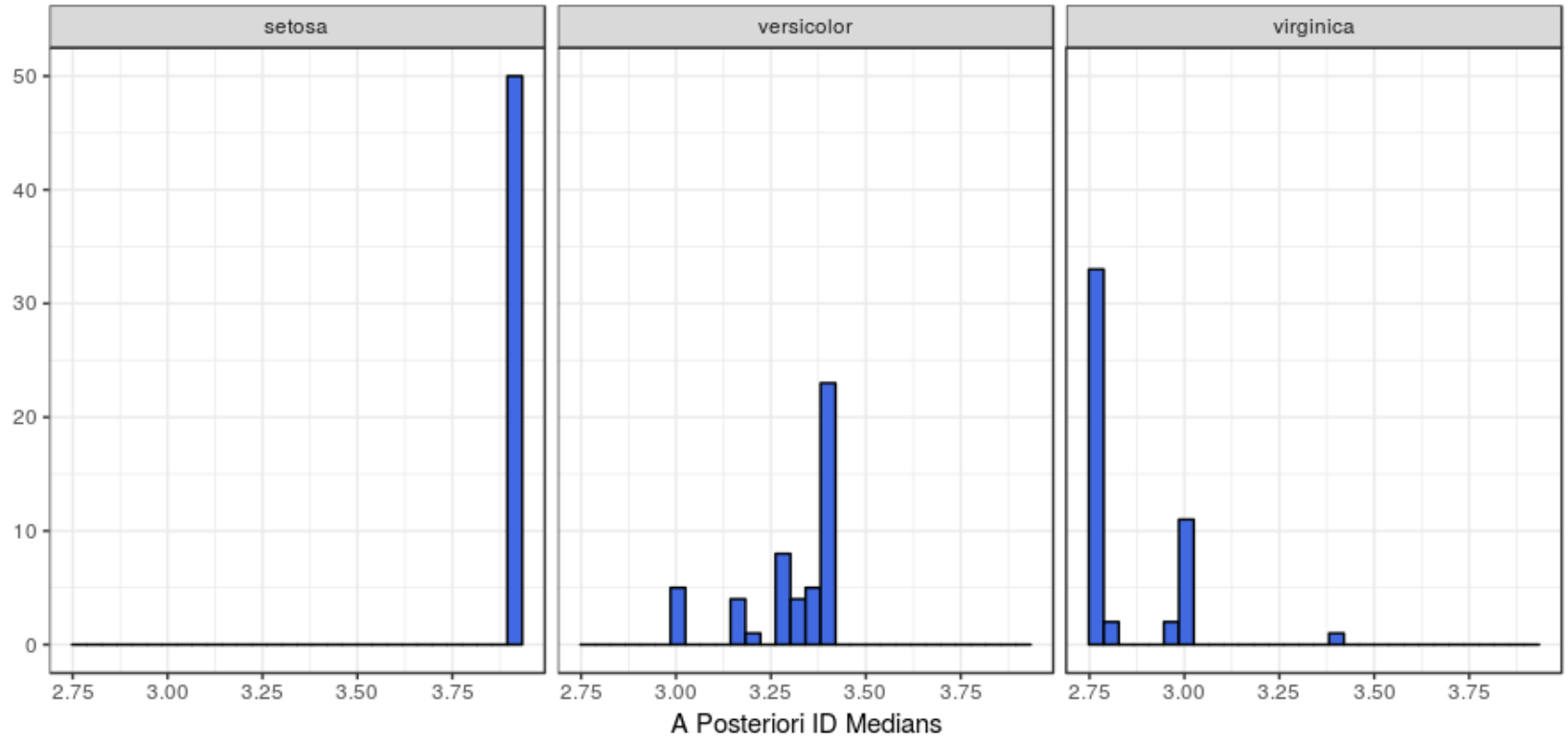


[1] Bayesian Distance Clustering - Duan and Dunson

We record $T = 50k$ iterations after $150k$ burn-in steps
Minimizing the Binder Loss (which measures the disagreements in all possible pairs of observations between the true and estimated clusterings - R function `mcclust::minbinder`),we find three cluster, almost coincident with the Flower Species. (Setosa - Versicolor - Virginica)
Here is the Pairwise Coclustering Probability Matrix.



For each observation μ_i, we obtaind a MCMC of Intrinsic Dimensions d_t, $t = 1, \ldots, T$. The distributions of the posterior medians, grouped per Species, are



We can conclude that the measurements of the `Versicolor` and `Virginica` Species of Iris are embedded in manifolds of dimension < 4

Application II : Identifiability of MCMC output

We can use BNP - Hidalgo to estimate the ID of the path of an MCMC chain to investigate the presence of identifiability issues in the model
Let us consider the following Bayesian linear regression model:

$$M1) \quad Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$$

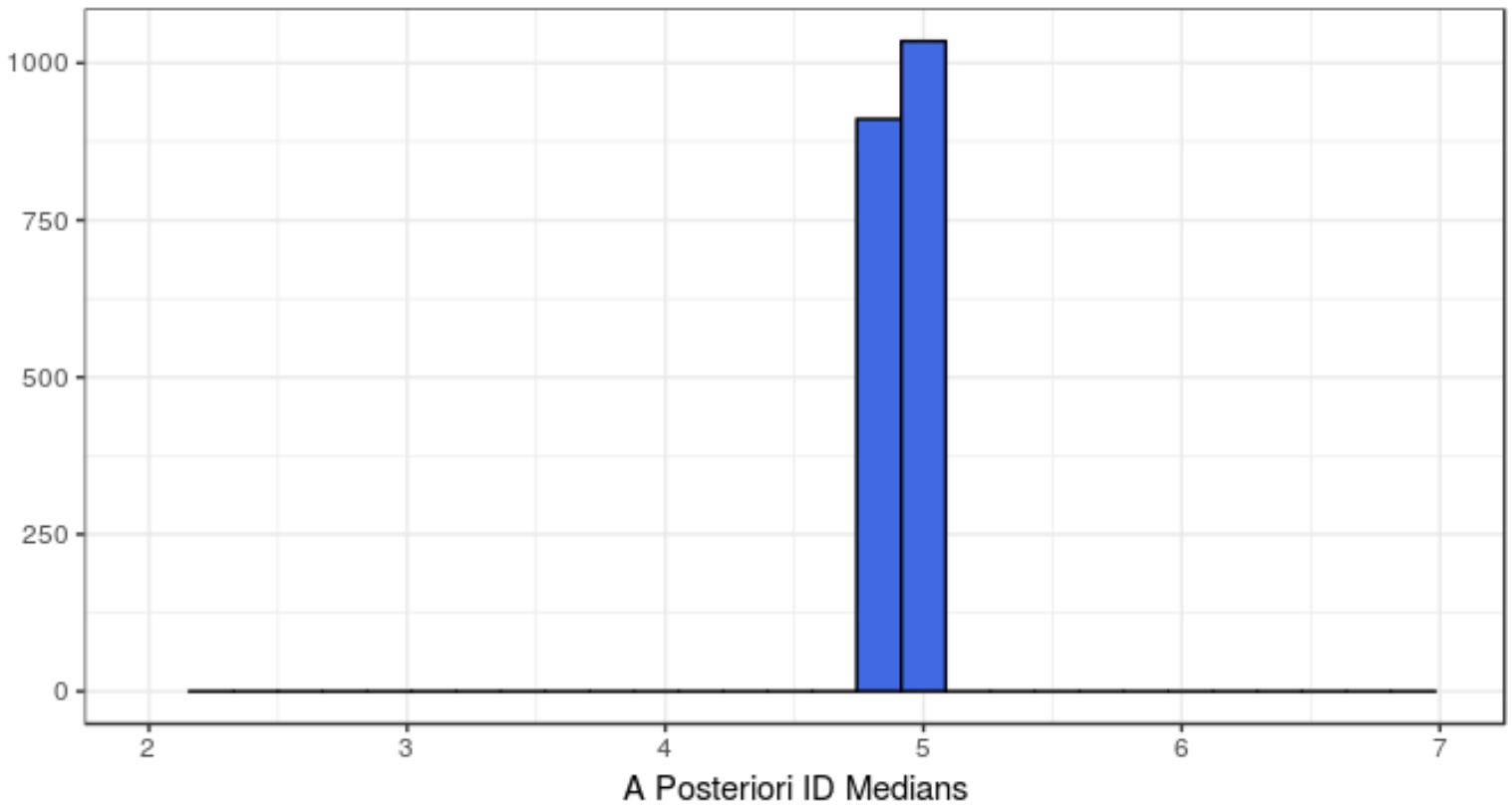
and its ill posed version:

$$M2) \quad Y_i = \beta_0 + \beta_1 + \beta_2 X_{2i} + \beta_3 X_{2i} + \beta_4 X_{4i} + \varepsilon_i$$

where $X_j \sim \mathcal{U}[a_j, b_j]$ on different intervals and $\varepsilon_i \sim \mathcal{N}(0, 1)$

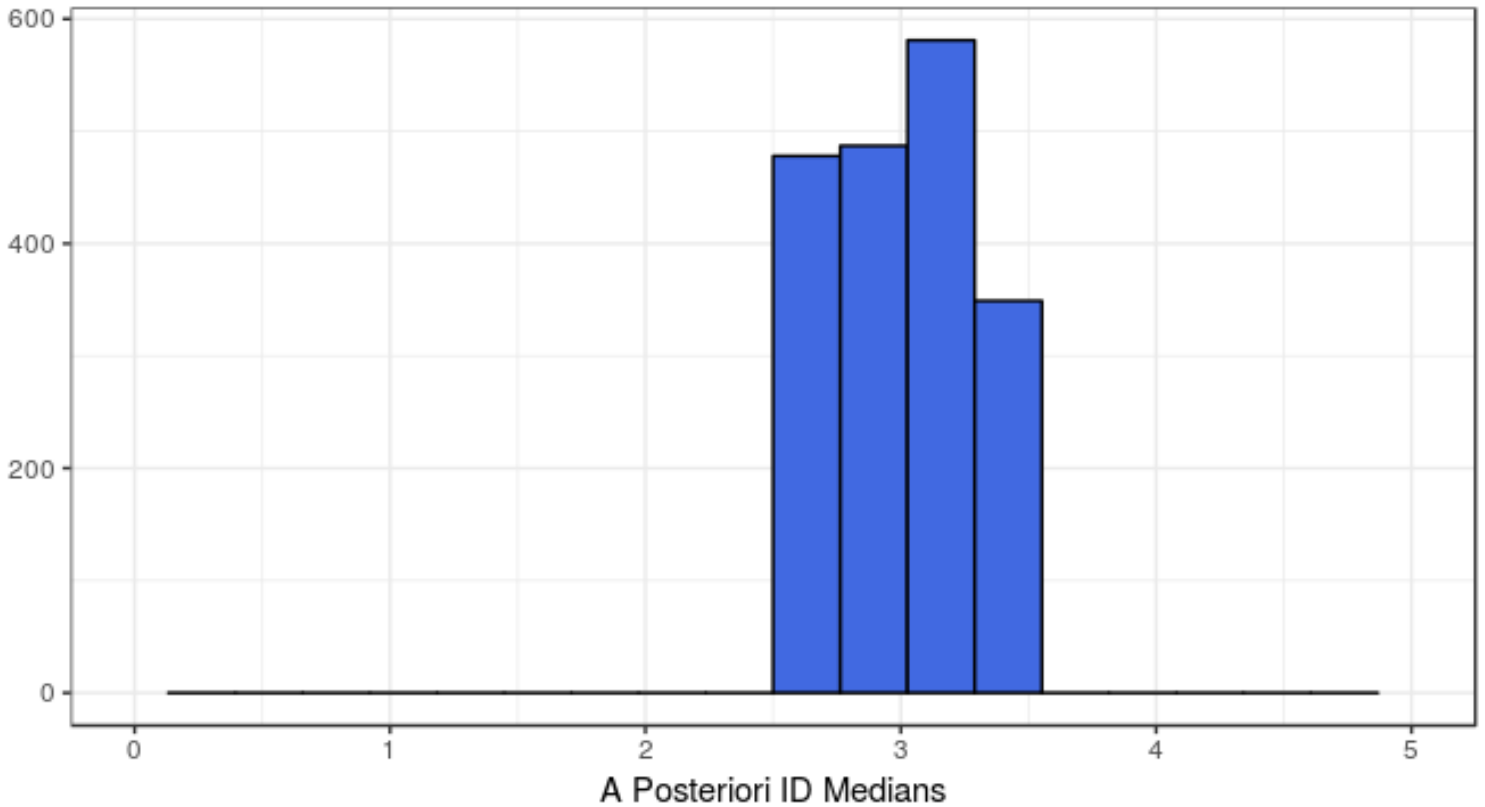
We estimate the model using the Hamiltonian no u-turn sampler (R package `Rstan`)
Every iteration of the Hamiltonian chain is treated as an observation embedded in $\mathbb{R}^5$. After computing the corresponding μ_i's, a MCMC sample of $T = 2k$ iterations is collected after a burn in of $B = 2k$ steps

For each observation μ_i, we obtain a MCMC of Intrinsic Dimensions d_t, $t = 1, \ldots, T$
The distribution of the posterior medians is



Not Identifiable Model

For each observation μ_i, we obtain a MCMC of Intrinsic Dimensions d_t, $t = 1, \ldots, T$
The distribution of the posterior medians is



Our approach detects that two dimensions out of five are actually redundant

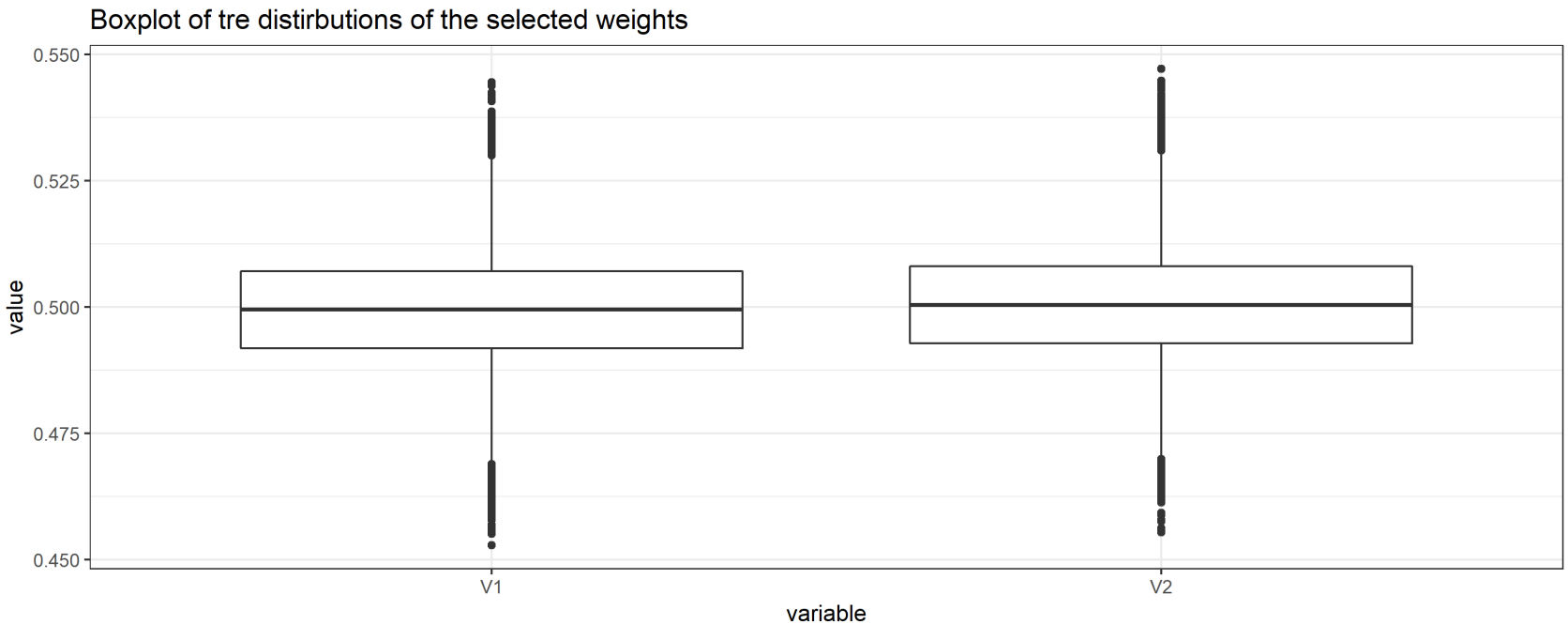
Application on Simulated Data

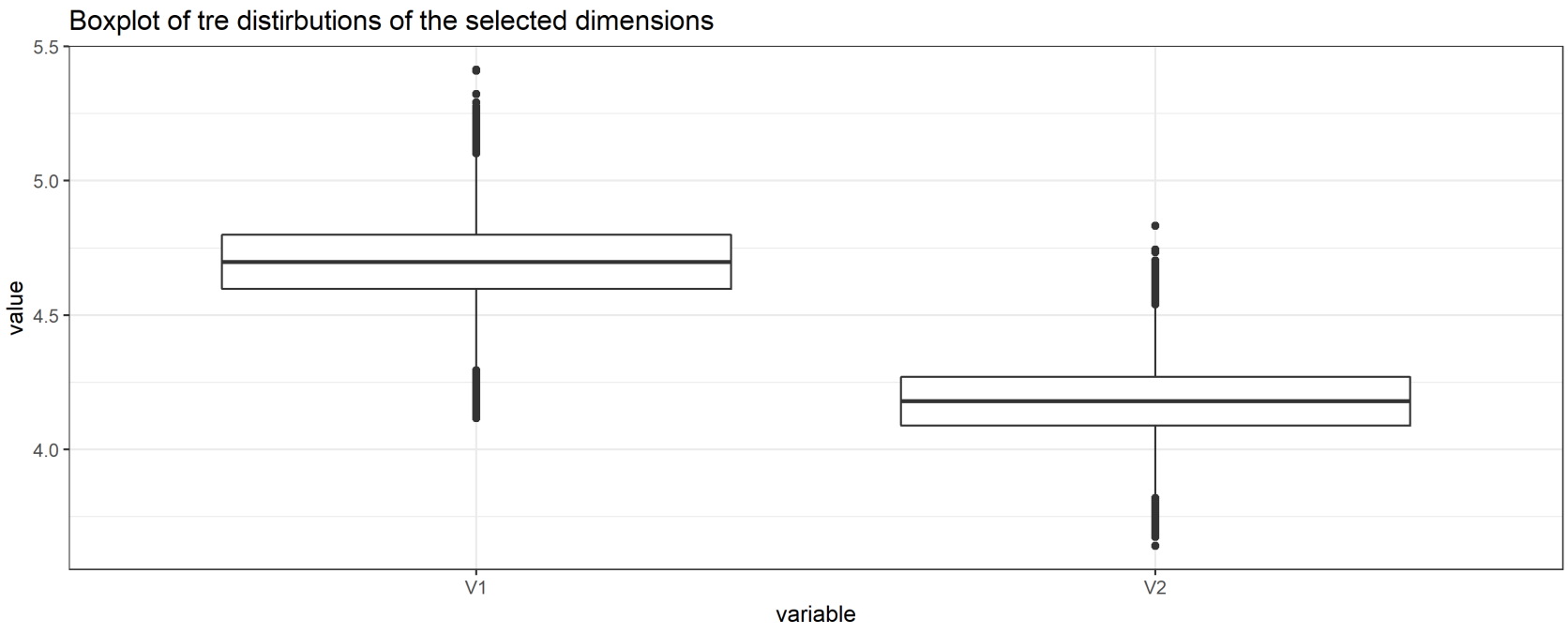
We also apply BNP-HIDALGO to the same simulated data on which the finite-dimensional version of HIDALGO was first evaluated

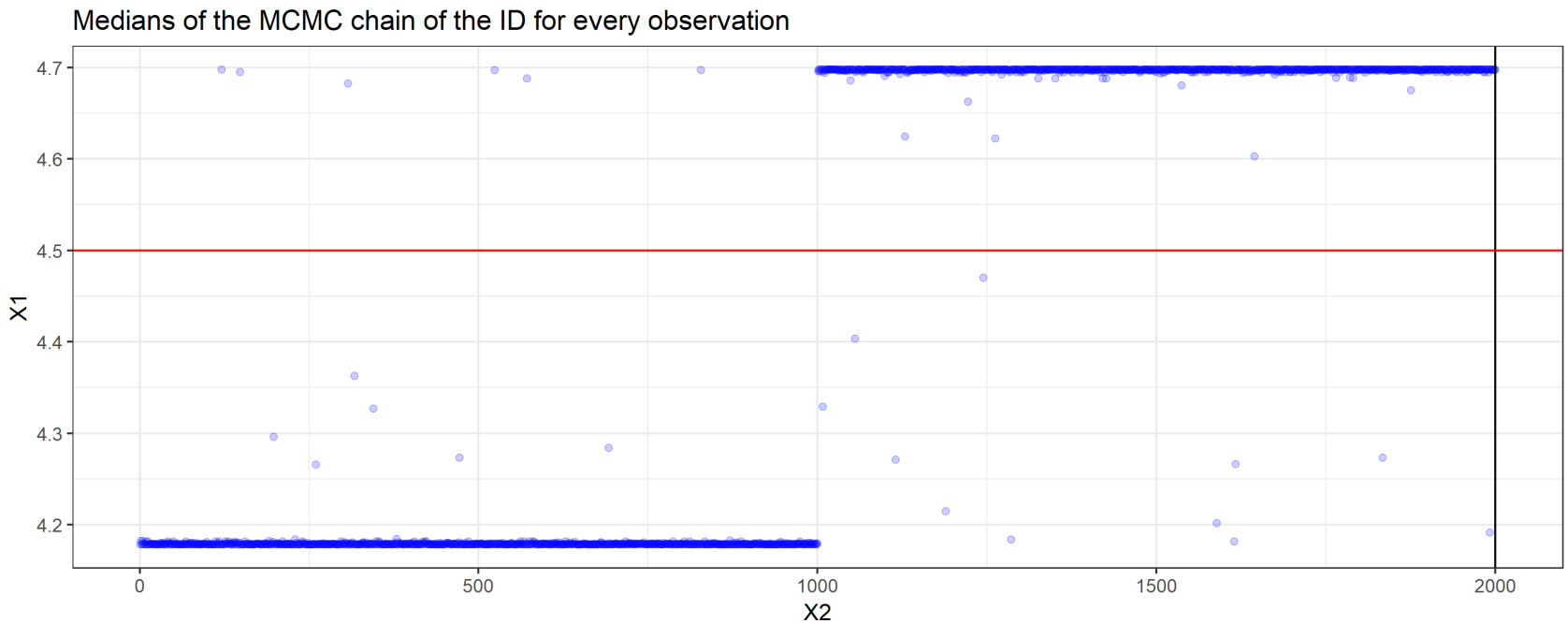
We adopted the same setting to favor the reproducibility of the results and the comparison between the two methods

Benchmark dataset:
1000 observation from a 4-dimensional Gaussian,
1000 observations from a 5-dimensional Gaussian,
both with unitary variance
The centroids are chosen to be at a distance from each other of 0.5, challenging the model with overlapping data

We set the number of neighbors $q = 3$
After a Burn-in period of 20k iterations, 10k samples are retained for posterior inference:




Boxplot of tre distirbutions of the selected dimensions
5.5
5.0
4.5
4.0
value
V1
V2
variable


Medians of the MCMC chain of the ID for every observation
4.7
4.6
4.5
4.4
4.3
4.2
X1
0
500
1000
1500
2000
X2

We decide to assign one observation to the dimension $\hat{d}_i$ equal to 4 or 5 following this criterion: for every observation, we collect the MCMC sample for d, namely d_i^t, with $t = 1, \ldots, 10000$.
We then compute the median over the iterations $\tilde{d}_i$. Then

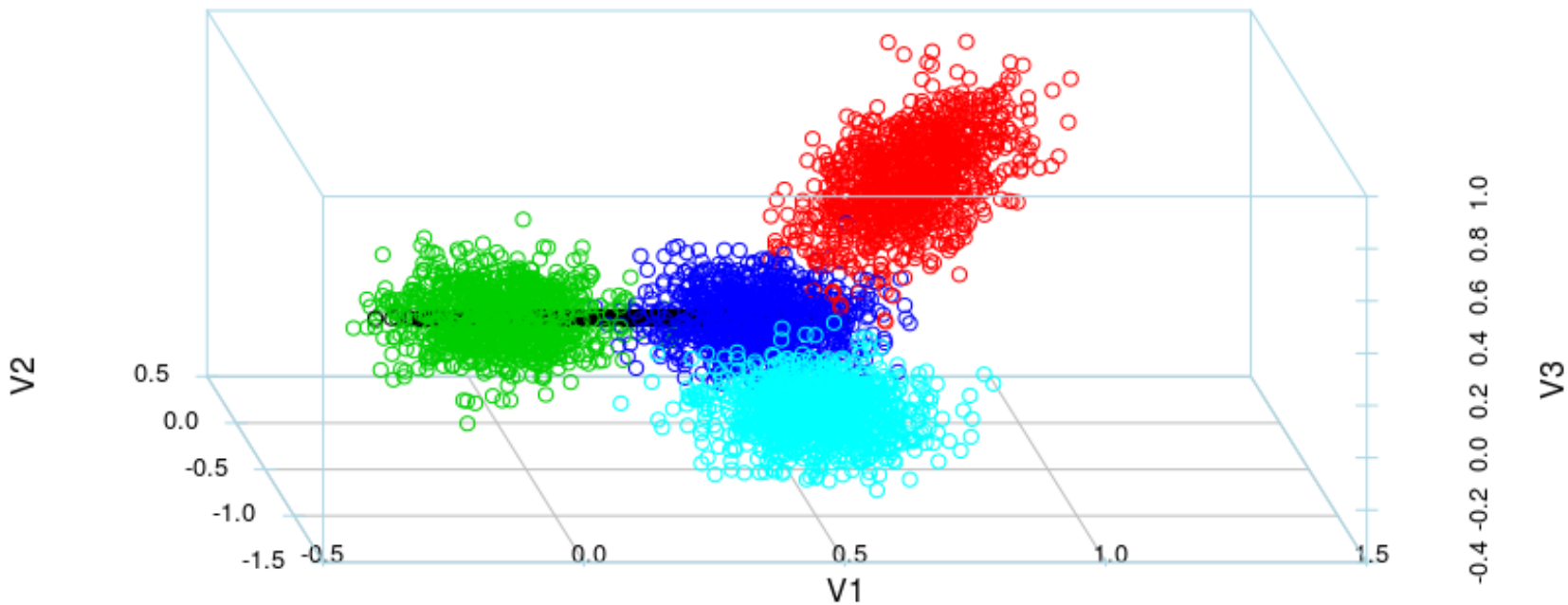
$$\hat{d}_i = \begin{cases} 4 & \text{if } \tilde{d}_i < 4.5 \\ 5 & \text{if } \tilde{d}_i \geq 4.5 \end{cases}$$

We end up with the following **confusion matrix**:

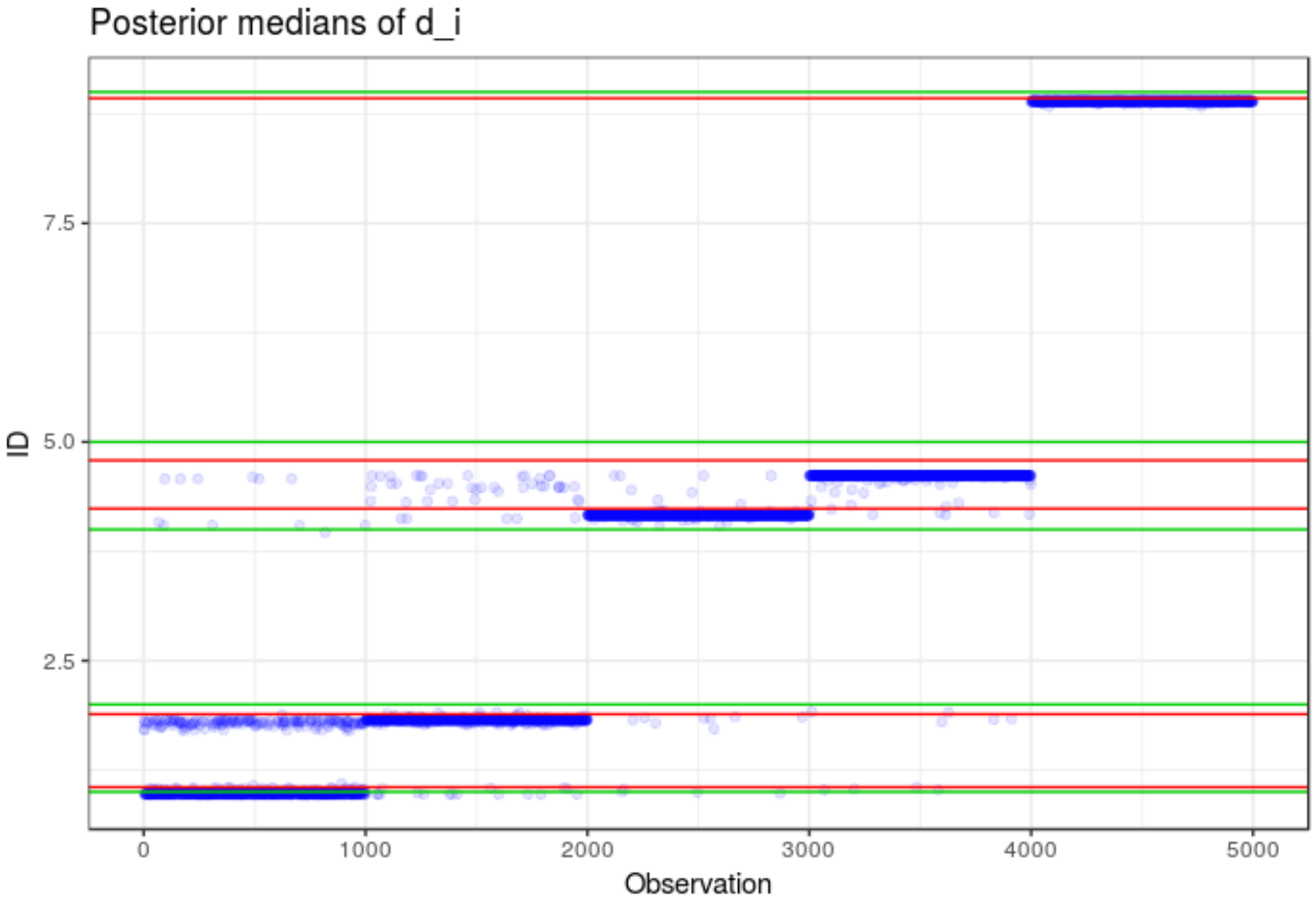
$\hat{d}_i$ vs d_i	4	5
4	**994**	**6**
5	**11**	**989**

A more challenging setting: 1000 observations generated from 5 Gaussian distributions of dimensions 1, 2, 4, 5 and 9, partially overlapping

Here they are projected on the first three dimensions:



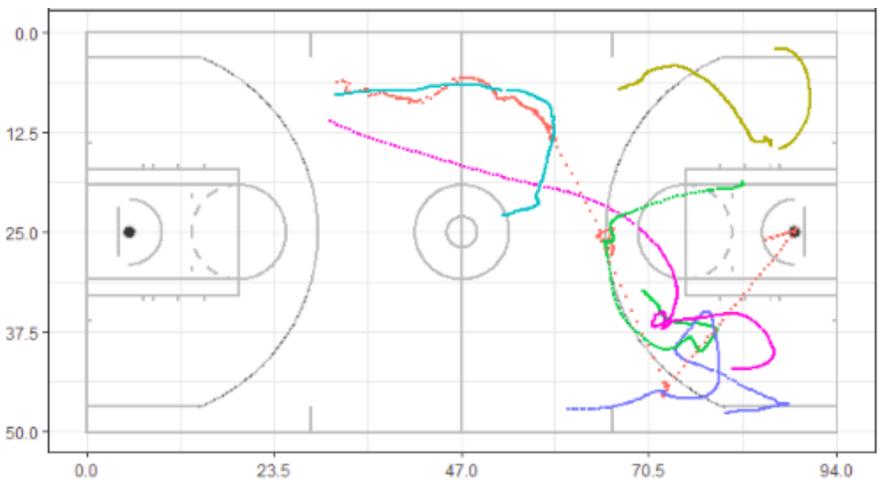
The model is able to estimate the IDs that are in the dataset
The red lines denote the MLE estimates in every subgroup
The green lines denote the actual ID

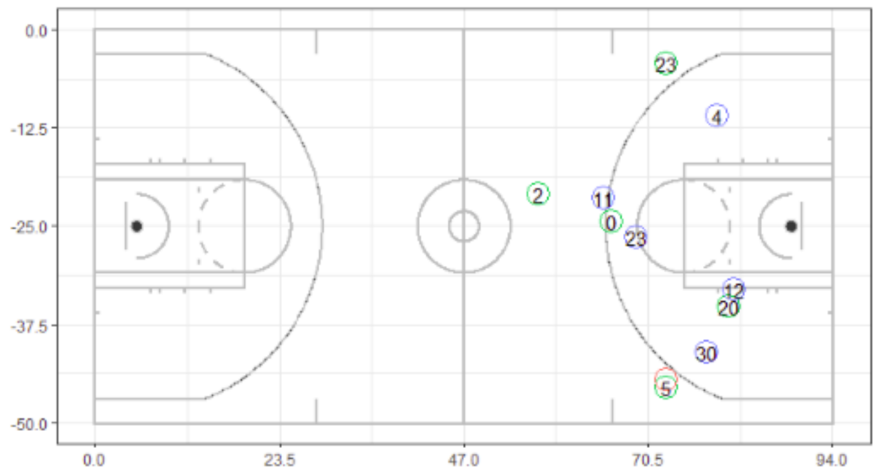


BNP Hildago: Application to Basketball data

Joint work with E. Santos-Fernandez and K. Mergensen

SportVU NBA player tracking technology: captures (at 25 frames per second) the coordinates of each player (x, y) and the ball (x, y, z)
In this analysis, we use the locations of the players and the ball when each **shot** was taken with a potential outcome (scored or missed)

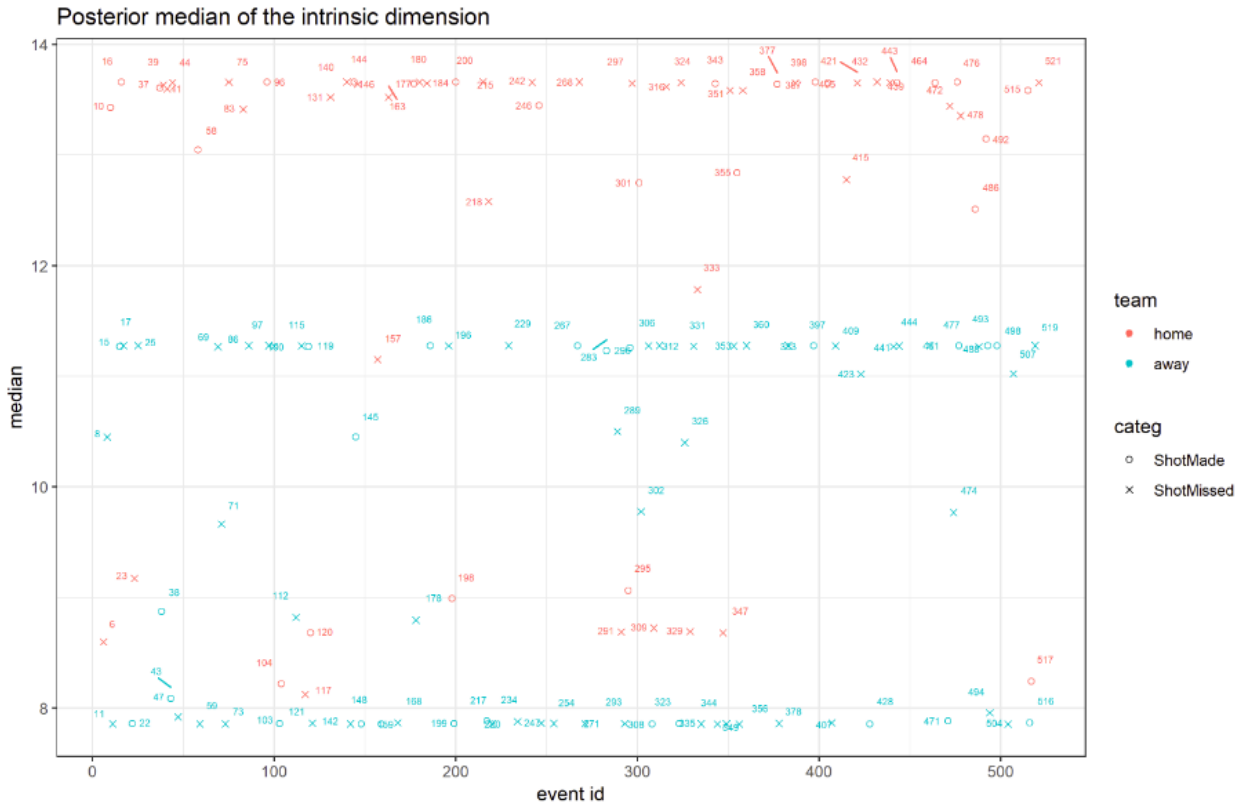




BNP Hildago: Application to Basketball data

Joint work with E. Santos-Fernandez and K.Mergensen

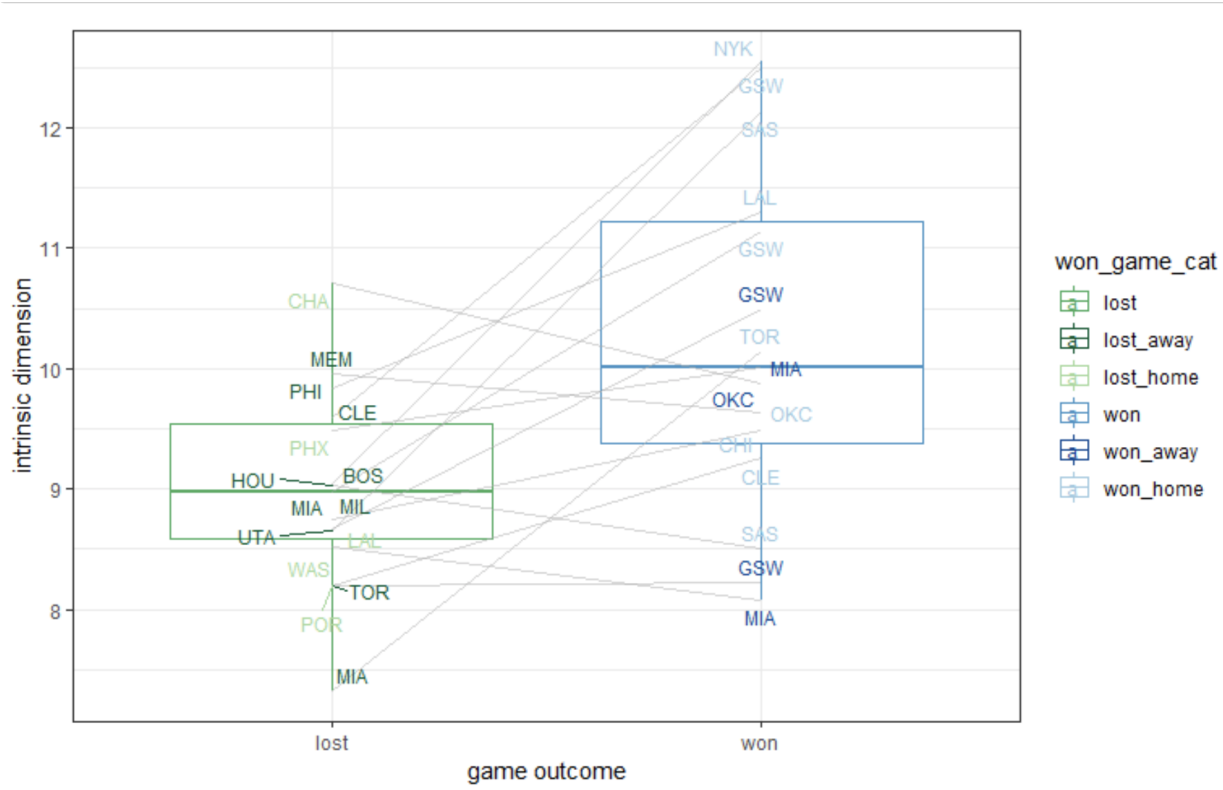
We are applying our methodology to undercover potential patterns in the IDs of the "configuration" on the field of the players when a shot is taken. Example: Golden State Warriors (home - 89) vs Cleveland (away - 83)



BNP Hildago: Application to Basketball data

Joint work with Edgar Santos-Fernandez and Kerrie Mergensen

Results on 15 games from the NBA season 2015-16
RHS: ID of the winning teams



The path to Hidalgo

- The problem of clustering led us to the problem of density estimation
- The problem of density estimation led us to the problem of ID estimation
- We developed a reliable ID estimator, TWO-NN, that limits the issue of density variations
- We realized that often the ID is not constant in the dataset: we extended the statistical framework of TWO-NN to comply with this case
- We developed Hidalgo, a method that finds groups of points (manifolds) of different ID
- Applications of Hidalgo to real datasets reveal that the topological information given by the ID discriminates points differing in important features

References

1. Allegra, Facco, Laio, Mira; *Data classification based on the local intrinsic dimension*, Submitted
2. Facco, d'Errico, Rodriguez, Laio; *Estimating the intrinsic dimension of datasets by a minimal neighborhood information*. Scientific reports 7, 12140 (2017).
3. Rodriguez, d'Errico, Facco, Laio, *Computing the Free Energy without Collective Variables*, J. Chem. Theory Comput. 2018, 14, 1206 1215
4. ...

Acknowledgments

Finance experts: Giovanni Barone-Adesi and Julia Reynolds from Università della Svizzera italiana, Institute of Finance