Bayesian dimensionality reduction via identification of data intrinsic dimensions: theory and applications

Antonietta Mira

Università della Svizzera italiana (USI), Lugano, Svizzera
and University of Insubria, Como, Italy

joint with

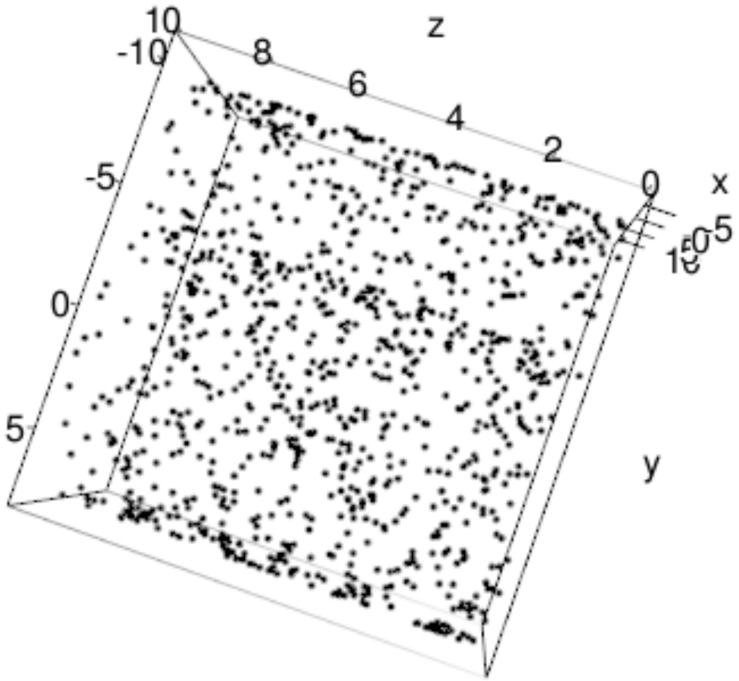
F. Denti
Bicocca University, Milano, Italy and USI, Lugano, Switzerland

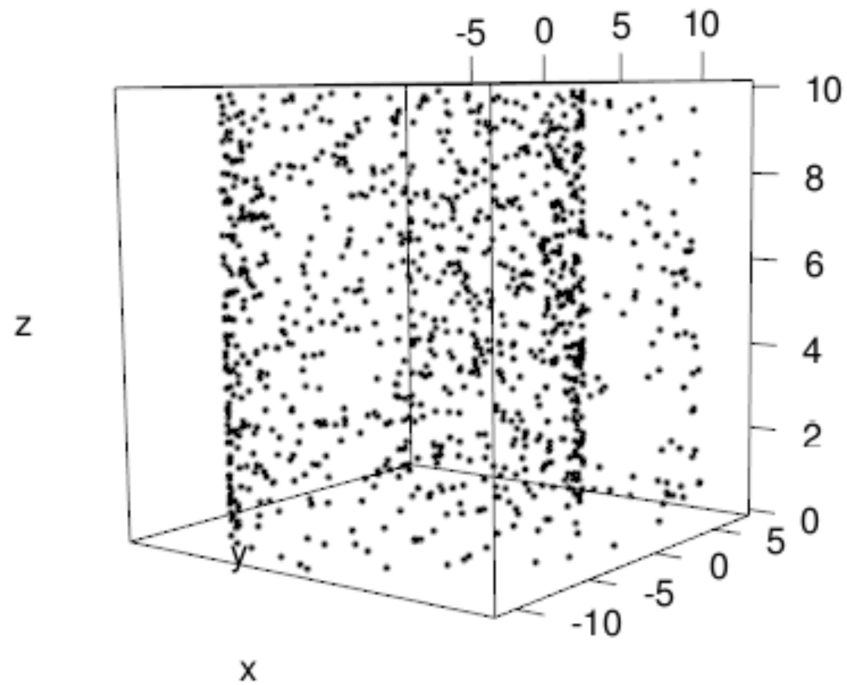
M. Allegra, E. Facco, A. Laio
SISSA, Trieste, Italy
E. Santos-Fernandez, K. Mengersen
School of Mathematical Sciences, QUT

BoB, 24.11.2019

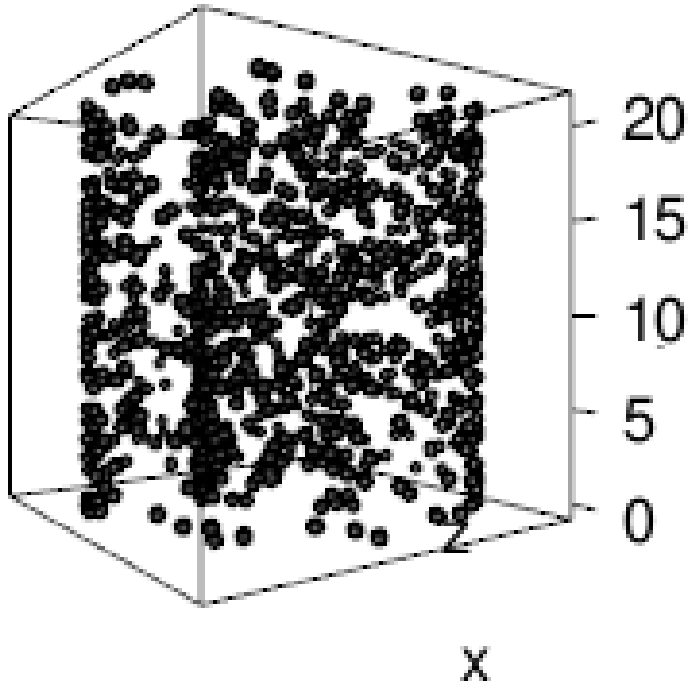
A matter of perspective

Can you guess the data generating mechanism?





And now?



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

Motivation: dimensionality reduction

In high-dimension (D), a small number of variable (d) is often sufficient to effectively describe the data while minimizing the information loss

This number, d, 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:

- folded vs unfolded state in protein configurations
- active vs non-active regions in brain imaging data
- patients vs controls in gene expression data

A simple topological feature uncovers a rich data structure

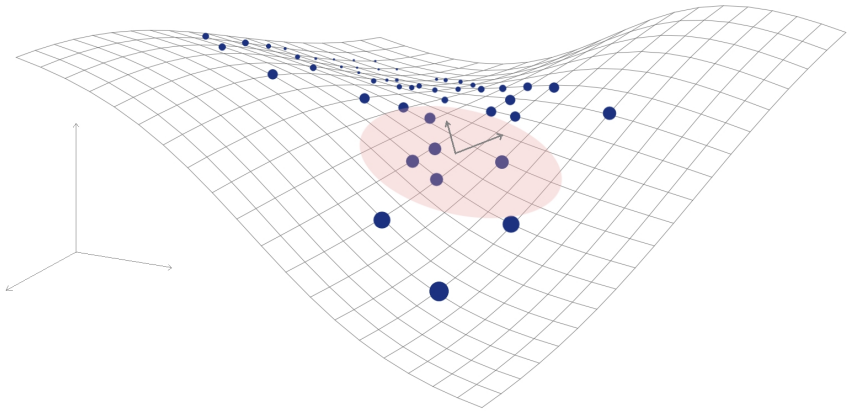
Other applications: finance, sport, MCMC

- firms with different financial risk in balance sheets data
- winning vs loosing teams in basketball data
- identified vs unidentified models in MCMC simulation
- attractors in chaotic systems

Again, a topological feature uncovers a rich data structure

Statistical inference based on Local Intrinsic Dimension

Number of independent directions of variation of the data is $d < D$

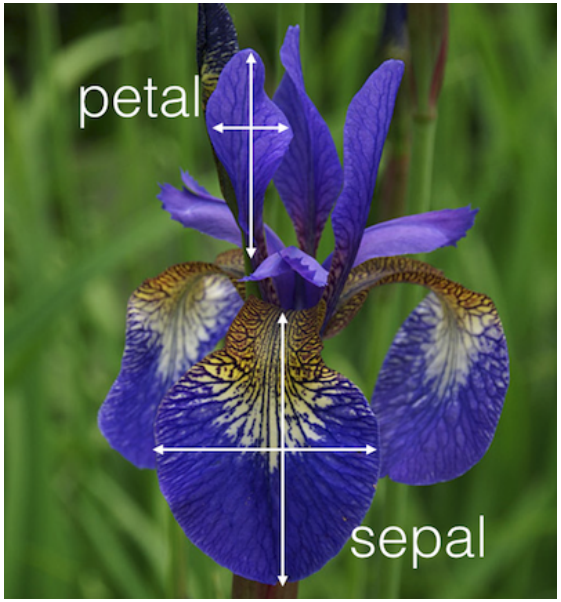


Accounting for the ID can improve statistical analysis such as identification of patterns and classification schemes which are computationally hard in high dimension D (many variables)

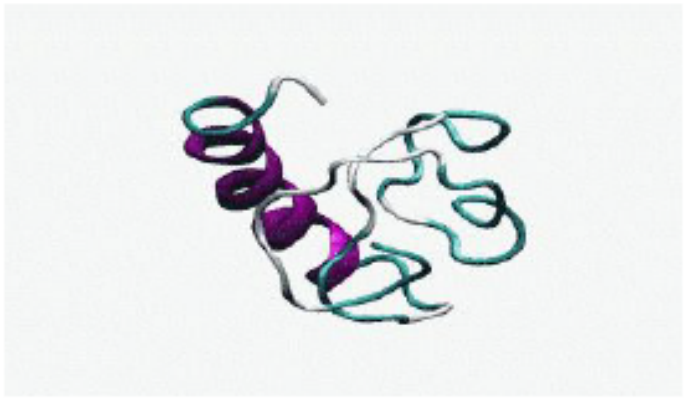
Pre-process step in **any** statistical analysis

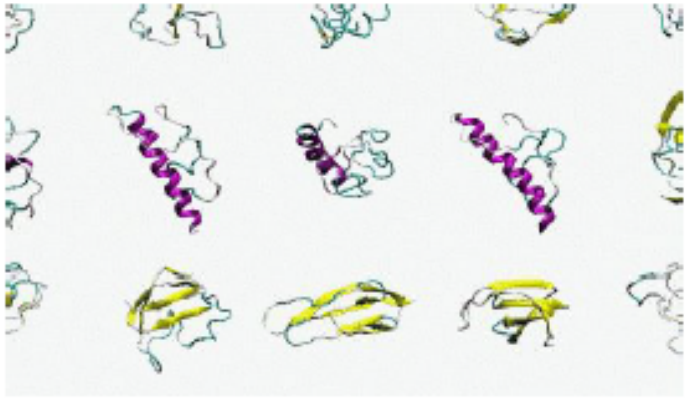
ID Toy Example: Iris Data

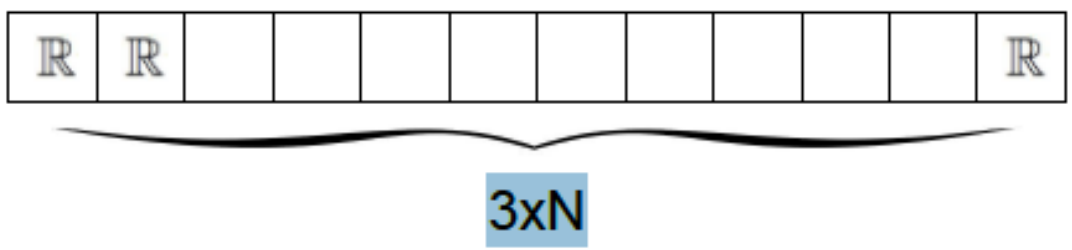
Three types of Iris flowers,
$D = 4$ recorded variables



ID Example I: Molecular Dynamics







How to estimate ID? A (first) Statistical Approach

The data are regarded as a configuration from PP with true intensity function $\rho(x)$
The distances between points in the dataset follow a scaling law that depends on $\rho(x)$ and d

The number of points at distance $< r$ from point i scales as

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

If $\rho(x)$ is everywhere constant:

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

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

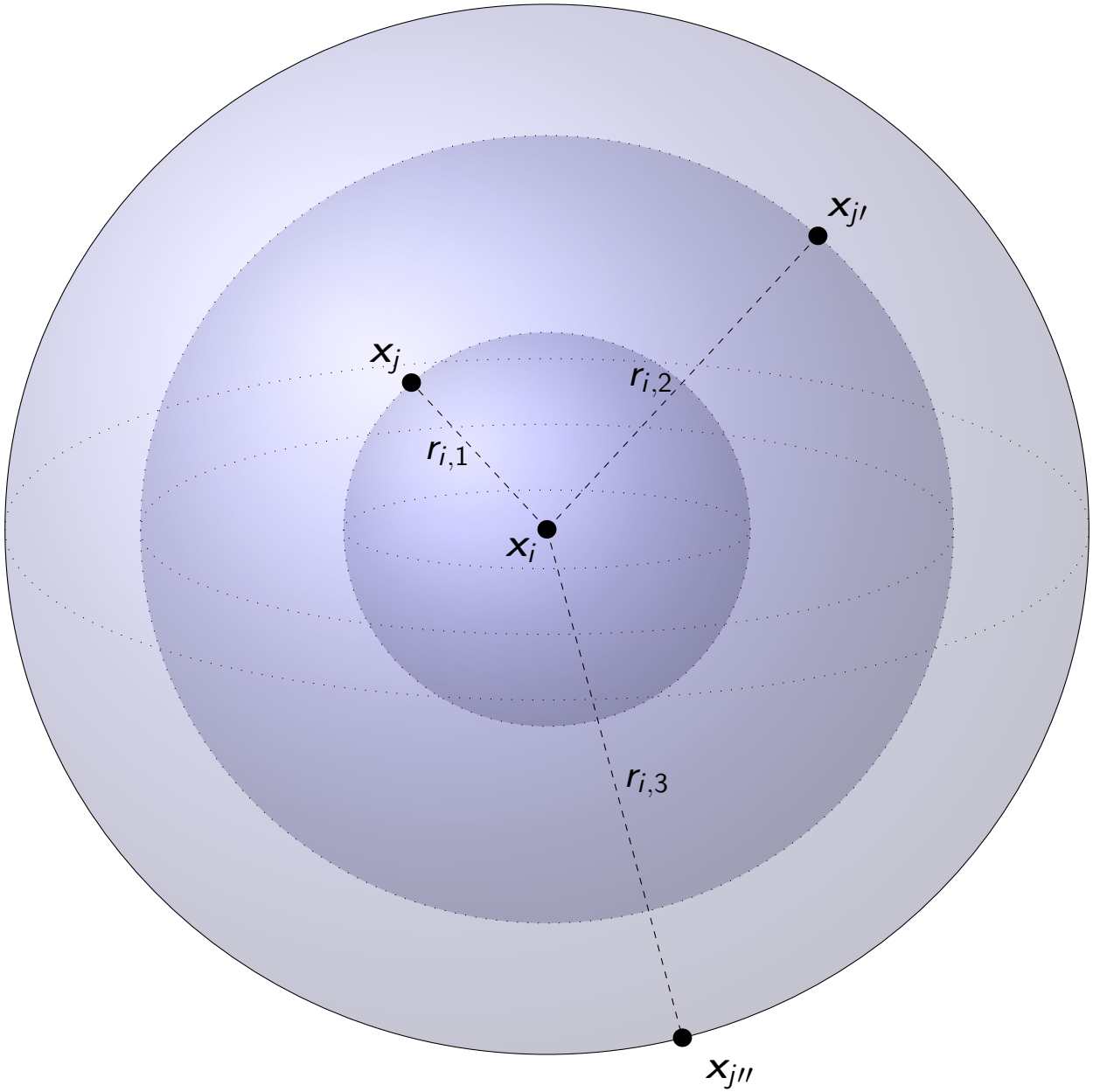
ID: TWO-NN Estimator

Idea: **Find a function of the distances that depends only on d**

Two assumptions:

- data are regarded as a configuration from PP with intensity function $\rho(x)$
- for all x_i, $\rho(x_i)$ is locally constant in the region containing the first 2 neighbors of x_i: locally homogenous PP

A pictorial representation



ID: TWO-NN Estimator

For every point i, 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

Infer the ID from the μ_i of all points collectively
$\boldsymbol{\mu} = (\mu_1, \cdots, \mu_N)$: fit a Pareto distribution
The joint distribution of all μ_i does not depend on ρ (assuming ρ is constant over scale of first 2 neighbors)

There are several ways of fitting:

- Fit empirical cumulative distribution of μ_i: $F(\mu_i) = 1 - \mu_i^{-d}$
- Equivalently, linear fit on $-\log(1 - F(\mu_i)) = d \cdot \log \mu_i$
- Estimate d by MLE in a Pareto distribution: $\hat{d} = \frac{1}{\sum_i log \mu_i}$

If assumptions are satisfied, the distribution of $\boldsymbol{\mu}$ is well fitted

ID: TWO-NN Estimator

Infer the ID from the μ_i of all points collectively
$\boldsymbol{\mu} = (\mu_1, \cdots, \mu_N)$: fit a Pareto distribution
The joint distribution of all μ_i does not depend on ρ (assuming ρ is constant over scale of first 2 neighbors)

There are several ways of fitting:

- Fit empirical cumulative distribution of μ_i: $F(\mu_i) = 1 - \mu_i^{-d}$
- Equivalently, linear fit on $-\log(1 - F(\mu_i)) = d \cdot \log \mu_i$
- Estimate d by MLE in a Pareto distribution: $\hat{d} = \frac{1}{\sum_i log \mu_i}$

If assumptions are satisfied, the distribution of $\boldsymbol{\mu}$ 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

Let's denote $\mu_i = r_{i2}/r_{i1}$ by μ_i^1 and let's remove the i subscript
We prove that:

$$\mu^{j-1} = \frac{r_j}{r_{j-1}} \sim Pareto(1, (j-1)d))$$

and, more importantly, we prove that the random variables $\mu^1, \ldots, \mu^{N-1}$ are all independent

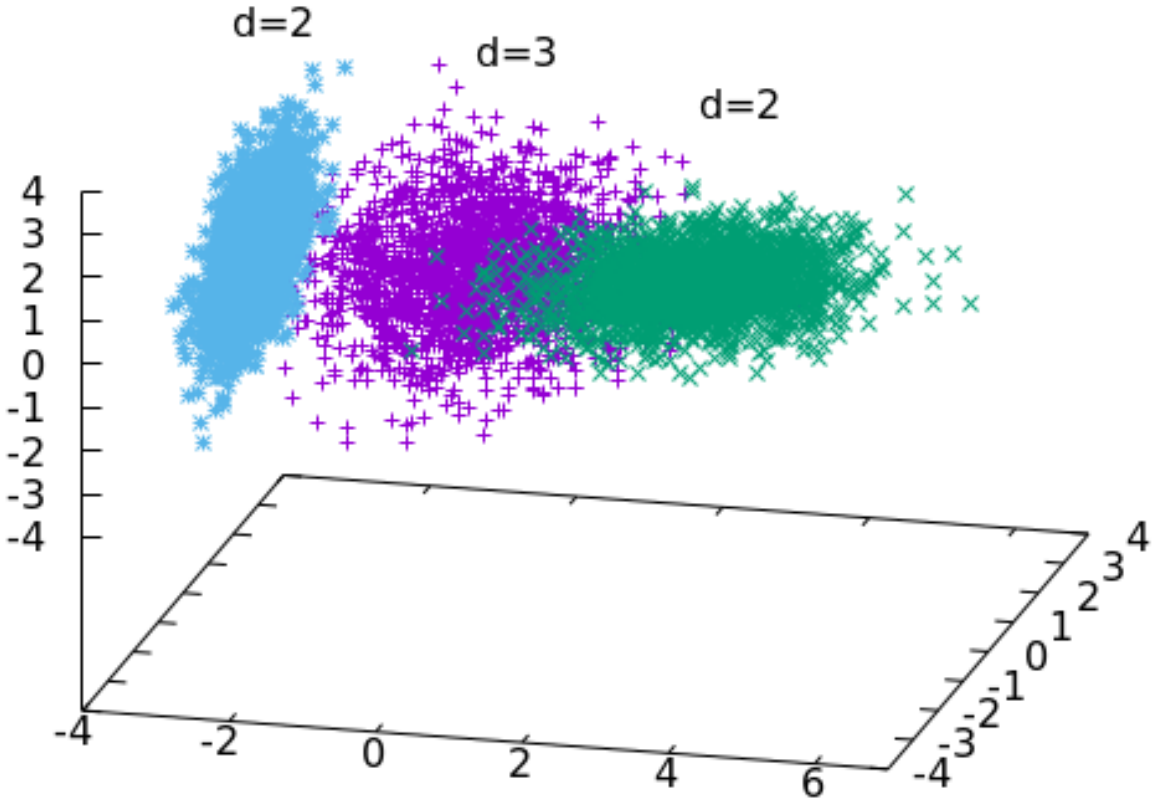
This allows us to extract more info from the data
For example, we can model together

$$(\mu^1, \mu^2) \sim Pareto(1, d) \times Pareto(1, 2d)$$

We can keep adding components to the vector until the homogeneity hypothesis holds

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 Z** $= Z_1, \ldots, Z_N$ which record the manifold membership of each point
Likelihood is seen as marginal of

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

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

Independent priors on $\mathbf{d}$ and $\mathbf{p}$

Prior on $\mathbf{d}$: $\quad d_k \sim Gamma(a_0, b_0), \quad k = 1, \ldots, K$

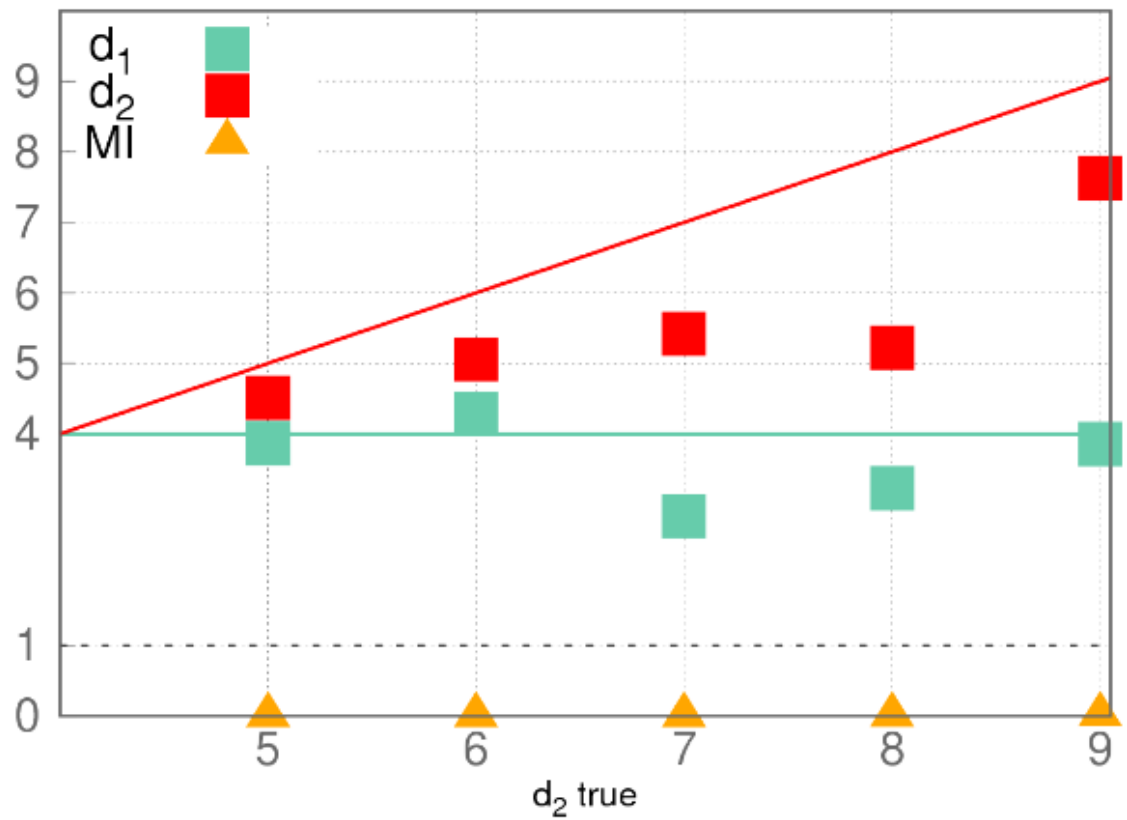
Prior on $\mathbf{p} \sim Dir(\alpha_1, \ldots, \alpha_K)$

Prior on $\mathbf{Z}|\mathbf{p} \sim$discrete distribution on $(1, \ldots K)$ w.p. $\mathbf{p}$

Simulation Study:

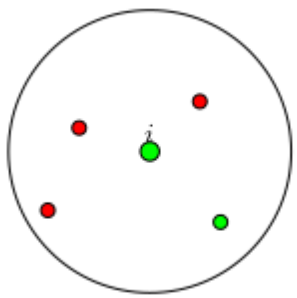
Comparison between two manifolds of dimension $d_1 = 4$ and $d_2 = 5, \ldots, 9$ with $\rho(x) =$ Gaussian

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



One more necessary assumption

The clustering induced in the data by the latent variables $\mathbf{z}$ plays a key role: the observations within each group concur to the estimation of a different value of d_{z_i}

If the clustering is inaccurate, so is the estimate

It is crucial to include a source of local homogeneity in the model, and this can be obtained via the following

Assumption: the different manifolds are separated in the space, and the neighborhood of a point should be more likely to contain points sampled from the same manifold than points sampled from a different manifold.

Therefore, we propose to extract from the original data $\mathbf{x}$ another source of information that can be used to penalize for local inhomogeneities: the $n \times n$ proximity matrix $\mathcal{N}^{(q)}$

The (i, j) entry $\mathcal{N}_{ij}^{(q)}$ of this binary matrix is
1 only if the observation j is one of the first q NNs of observation i
0 otherwise
Notice that $\sum_j \mathcal{N}_{ij}^{(q)} = q$
To induce local uniformity, we model

$$f\left(\mathcal{N}_{ij}^{(q)} = 1 | z_i = z_j\right) \propto \zeta_0$$

$$f\left(\mathcal{N}_{ij}^{(q)} = 1 | z_i \neq z_j\right) \propto \zeta_1$$

where $\zeta_0 > 0.5$ and $\zeta_1 < 0.5$
These inequalities imply that points assigned to the same manifold have more chances to be neighbors

For simplicity, we set $\zeta_0 = \zeta$ and $\zeta_1 = 1 - \zeta$

Denote with $\mathcal{N}_i^{(q)}$ the i-th row of the adjacency matrix
We regard $\mathcal{N}_i^{(q)}$ as independent and model them as:

$$f\left(\mathcal{N}^{(q)}|\mathbf{z},\zeta\right) = \prod_i f\left(\mathcal{N}_i^{(q)}|\mathbf{z},\zeta\right) = \prod_i \frac{\zeta^{n_i^{in}(\mathbf{z})}(1-\zeta)^{q-n_i^{in}(\mathbf{z})}}{\mathcal{Z}} \quad (1)$$

with $\zeta \in (0.5, 1)$ is the parameter enforcing uniformity between neighbors ($\zeta = 0.5$ implies no additional term in the likelihood)

$n_i^{in}(\mathbf{z}) = \sum_j n_{ij}\mathbb{I}_{z_j=z_i}$ is the number of the q NNs of $\mathbf{x}_i$ that are clustered together with observation i

$\mathcal{Z}$ is the normalization constant

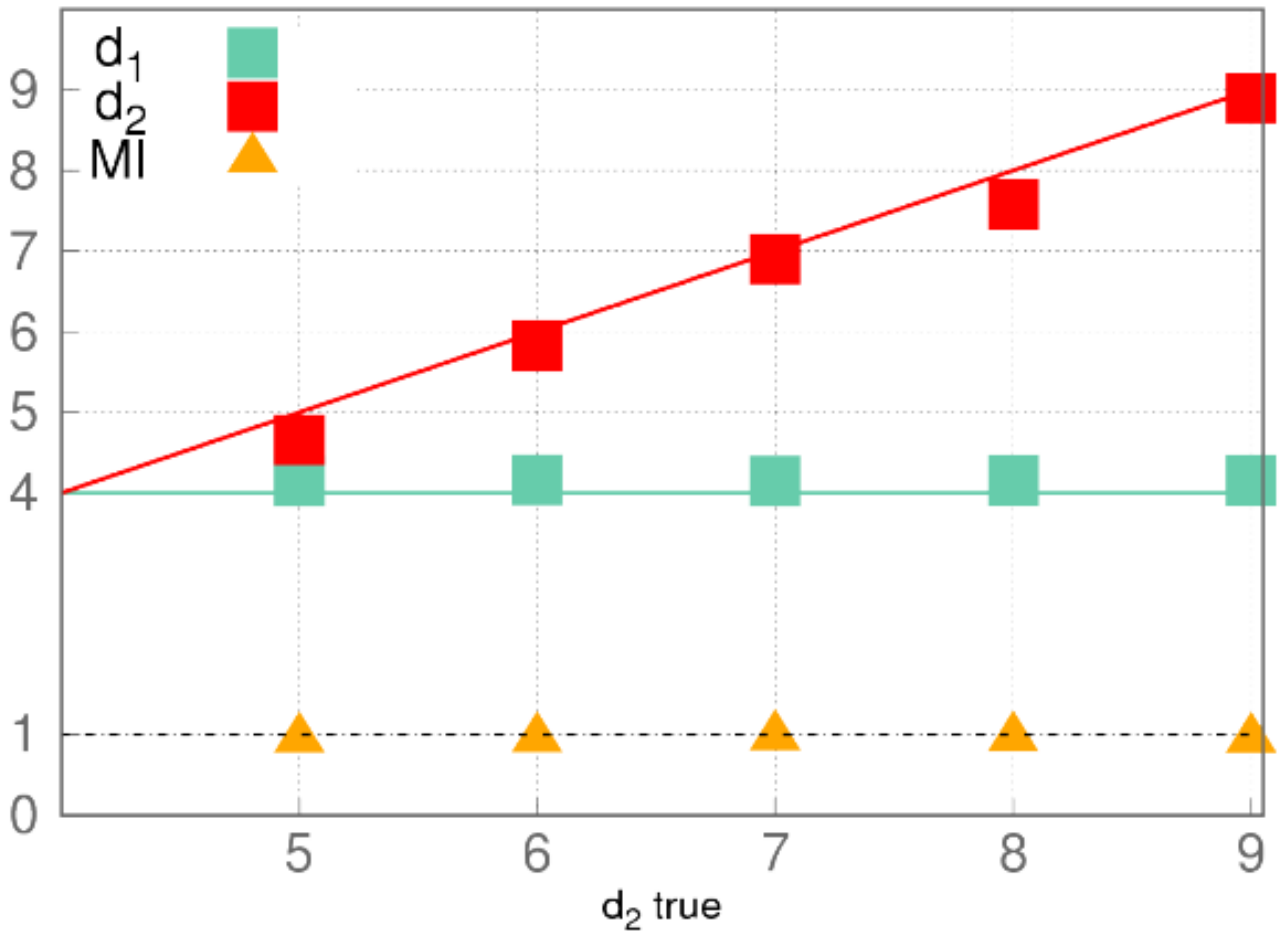
This additional term removes the independence between the cluster labels and helps better estimating the ID

The resulting likelihood for $\mu = (\mu_1, \ldots, \mu_n)$ is

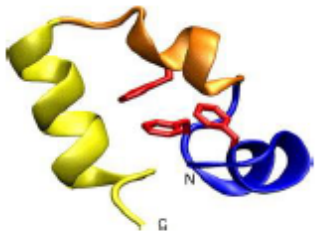
$$\mathcal{L}(\mu|\mathbf{d}, \mathbf{z}, \zeta) = \prod_{i=1}^{n} \mathcal{P}(\mu_i|d_{z_i}) \times f\left(\mathcal{N}_i^{(q)}|\mathbf{z}, \zeta\right). \tag{2}$$

Enforcing uniform neighborhoods

Thanks to this additional term in the LHD we get correct estimates of $\mathbf{d}, \mathbf{p}$ and $\mathbf{Z}$:



Example: molecular dynamics



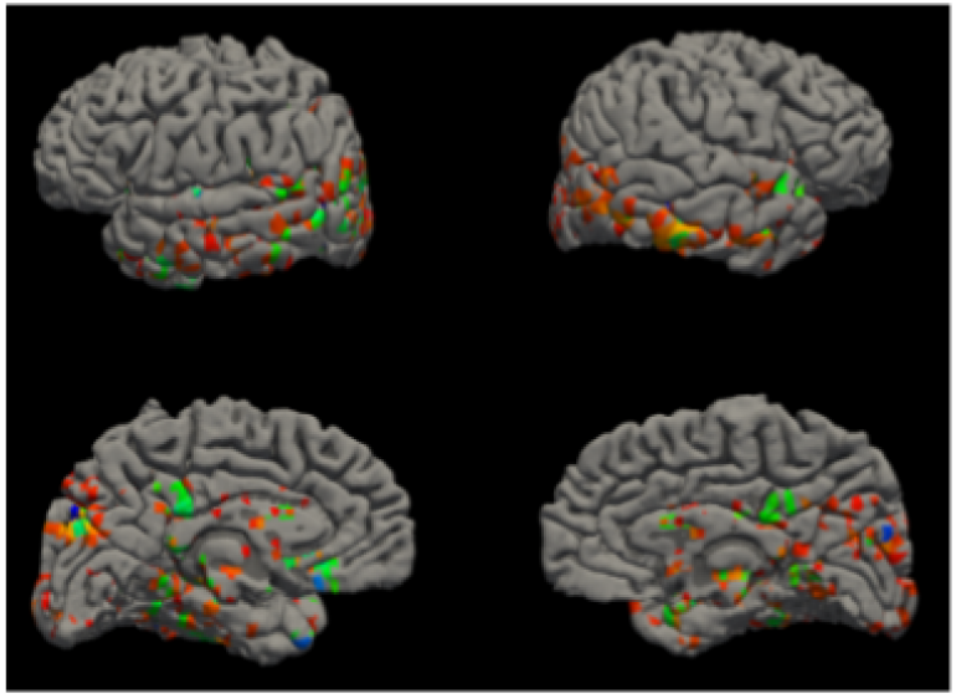
- consider a MD of unfolding/refolding villing headpiece
- for each of the N ~ 32000 configurations, D=32 dihedral angles.

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!

fMRI time series: ID-based classification on the BOLD time series of $\sim 30'000$ voxels in an fMRI experiment with 202 scans.
We find two manifolds $d = 16$ and $d = 32$



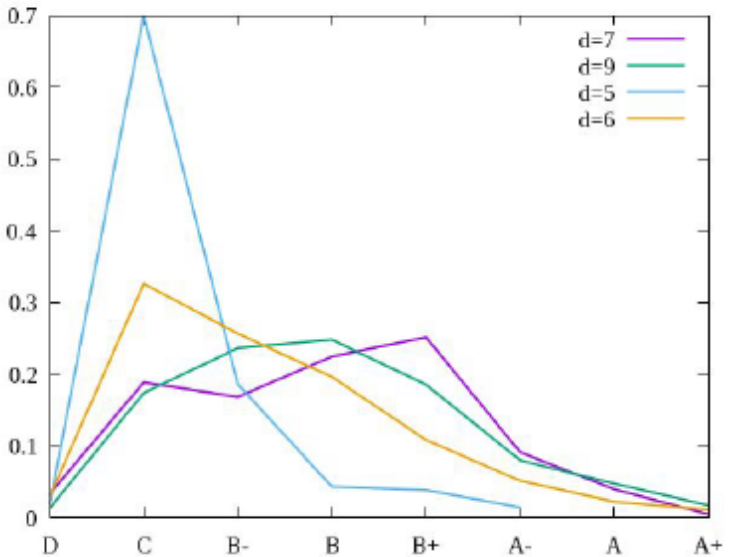
Red: high-ID voxels, Blue: task relevant volxels, Green: intersections
Task-relevant voxels are in the manifold with higher ID, while the low-dimensional manifold mostly includes "noise" voxels

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!

- 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 shortest 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

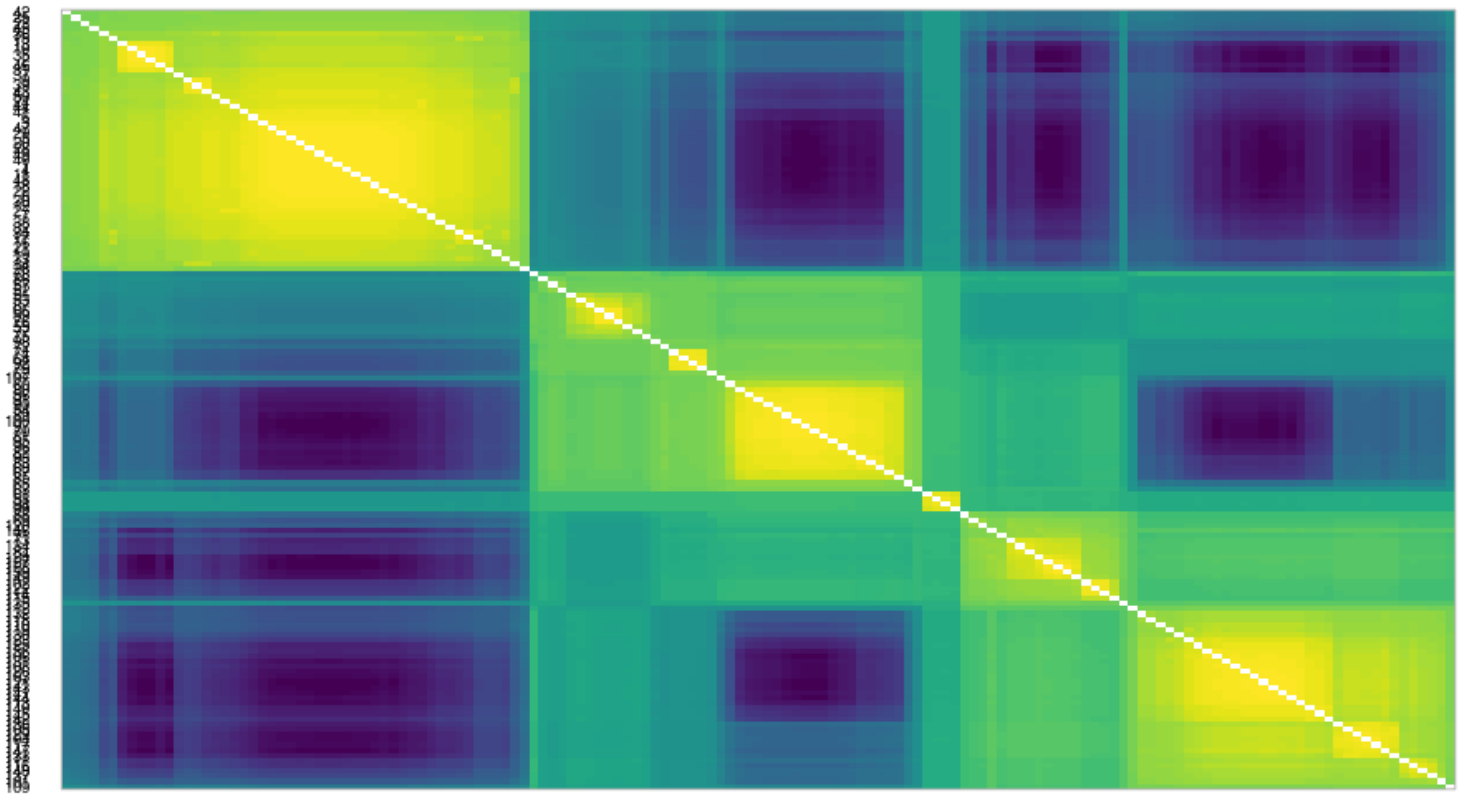
If we introduce a latent variable $\mathbf{Z}$ which denotes, for every observation, the assigned component of the mixture, we can rewrite the model as:

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

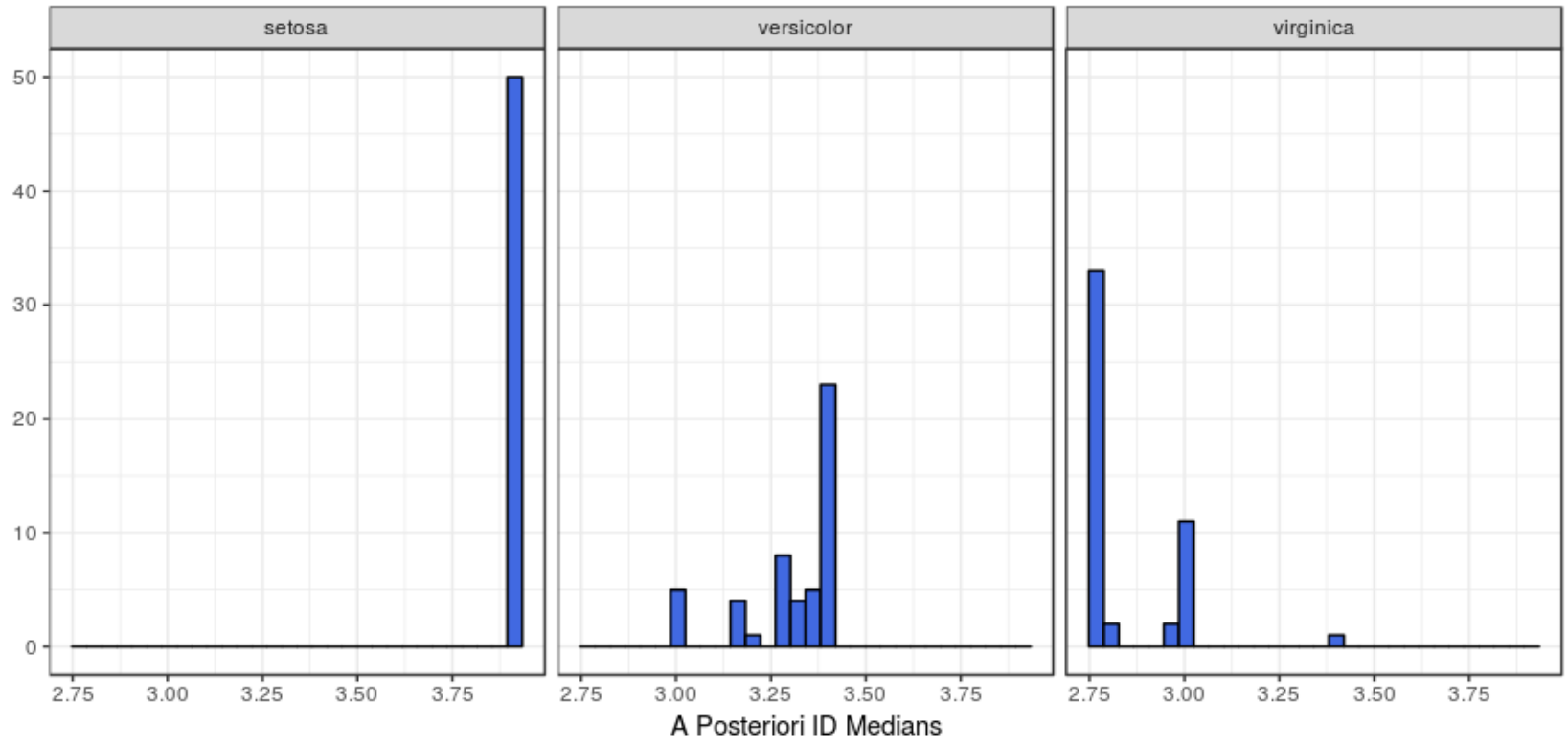
where with SB we denote the usual stick breaking prior, G_0 is a $Gamma(\alpha, \beta)$ and Q is a distribution with a density defined as $\mathcal{L}\left(n^{in}|\mathbf{Z}\right) = \prod_i \frac{\zeta^{n_i^{in}}(1-\zeta)^{n_i^{out}}}{\mathcal{Z}}$

BNP Hildago: Toy Example - Iris

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 $K = 3$ clusters, almost coincident with the Flower Species: Setosa - Versicolor - Virginica
Here is the Pairwise Coclustering Probability Matrix



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

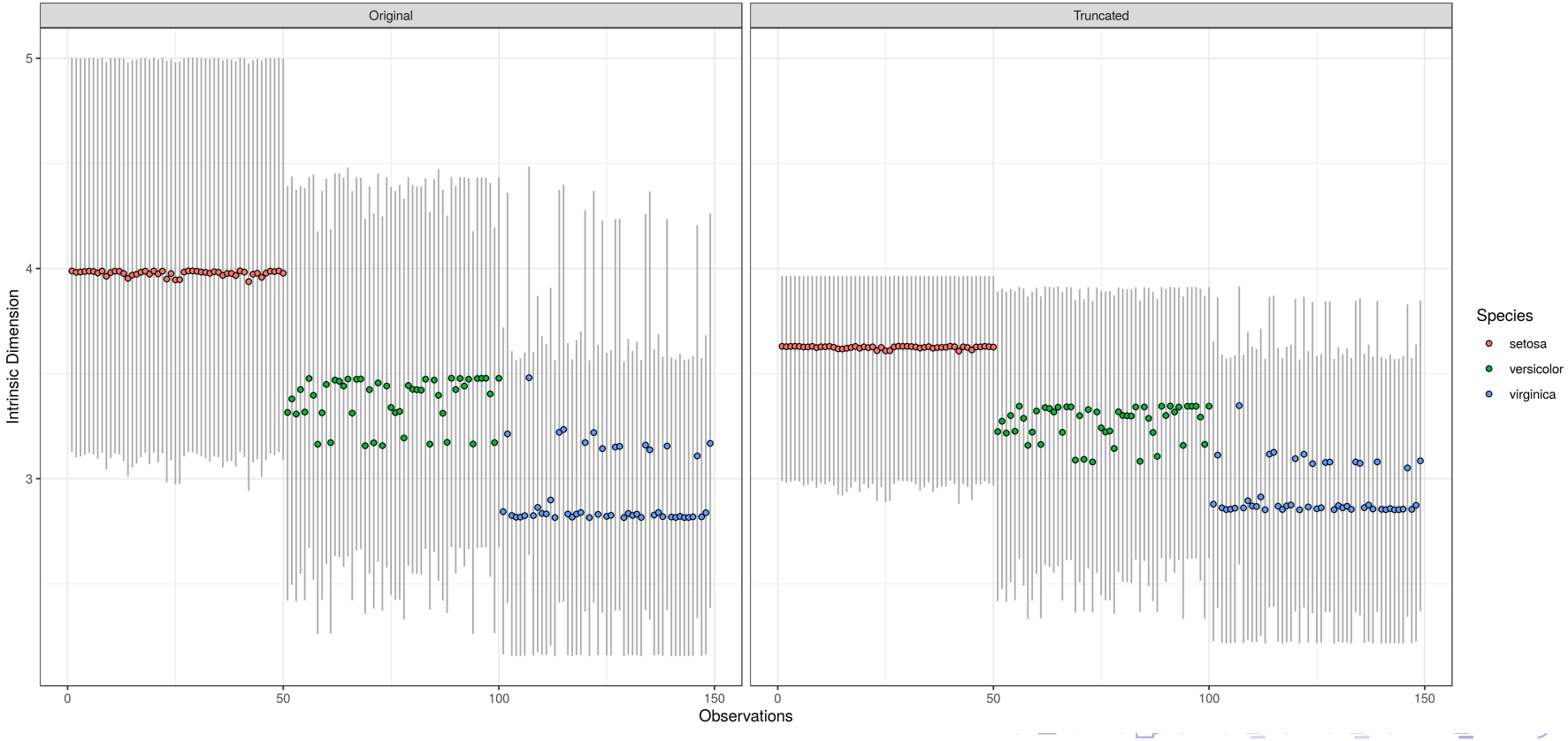


We can conclude that the measurements of the `Versicolor` and `Virginica` Iris are embedded in manifolds with ID, $d < 4$

From Gamma to truncated Gamma prior

It is interesting how the three different Species of flowers show different intrinsic dimensions.

Problem : the estimated ID for such a small dataset is above the maximum dimension D. We propose to substitute the Gamma prior on d_k with a **Truncated Gamma** over $(0, D)$.



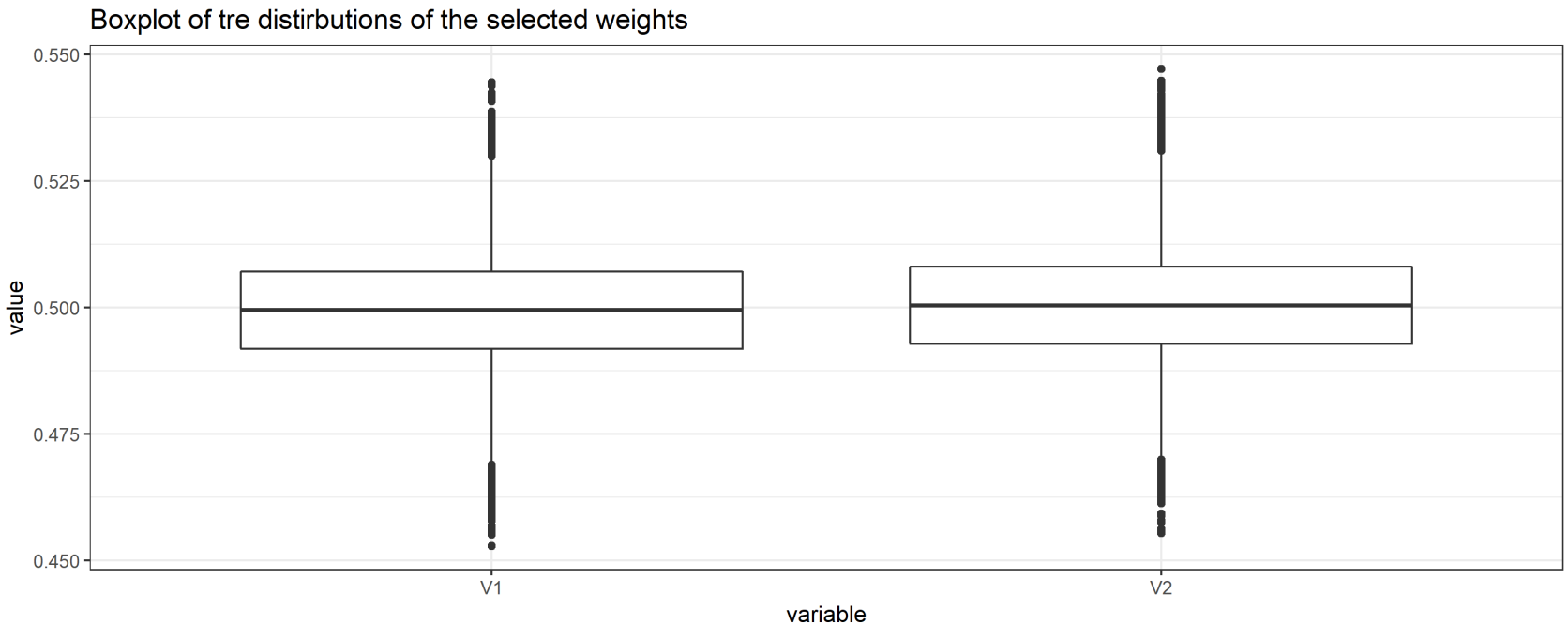
Benchmark dataset:

1000 observation from a 4-dimensional Gaussian,
1000 observations from a 5-dimensional Gaussian,
both with unitary variance

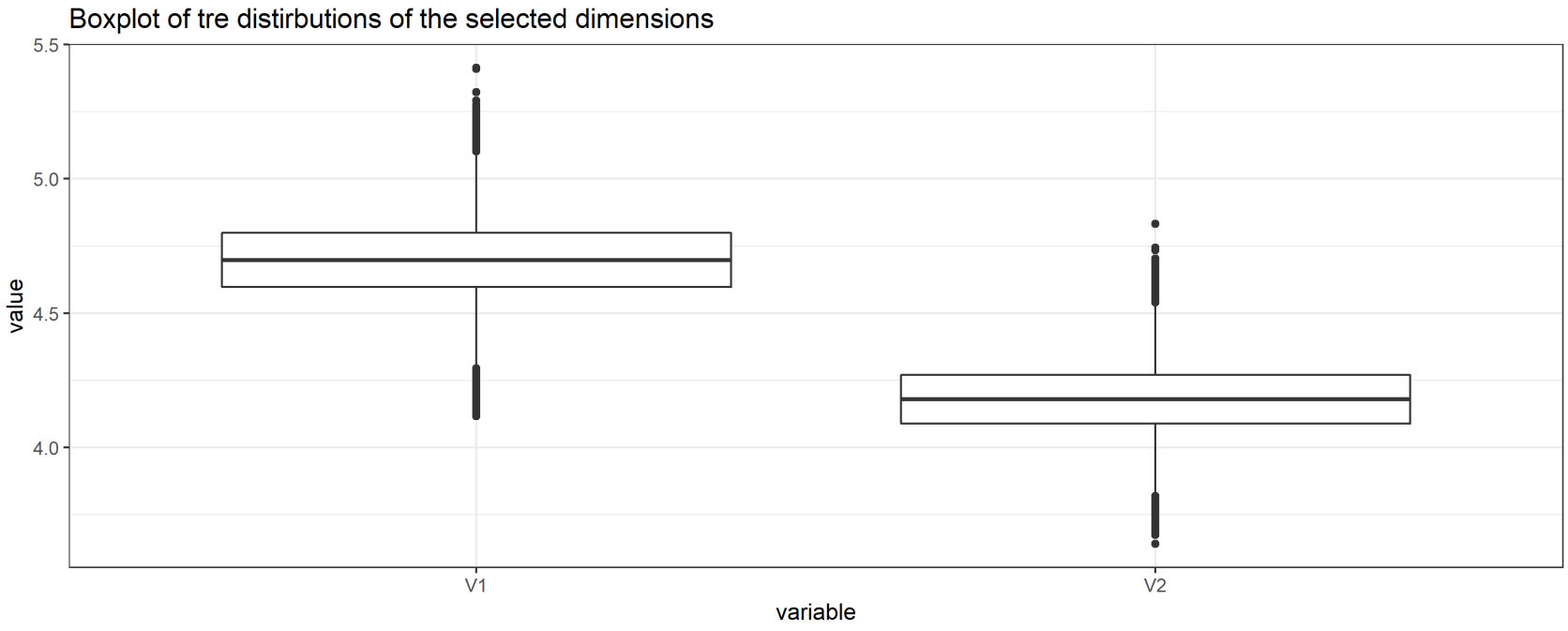
The centroids are chosen to be at a distance from each other of 0.5, challenging the model with overlapping data

Estimated data weights

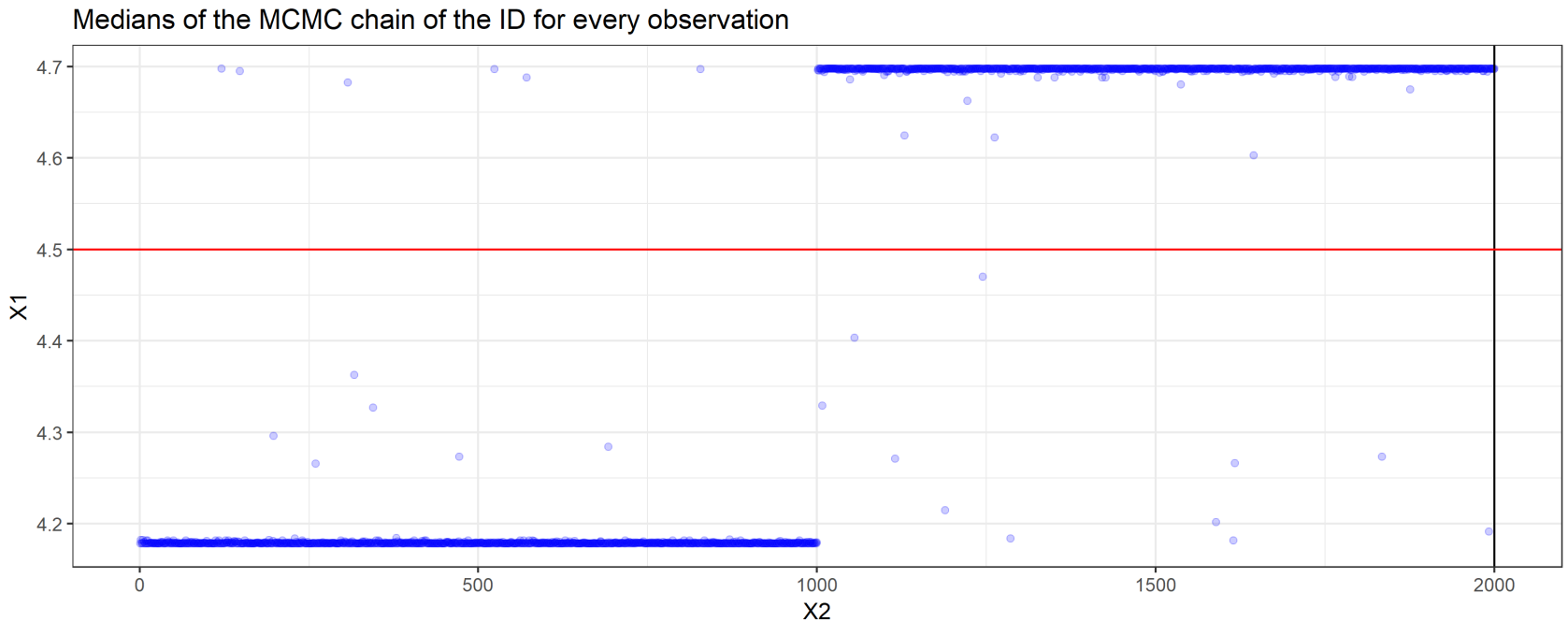
We set the number of neighbors $q = 3$
After a Burn-in period of 20k iterations, 10k samples are retained for posterior inference:



Estimated ID



Estimated data membership



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

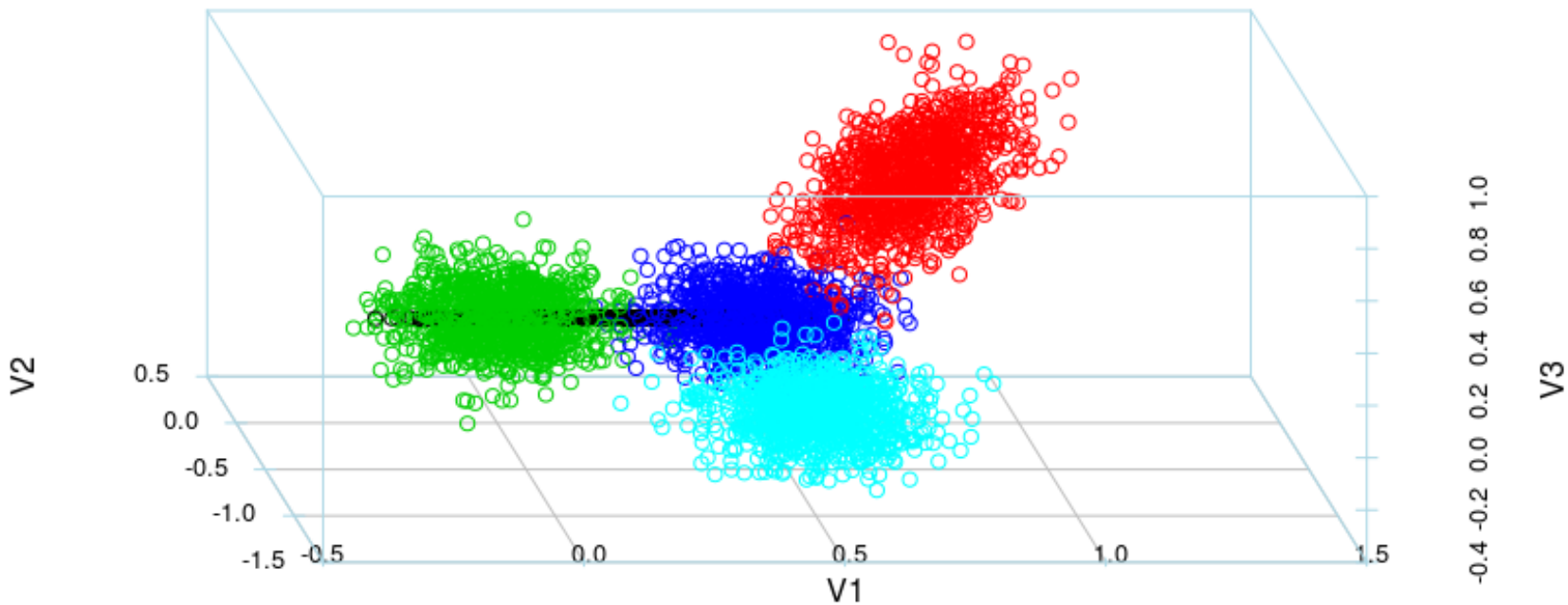
$$\hat{d}_i = \begin{cases} 4 & \text{if } \tilde{d}_i < 4.5 \\ 5 & \text{if } \tilde{d}_i \geq 4.5 \end{cases}$$

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

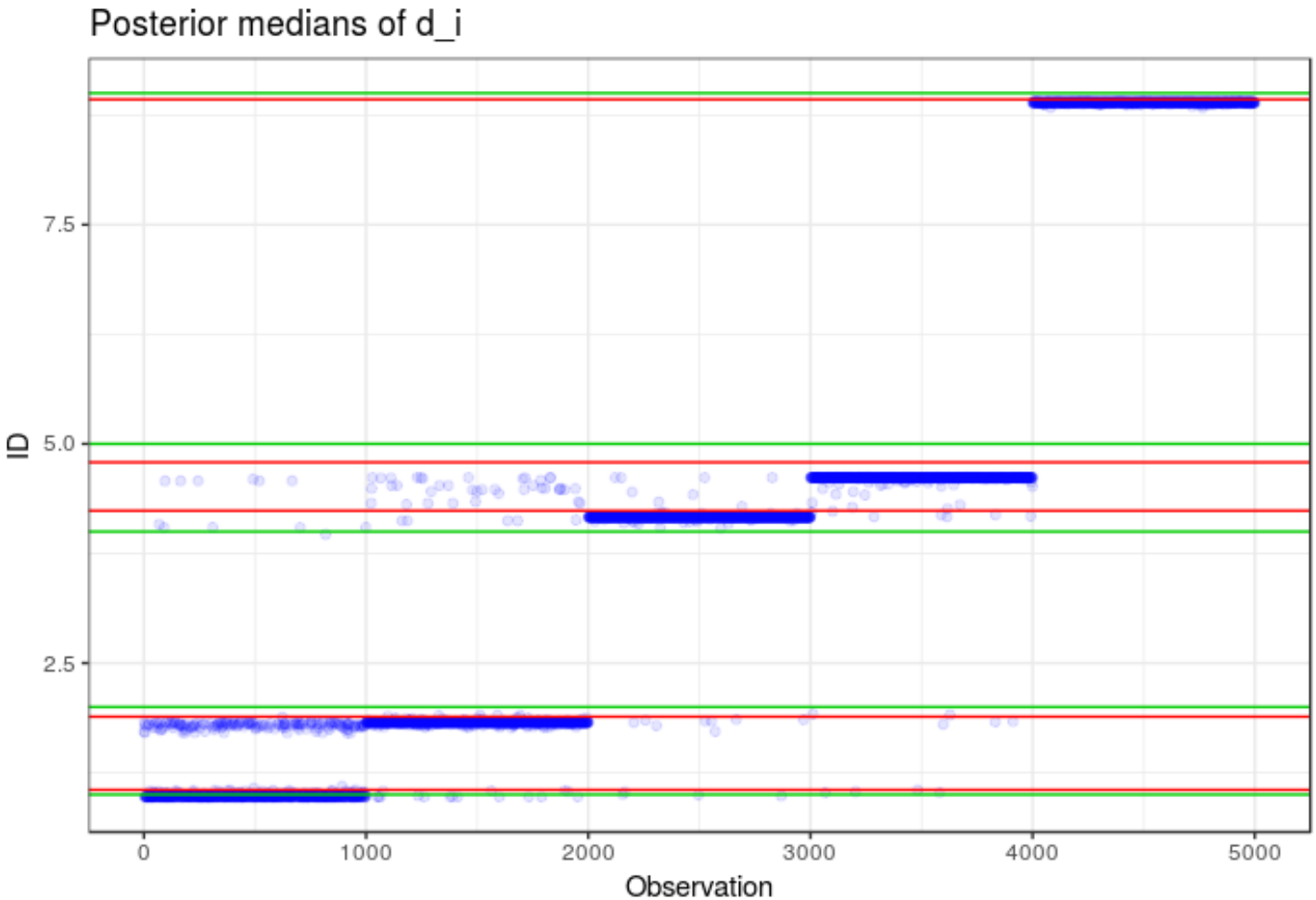
$\hat{d}_i$ vs d_i	4	5
4	**994**	**6**
5	**11**	**989**

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:



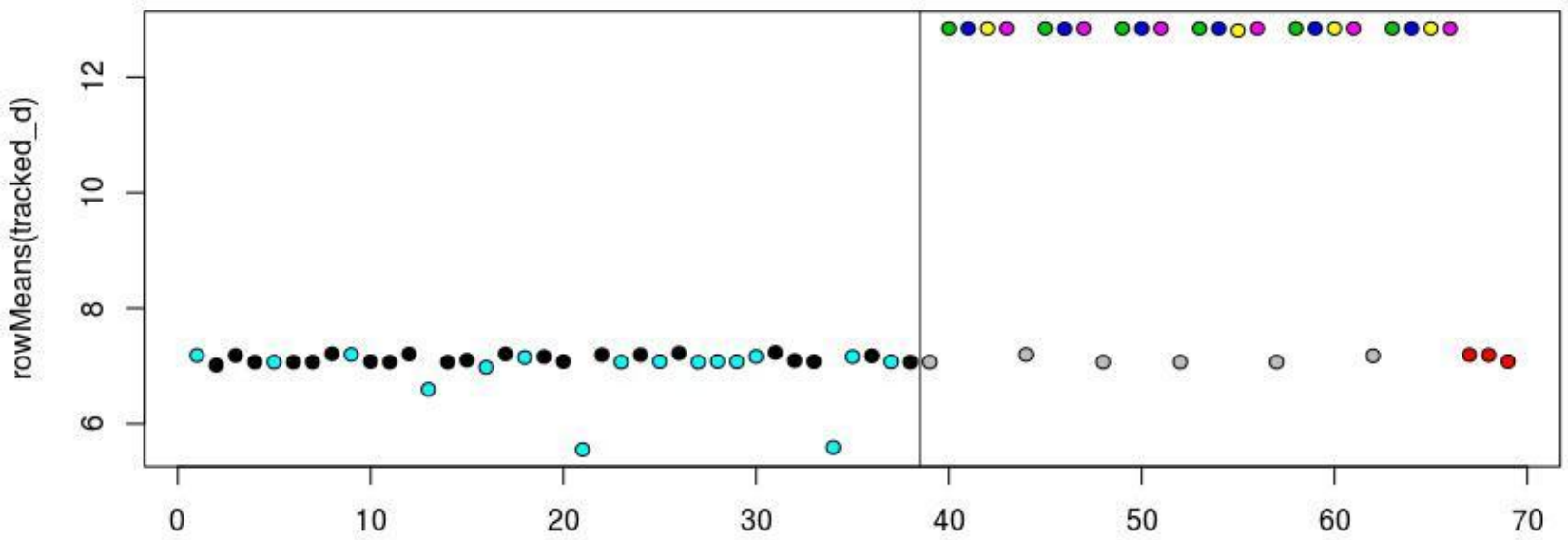
The model is able to estimate the IDs that are in the dataset
The red lines denote the MLE estimates in every subgroup
The green lines denote the actual ID



BNP Hildago: Application to gene expression data

Joint work with Luciano Cascione, Institute of Oncology Research, USI

- Data : $\approx$ 16900 genes expressions recorded for 69 samples of tissue
- The first 38 samples are affected by *Diffuse large B-cell lymphoma* - **DLBCL**. The remaining 31 constitute the control group



- Plot: 69 tissues colored by cell typology
  The vertical line divides case / control groups
- The y-axis: **posteriori medians** of the ID for each tissue sample

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:

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

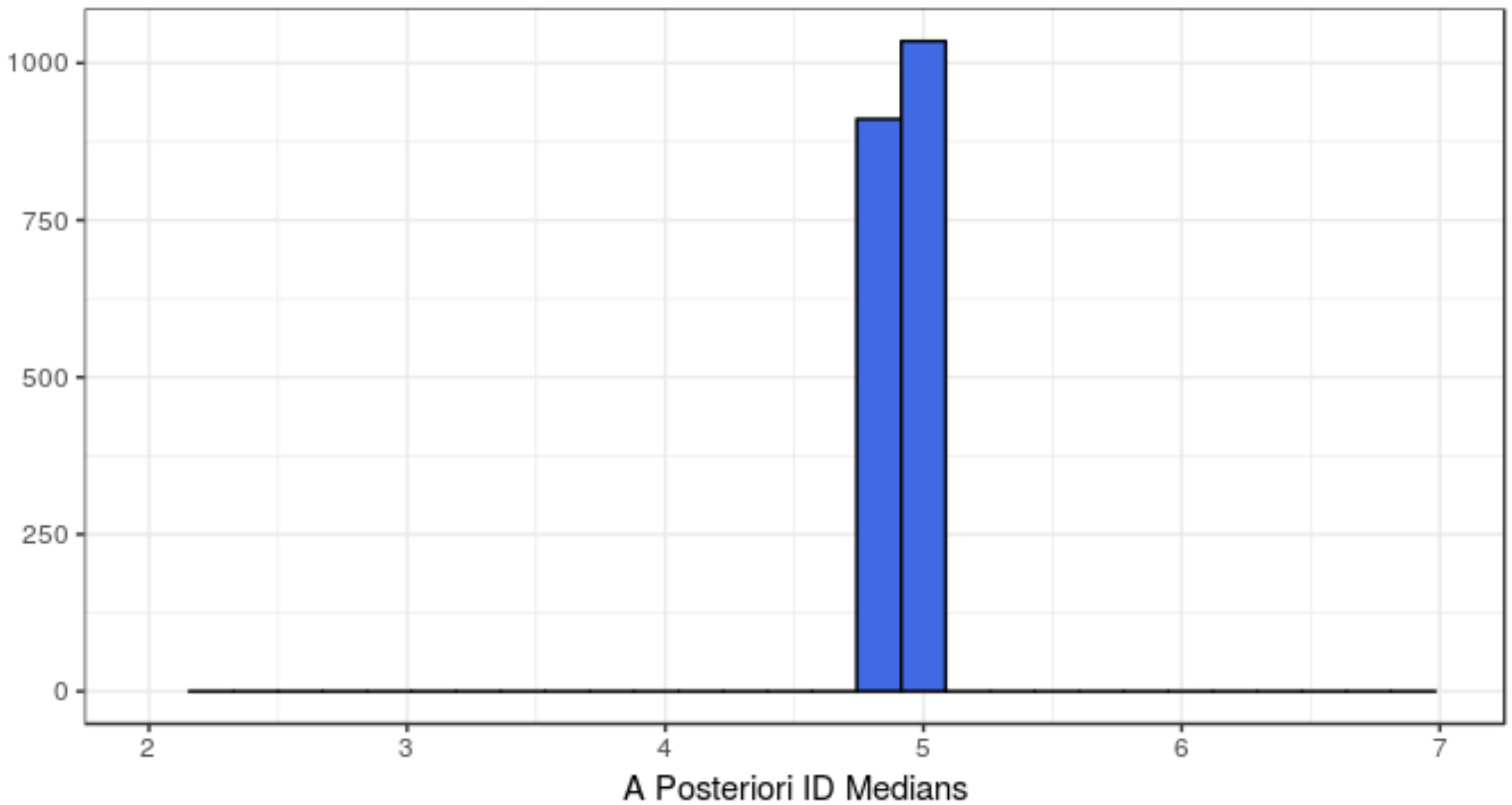
and its ill posed version:

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

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

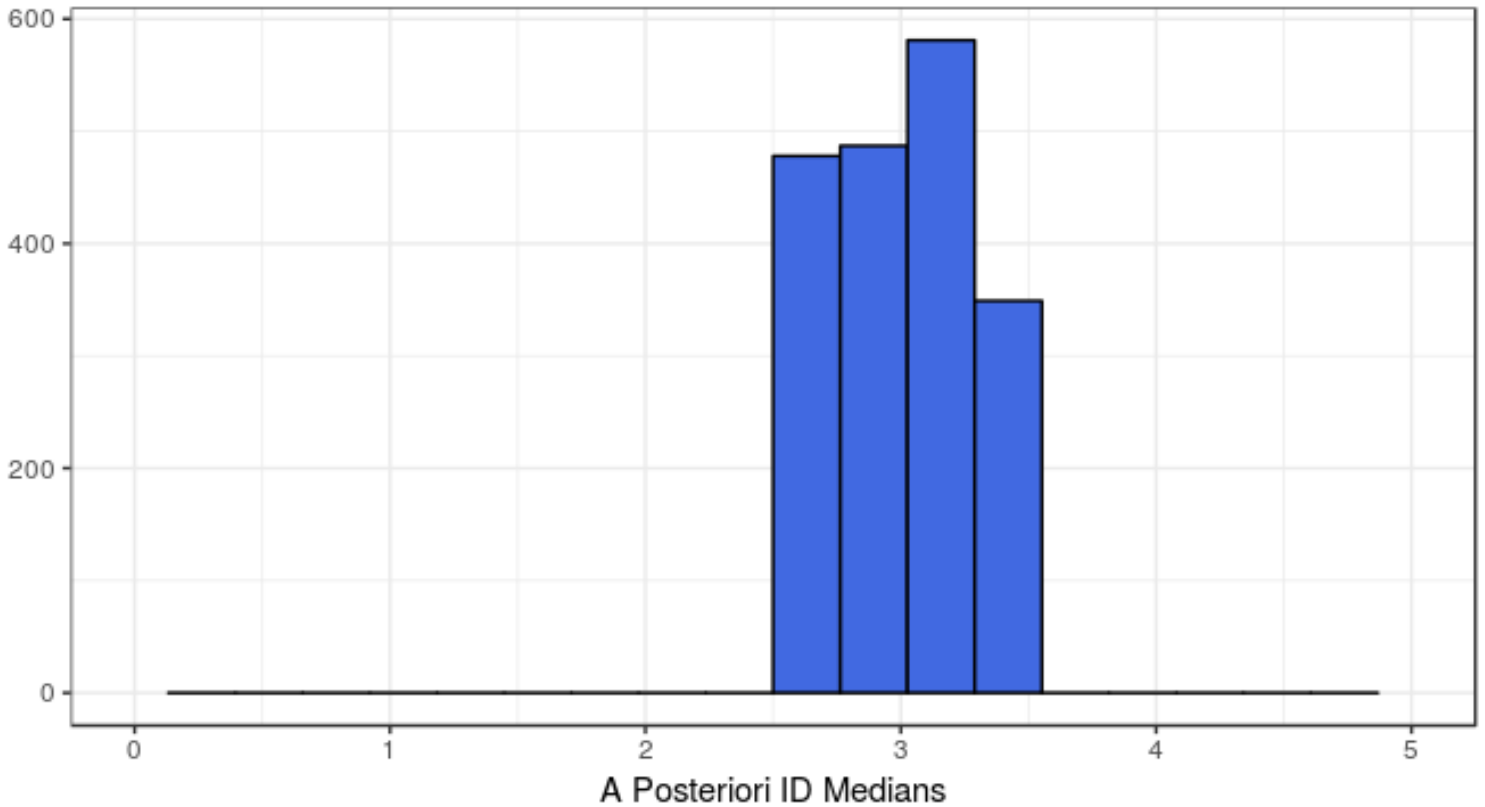
We estimate the model using the Hamiltonian no u-turn sampler (R package `Rstan`)
Every iteration of the Hamiltonian chain is treated as an observation embedded in $\mathbb{R}^5$. After computing the corresponding μ_i's, a MCMC sample of $T = 2k$ iterations is collected after a burn in of $B = 2k$ steps

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

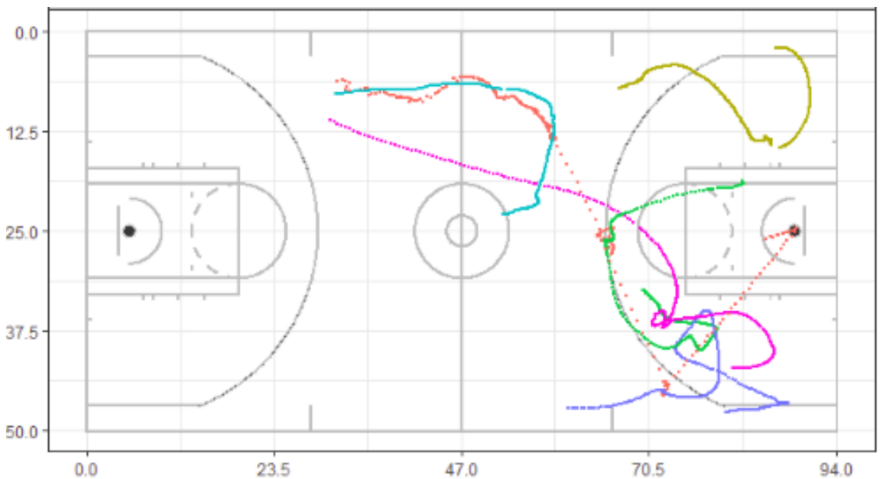


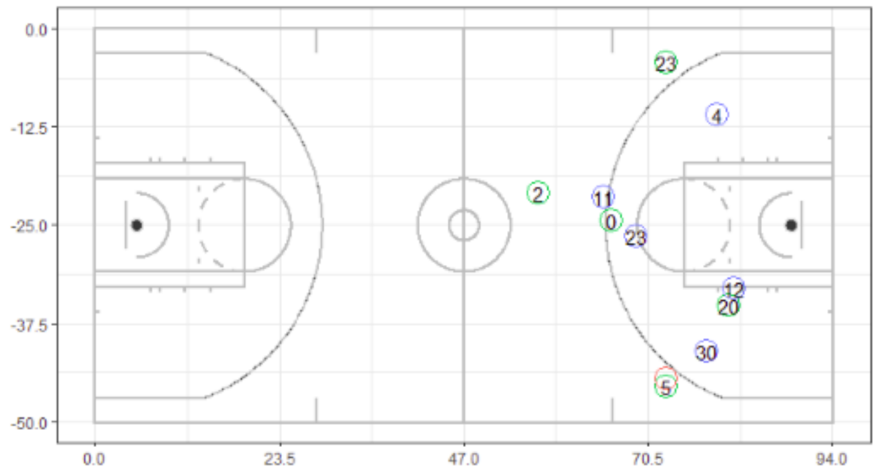
Our approach detects that two dimensions out of five are actually redundant

BNP Hildago: Application to Basketball data

Joint work with Edgar Santos-Fernandez and Kerrie Mengersen (QUT)

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)

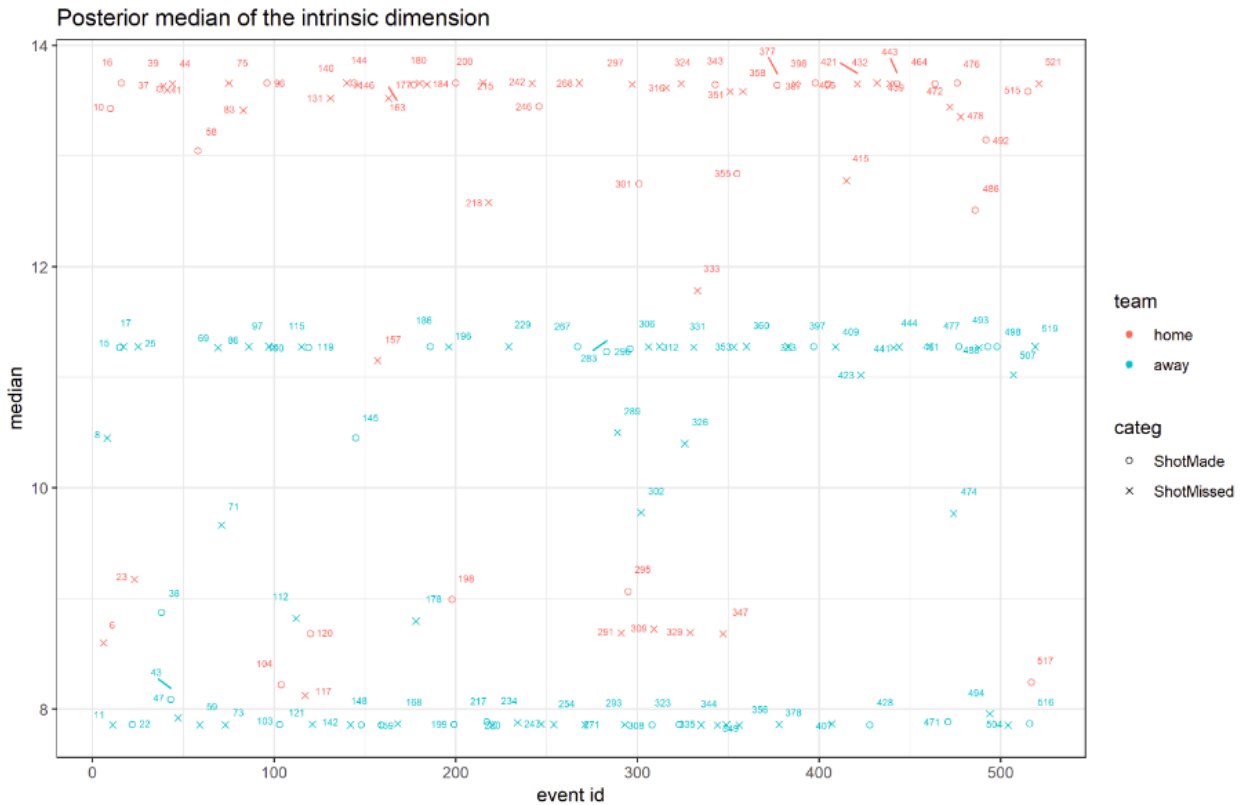




BNP Hildago: Application to Basketball data

Joint work with Edgar Santos-Fernandez and Kerrie 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: Cleveland vs Golden State Warriors



The path to Hidalgo

- CLUSTERING: Science 2014, Rodriguez and Laio
- TWO-NN: Nature Scientific Report 2017, a reliable ID estimator when ID is constant
- 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
- BNP-HIDALGO: from finite to infinite mixtures

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. ...