Cluster analysis on high dimensional data

Roland Winkler (roland.winkler@dlr.de), Frank Klawonn, Rudolf Kruse

November 20, 2010

Deutsches Zentrum
DLR für Luft- und Raumfahrt e.V.
in der Helmholtz-Gemeinschaft

How much Melon do you get for your Money?

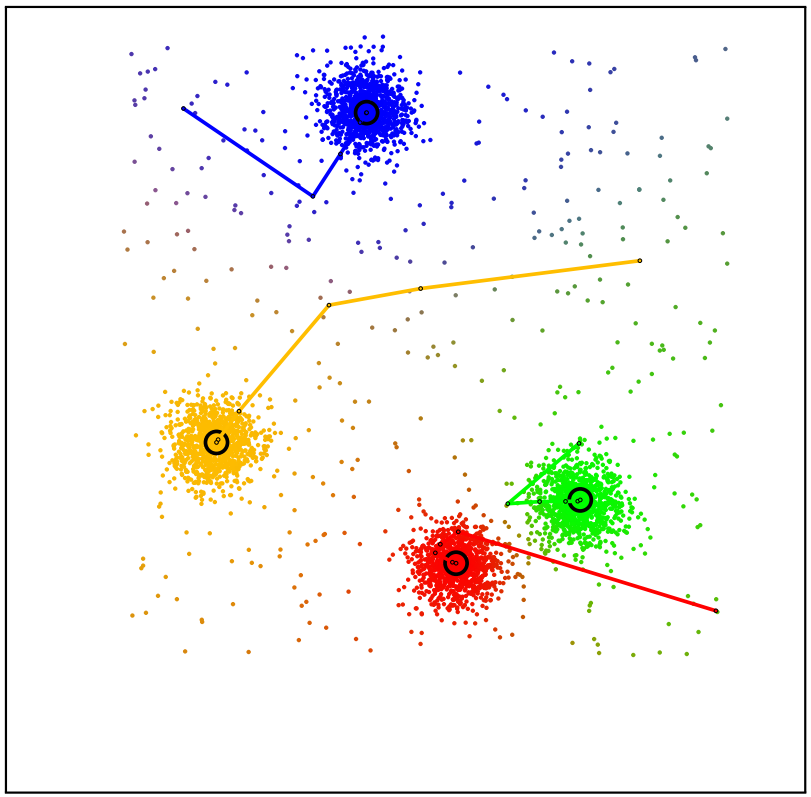
- Imagine an n-dimensional, perfectly spherical watermelon M with radius of $r = 20cm$ and a white skin part of $2cm$ thickness
- The price of the melon is linear to its hypervolume
- Interested in the eatable, red part of the melon $\bar{R}_n = \frac{V_n(18)}{V_n(20)}$
- $\bar{R}_3 = 0.73$, $\bar{R}_7 = 0.48$, $\bar{R}_{20} = 0.12$, $\bar{R}_{50} = 0.005$, $\bar{R}_{100} = 27 \cdot 10^{-6}$

Current Section

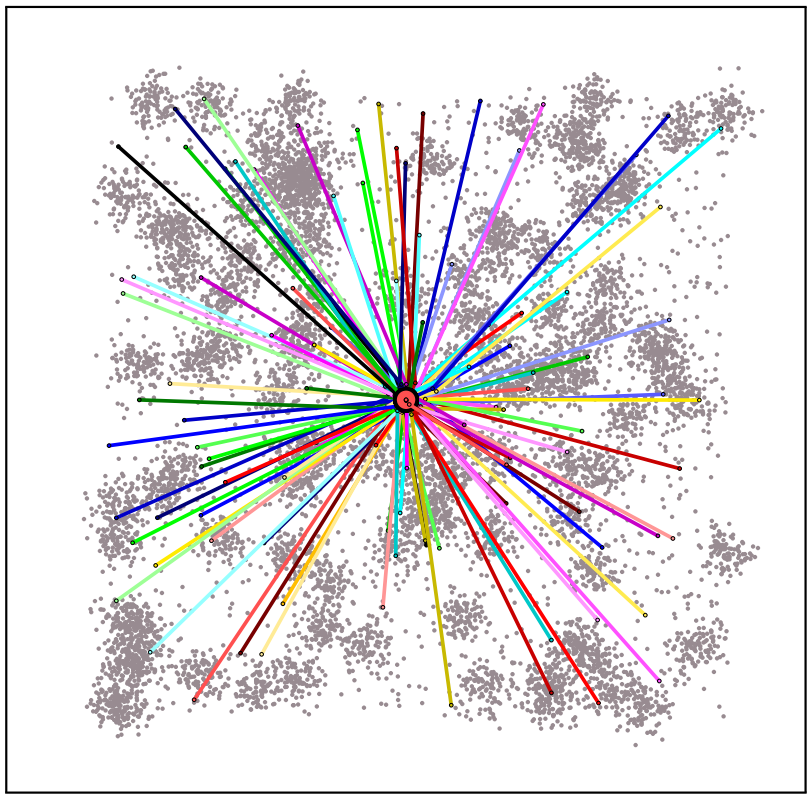
Clustering

”Data Objects of the same cluster should be as similar as possible while data objects of different clusters should be as different as possible.”

Example: FCM



Data set 1: 2 Dim



Data set 2: 50 Dim

Applications for high dimensional clustering

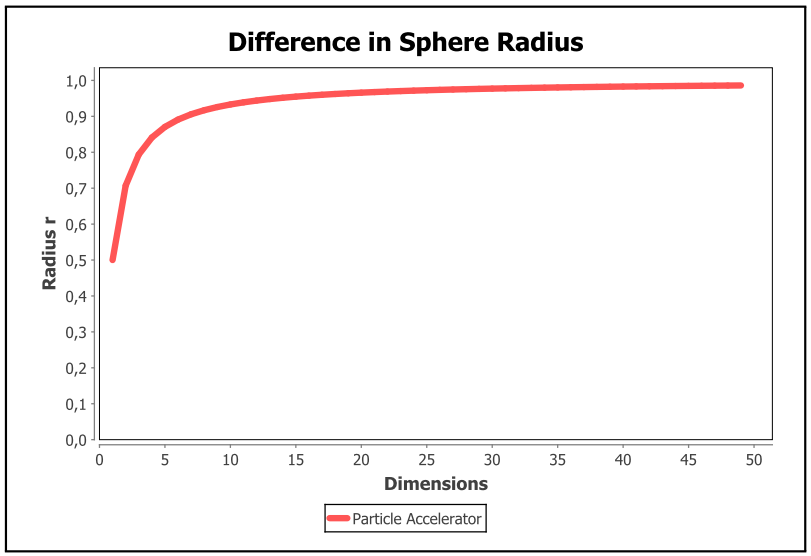
- Text mining
- Media clustering
  - Pictures
  - Music
  - Movies
- Image recognition
- Gene analysis
- Molecule clustering (Pharmacy)
- Physical data
  - Astronomy
  - Particle accelerator

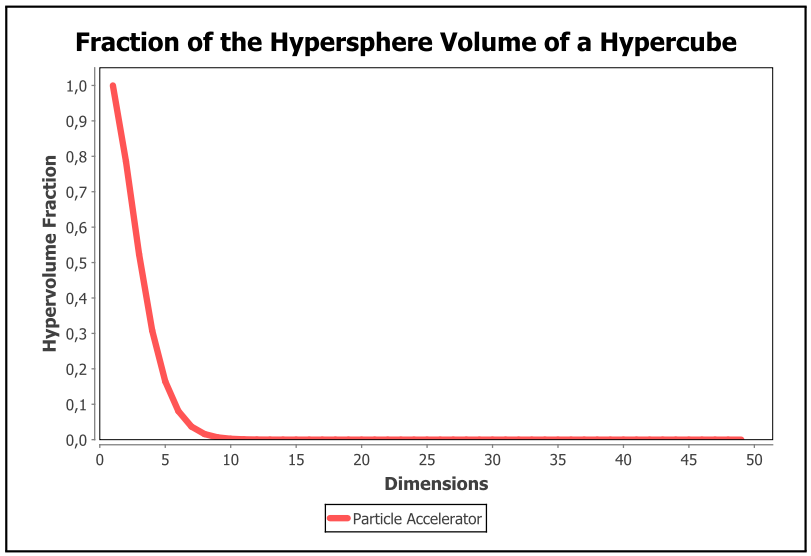
Challenges with high dim data sets in clustering

- Huge space that is very thin populated
- A meaningful scale between dimensions is difficult
- Not all recorded dimensions might be useful for clustering, some dimensions might only be locally useful
- Clusters might be in (affine) subspaces
- The fraction of data objects with missing values is sometimes large
- More outliers that are harder to recognize
- **Curse of Dimensionality: Distance concentration**

Hypervolume

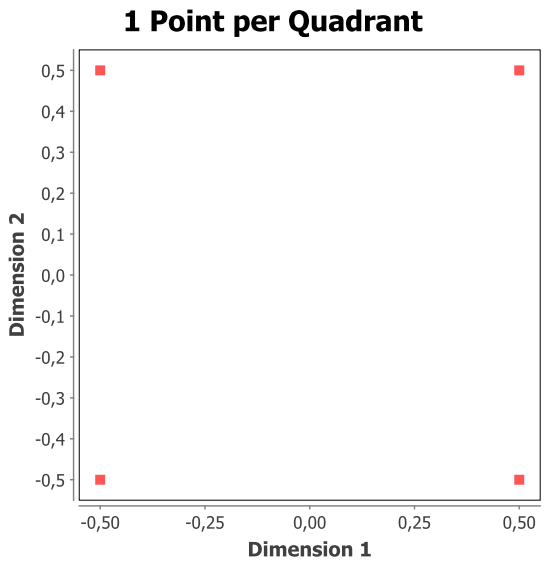
- First example: Let $S_1(1), S_2(r) \subset \mathbb{R}^n$ be hyperspheres, with $V_{S_2} = \frac{1}{2} \cdot V_{S_1}$ than $r = f(n) = \left(\frac{1}{2}\right)^{\frac{1}{n}}$.
- Second example: Let there be hypersphere $S(r) \subset \mathbb{R}^n$ and hypercube $C(2 \cdot r) \subset \mathbb{R}^n$. $g(n) = \frac{V_S}{V_C}$





Populate high dimensional space

- Put one Data object in each quadrant
- Exponentially (2^n) increasing number of data objects
- For 100 dimensions, that are $2^{100} \approx 1.3 \cdot 10^{30}$ data objects
- Datasets are usually much smaller than that



Distances in high dimensional space

- Distances increase with dimensionality because there are more values in which data objects can differ

  $$d(x, y)^2 = \sum_{i=1}^{d} (x_i - y_i)^2$$

- Clusters become indistinguishable if the point of view is not already inside a cluster (distance concentration)
- Cluster algorithms tend to fail due to this problem
- Because of the general increased distances, outliers are hard to distinguish from 'good' data objects

Distance concentration (in math)

Let F_m, $m = 1, 2, \ldots$ be a sequence of m-dimensional random variables and $S^{(m)} = \{x_1^{(m)}, \ldots, x_n^{(m)}\}$ be a sample of n independent data objects, distributed as F_m. Furthermore, let $\|\cdot\| : dom(F_m) \longrightarrow \mathbb{R}$ a metric, $p > 0$, $E(\|S^{(m)}\|^p)$ and $V(\|S^{(m)}\|^p)$ be finite, $E(\|S^{(m)}\|^p) > 0$ and n large enough so that $E(\|S^{(m)}\|^p) \in [dist_{\min}(S^{(m)})^p, dist_{\max}(S^{(m)})^p]$. Than

$$\lim_{m\to\infty} \frac{V(\|S^{(m)}\|^p)}{E(\|S^{(m)}\|^p)^2} = 0 \iff$$

$$\lim_{m\to\infty} P\left((1+\varepsilon) \cdot dist_{\min}(S^{(m)})^p > dist_{\max}(S^{(m)})^p\right) = 1, \forall \varepsilon > 0$$

Proof:

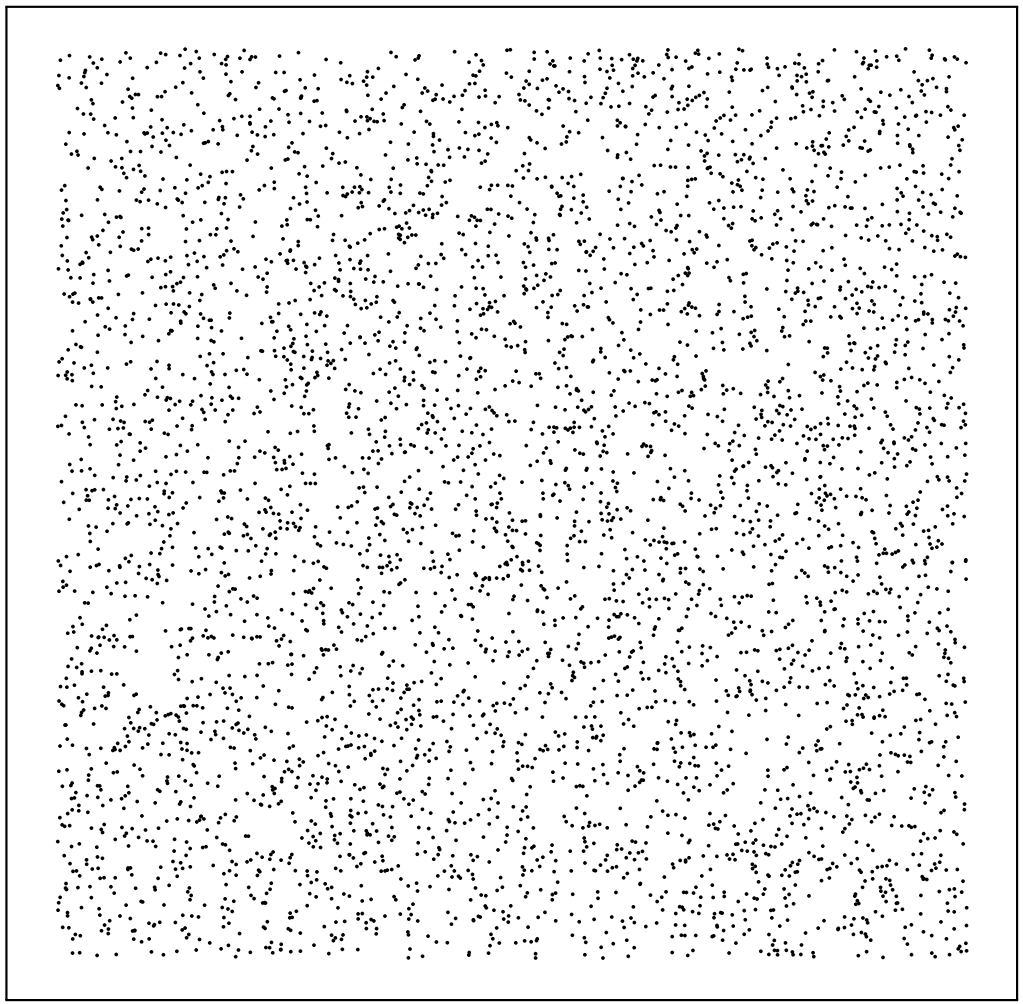
$(\Rightarrow)$ (Beyer et al., 1999)

$(\Leftarrow)$ (Durrant and Kabán, 2009)

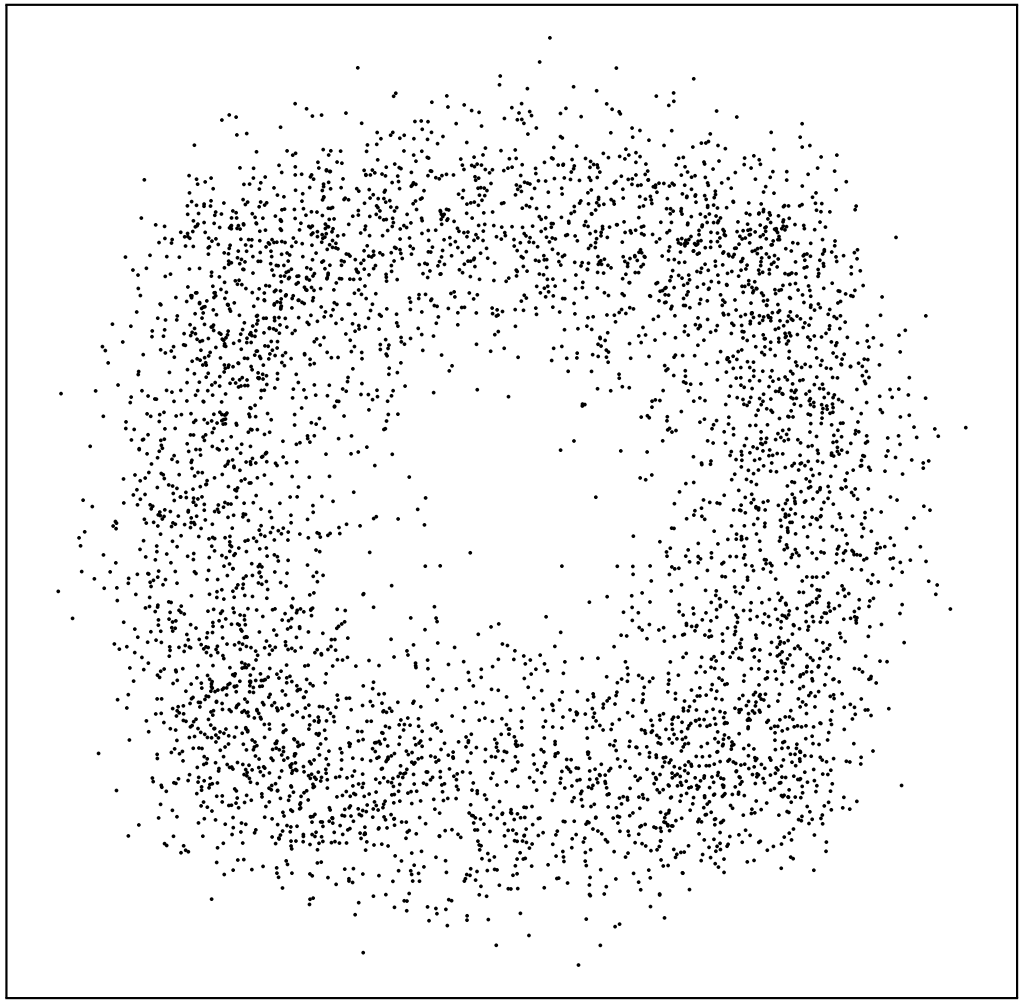
Distance concentration (in words)

- With increasing dimensionality distances become larger
- If the variance of data object location does not increase accordingly, distances to all data objects become identical
  - Independent on the point of view
  - Almost independent on the underlying data distribution (finite variance)
  - Almost independent on the sample size

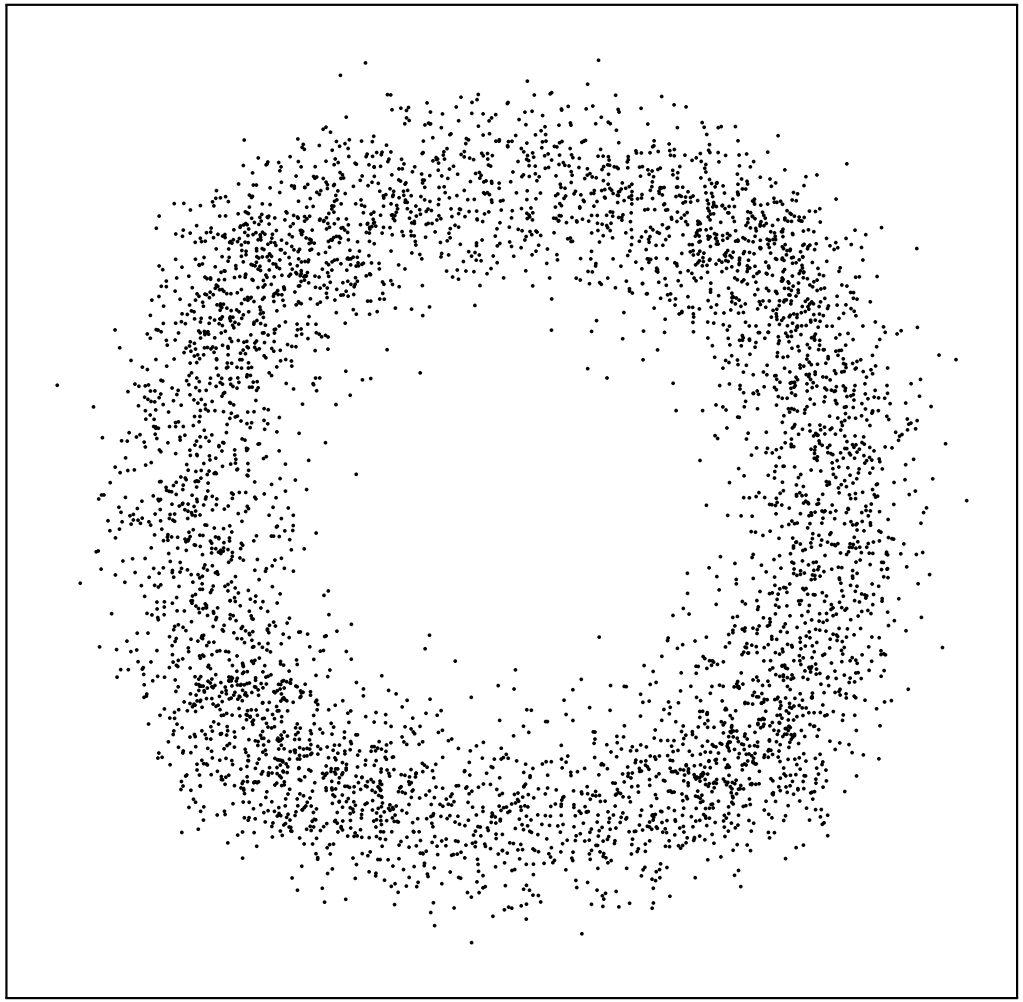
2 dimensions uniform distribution



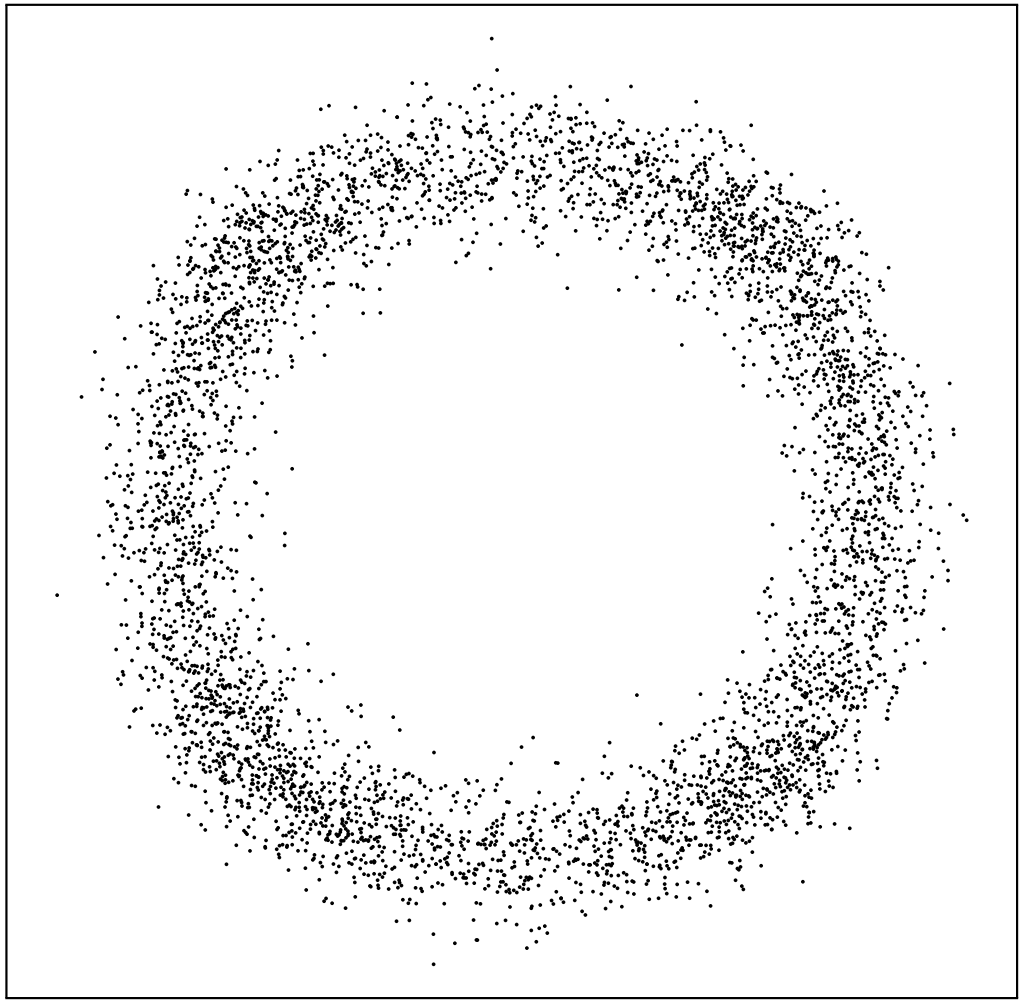
5 dimensions uniform distribution



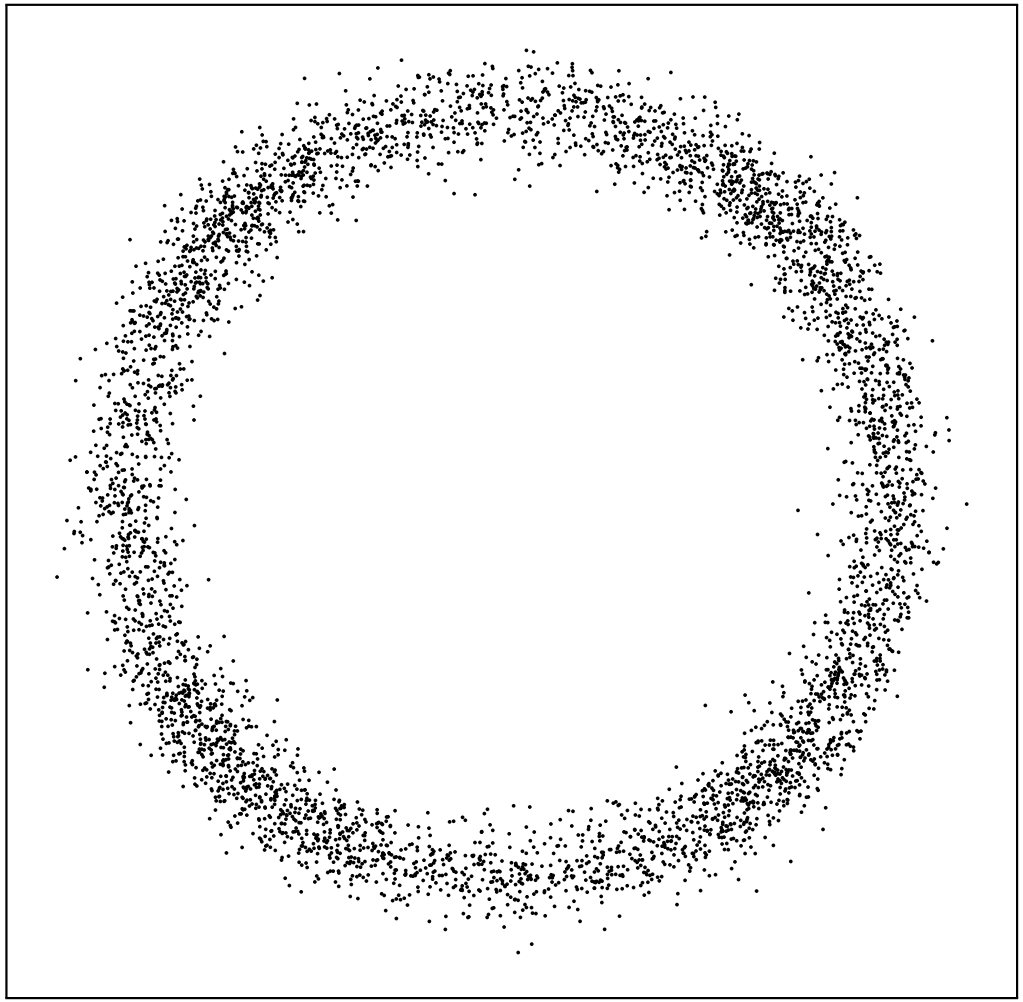
10 dimensions uniform distribution



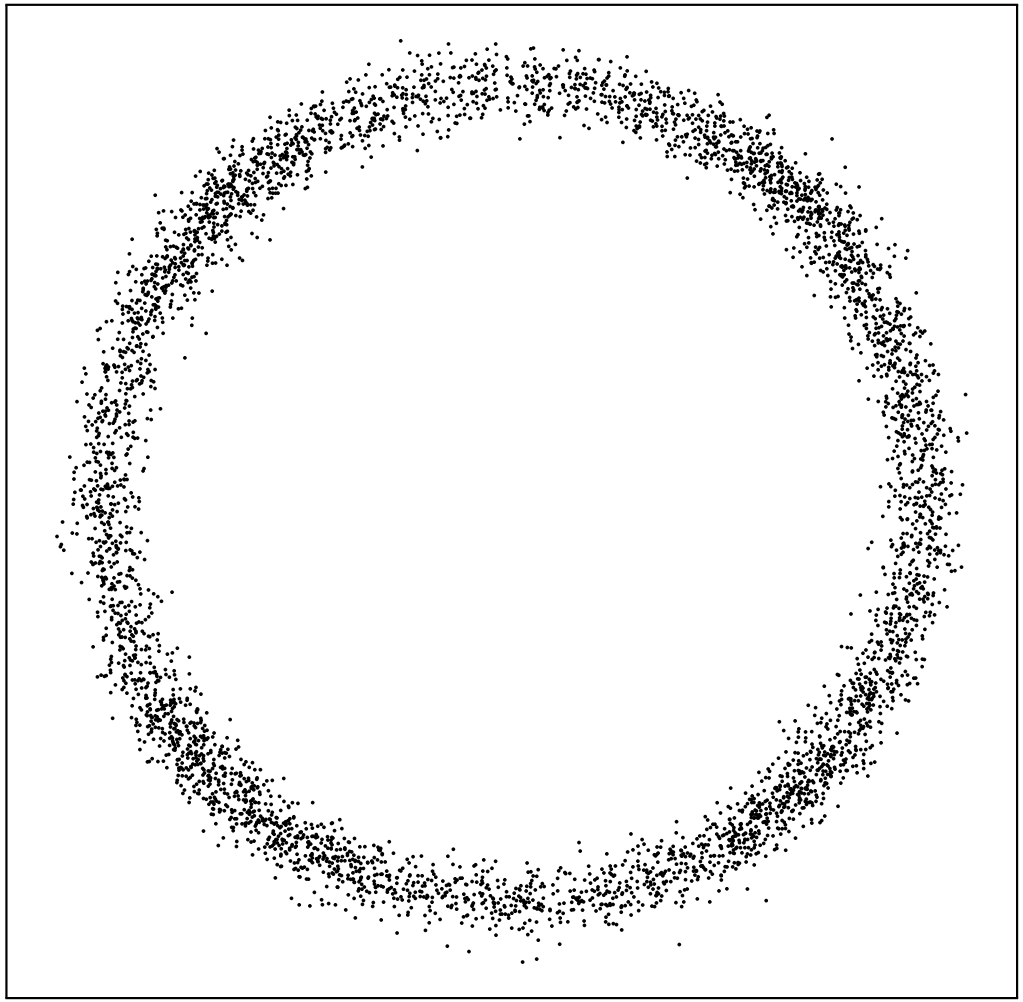
20 dimensions uniform distribution



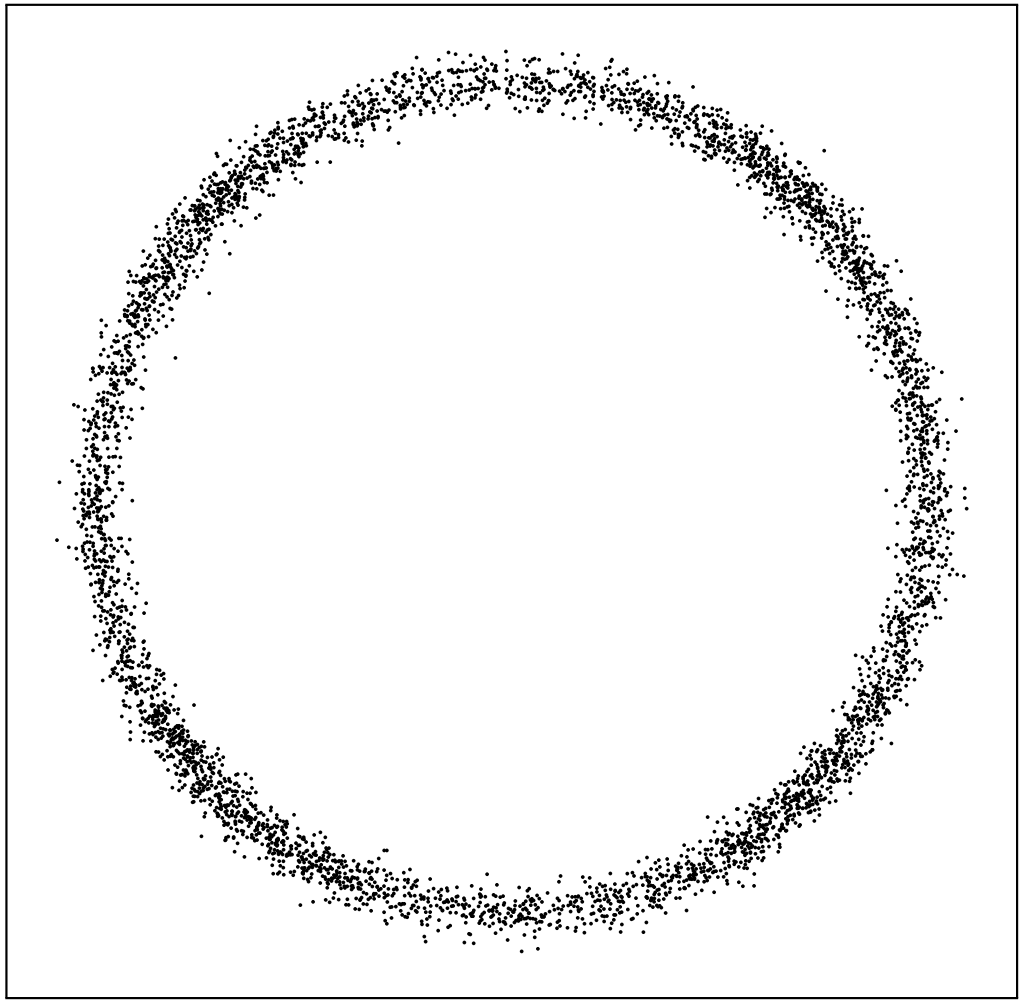
50 dimensions uniform distribution



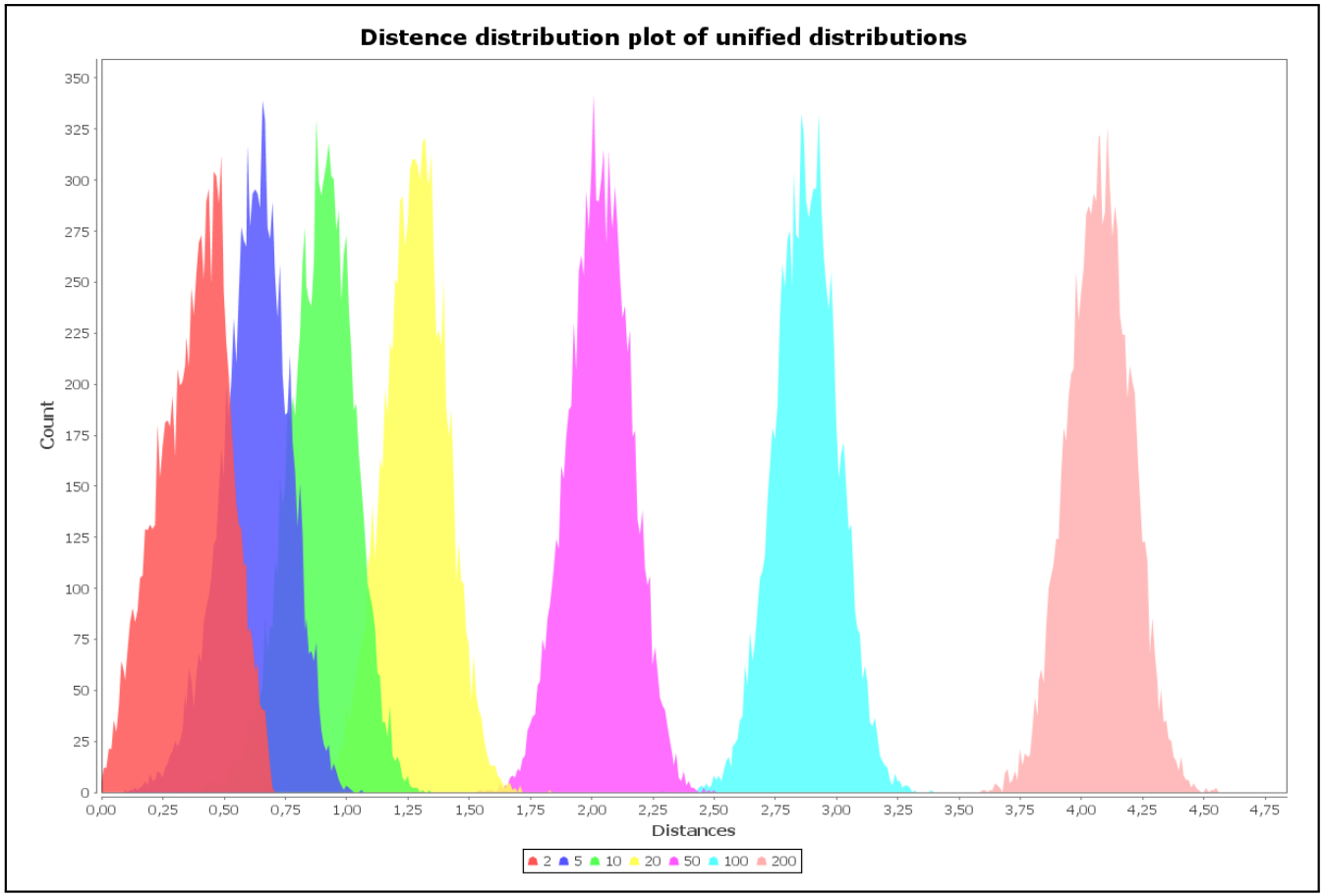
100 dimensions uniform distribution



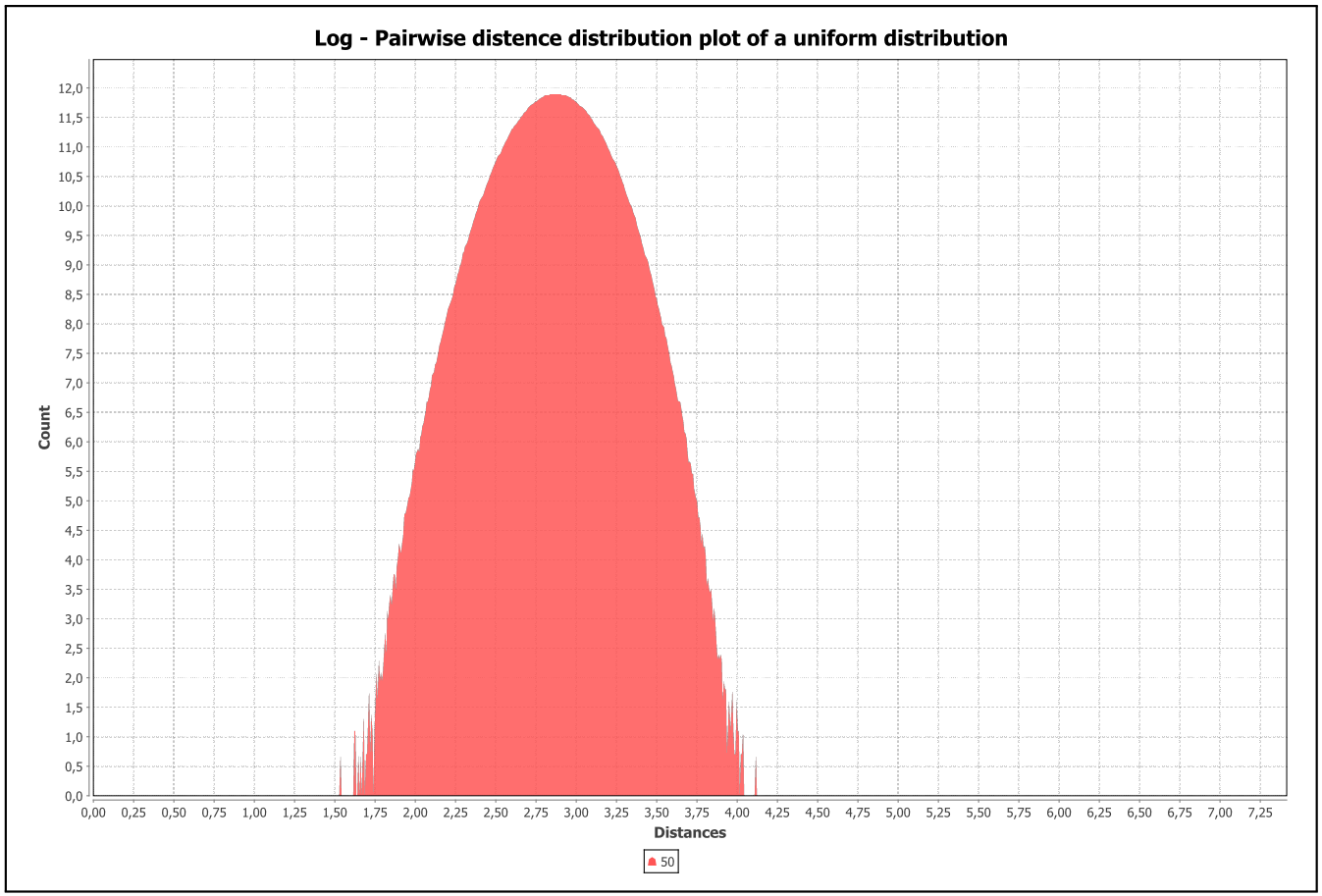
200 dimensions uniform distribution



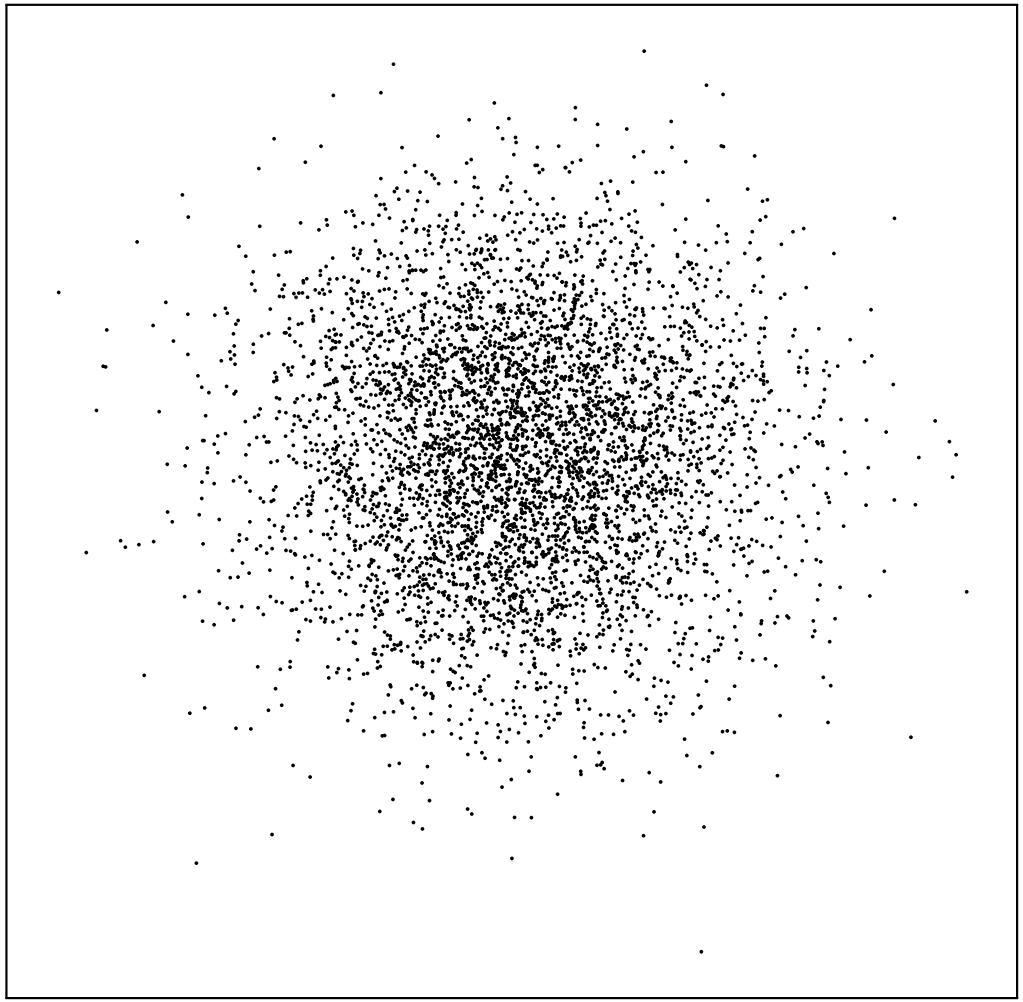
Multidimensional distance uniform distribution



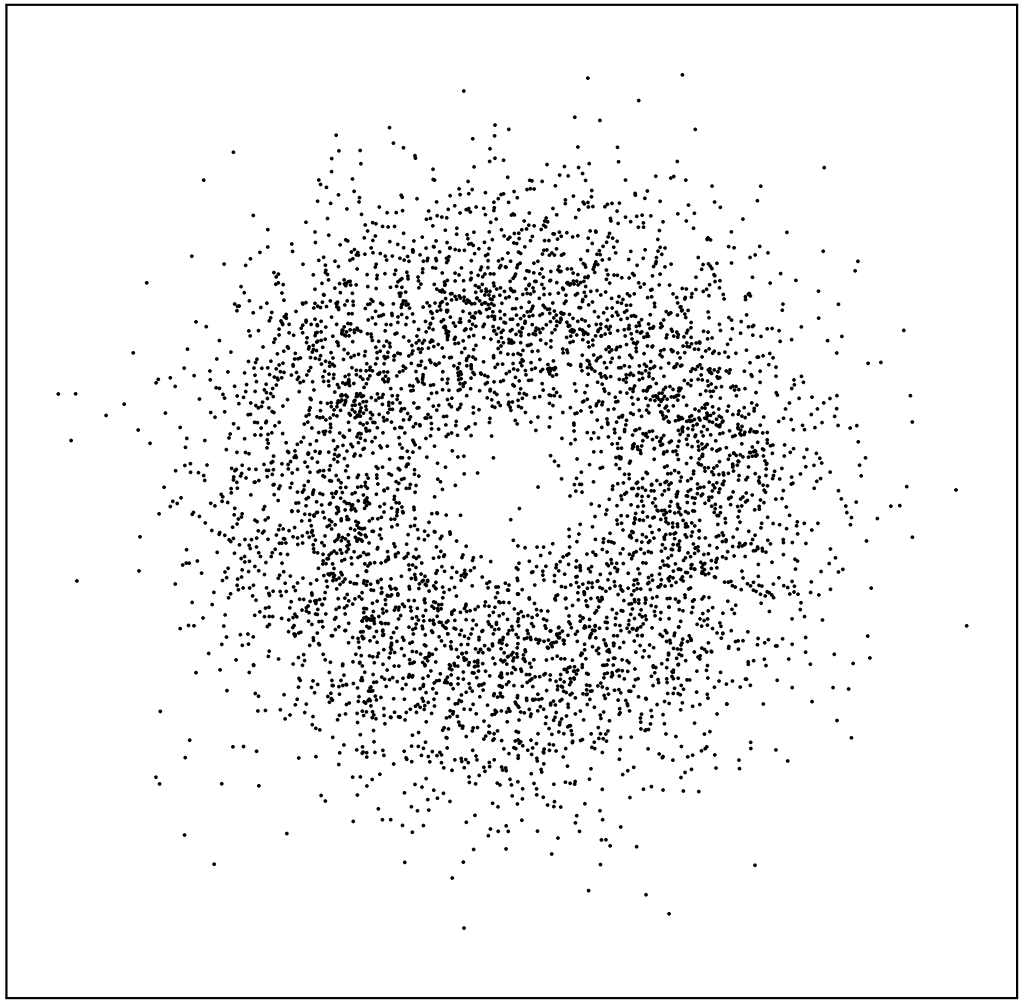
Pairwise distance diagram of a uniform distribution



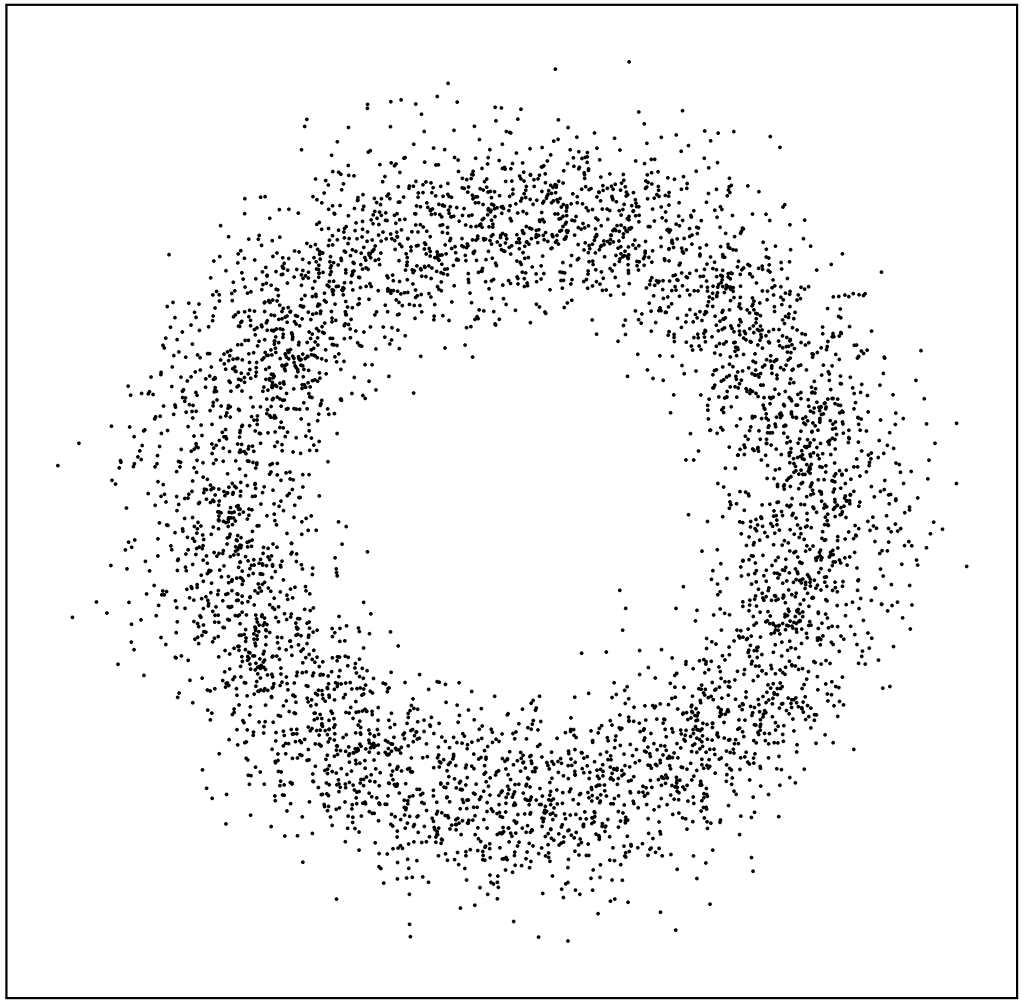
2 dimensions Gaussian distribution



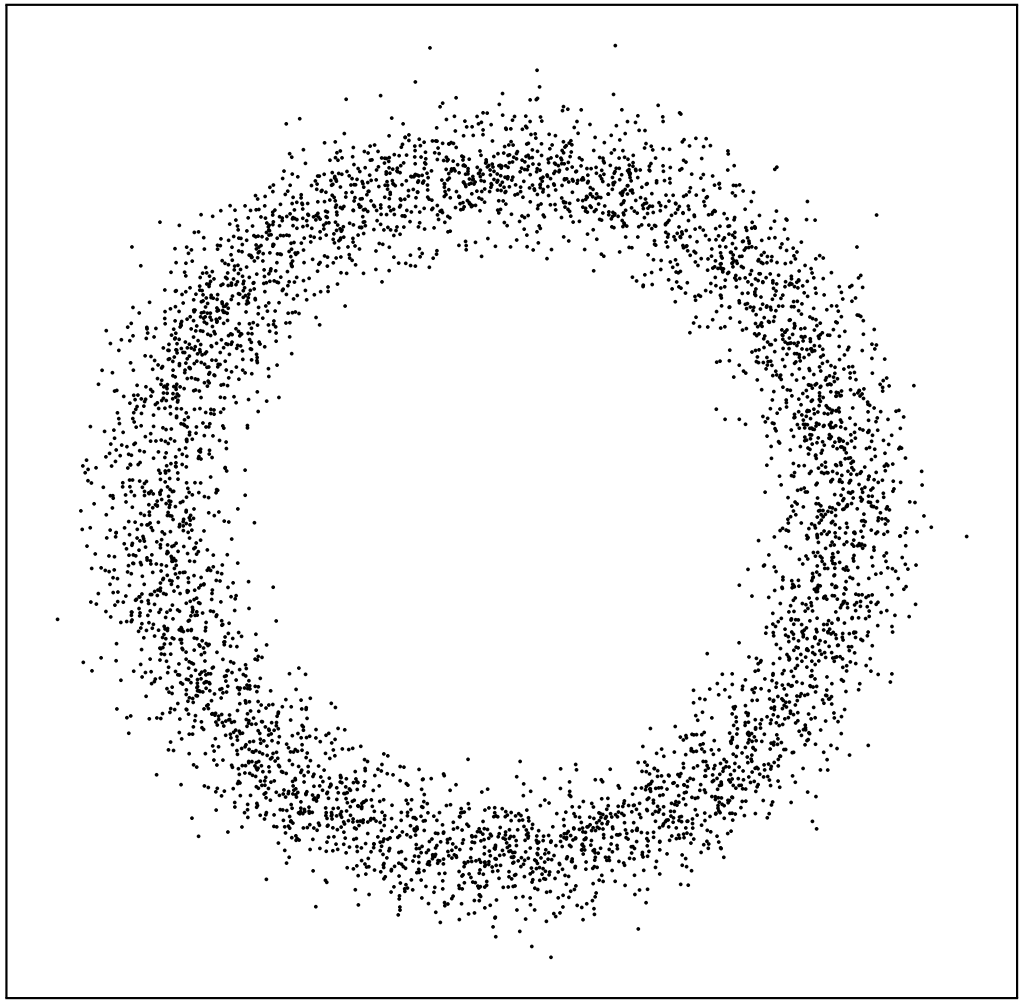
5 dimensions Gaussian distribution



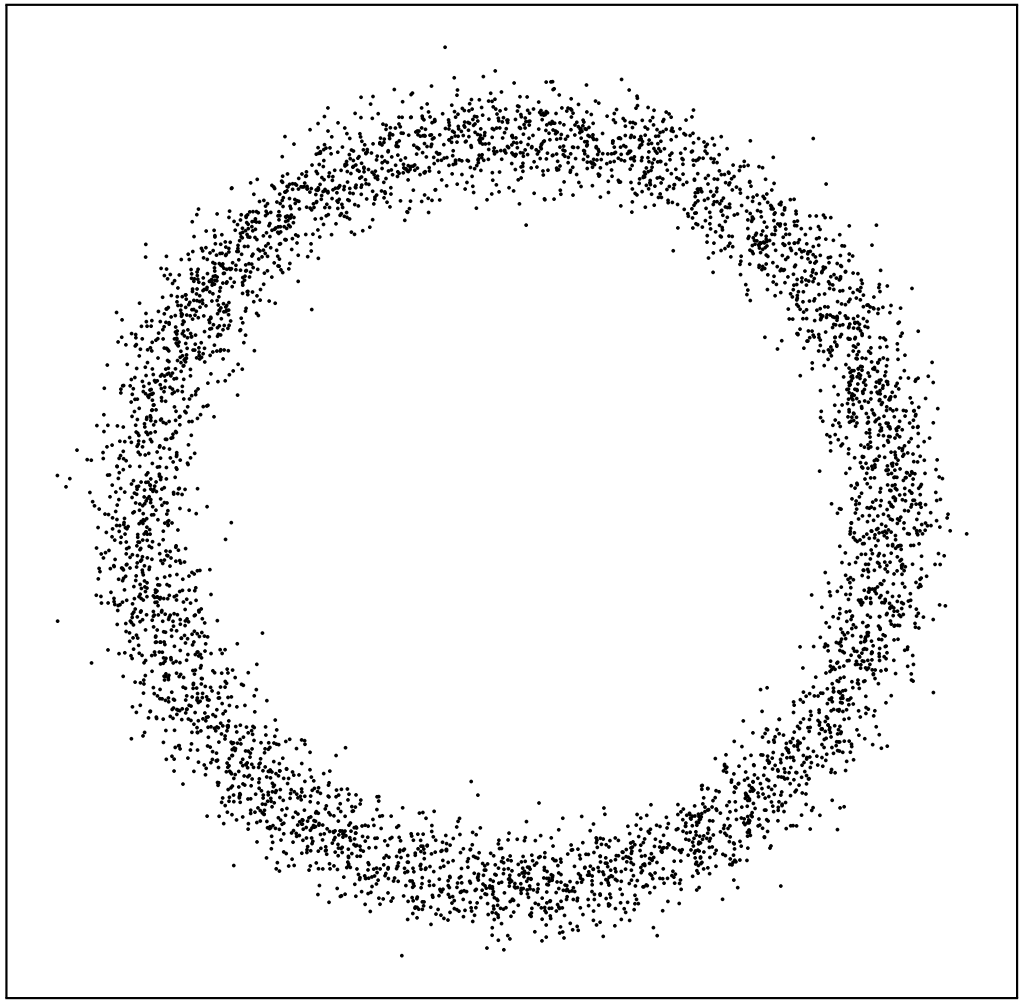
20 dimensions Gaussian distribution



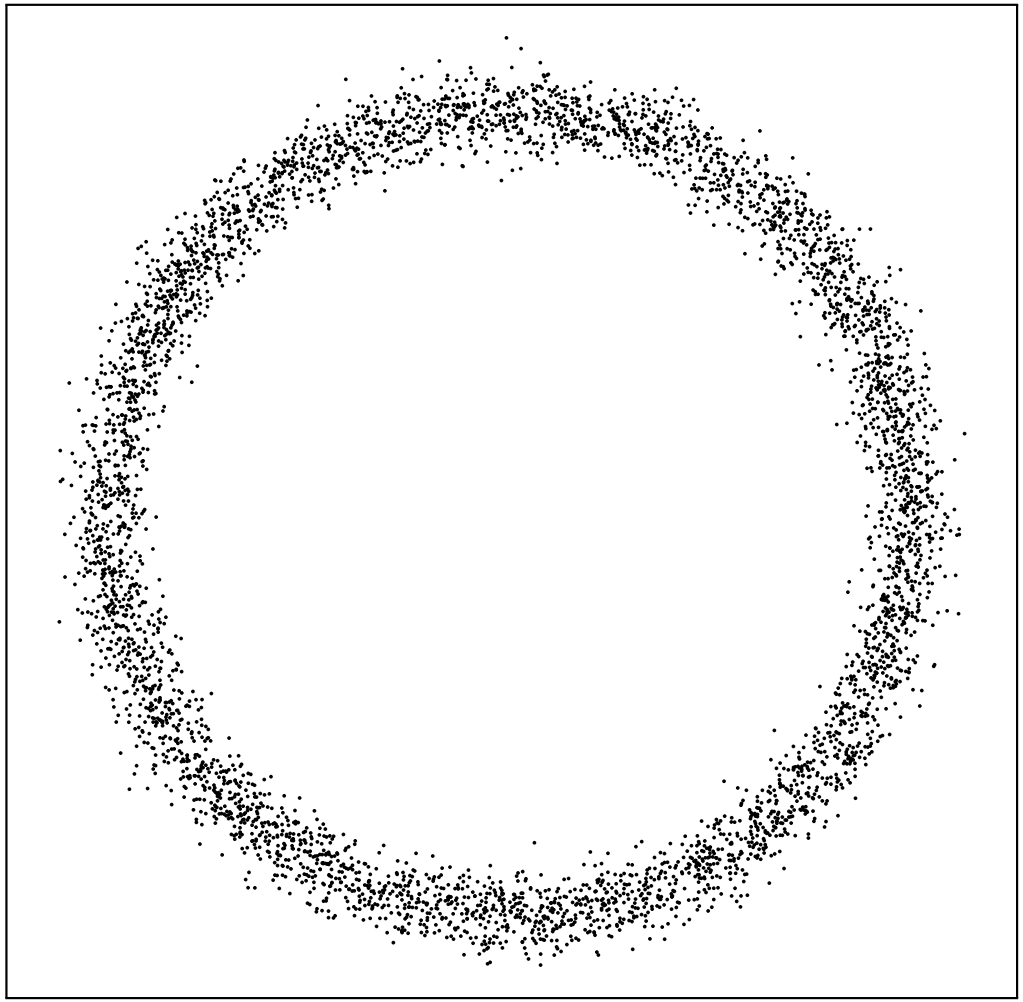
50 dimensions Gaussian distribution



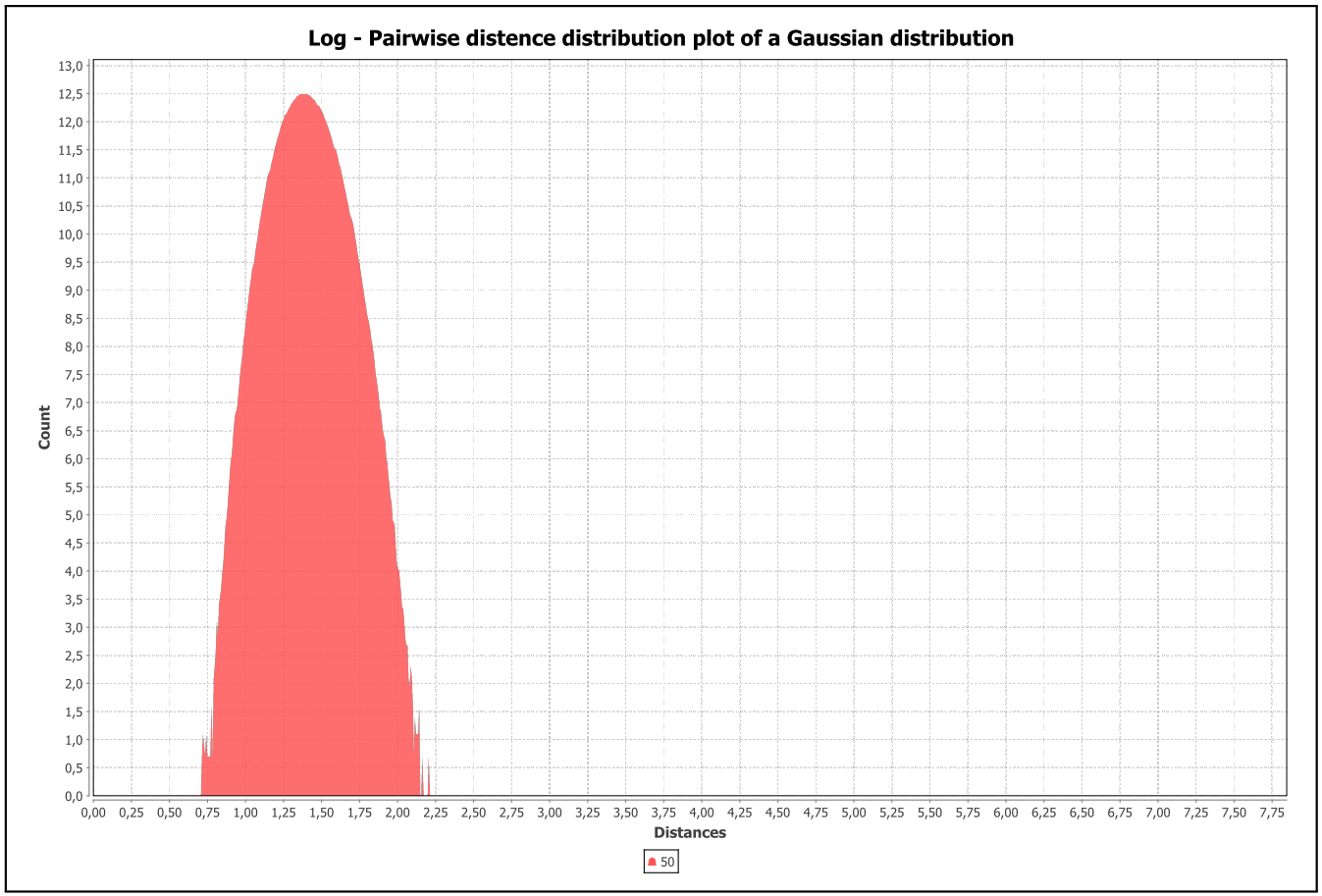
100 dimensions Gaussian distribution



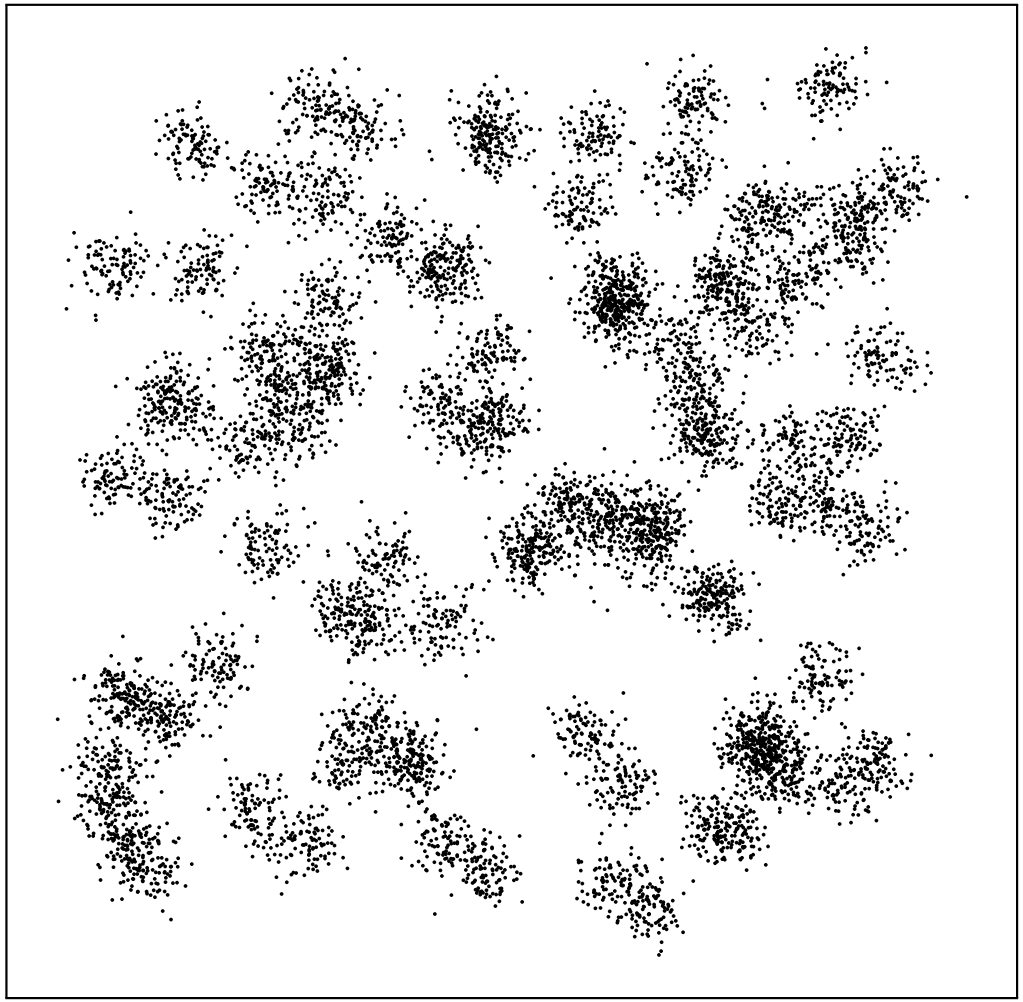
200 dimensions Gaussian distribution



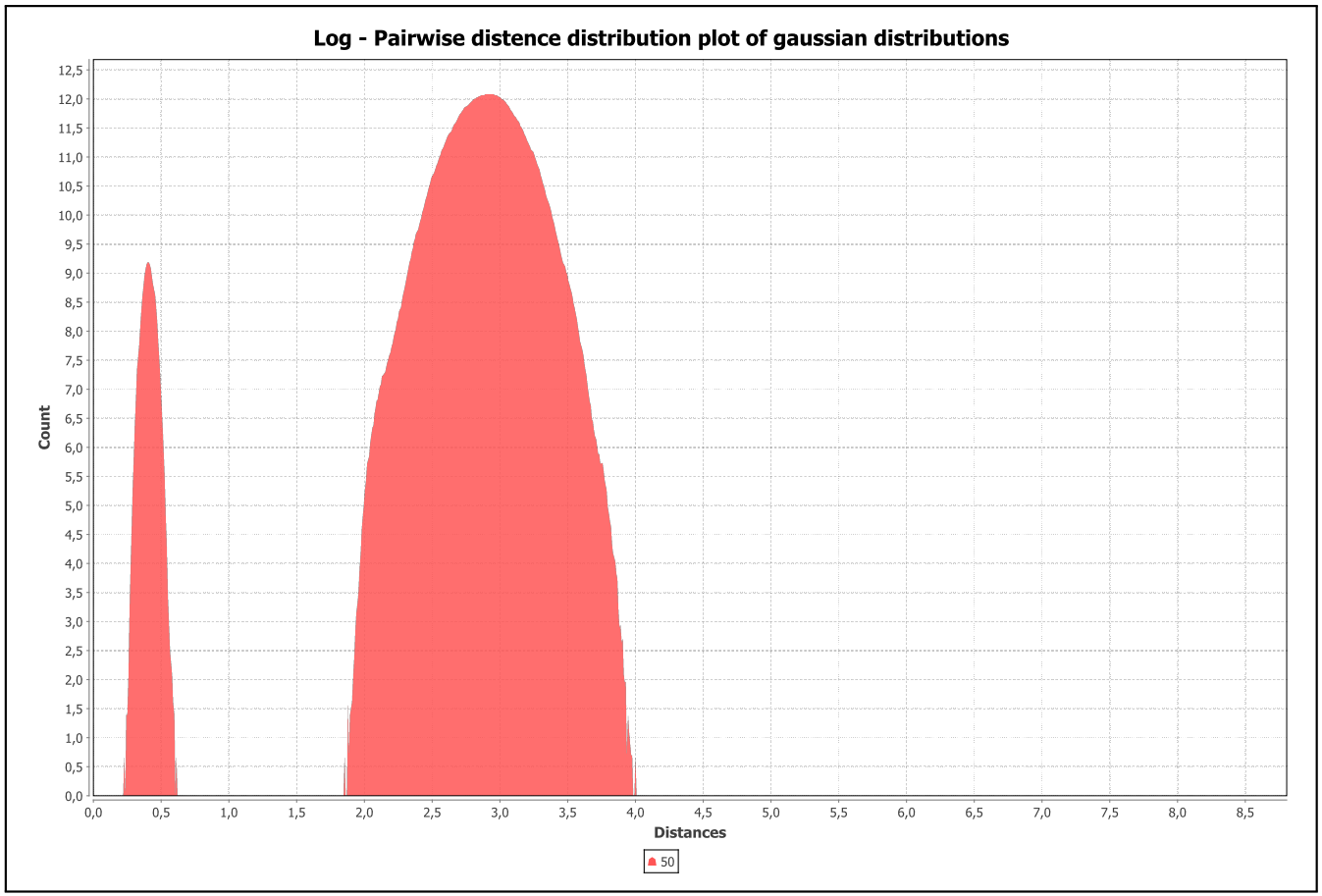
Pairwise distance diagram of a Gaussian distribution



100 cluster in a 50 dimensional data set



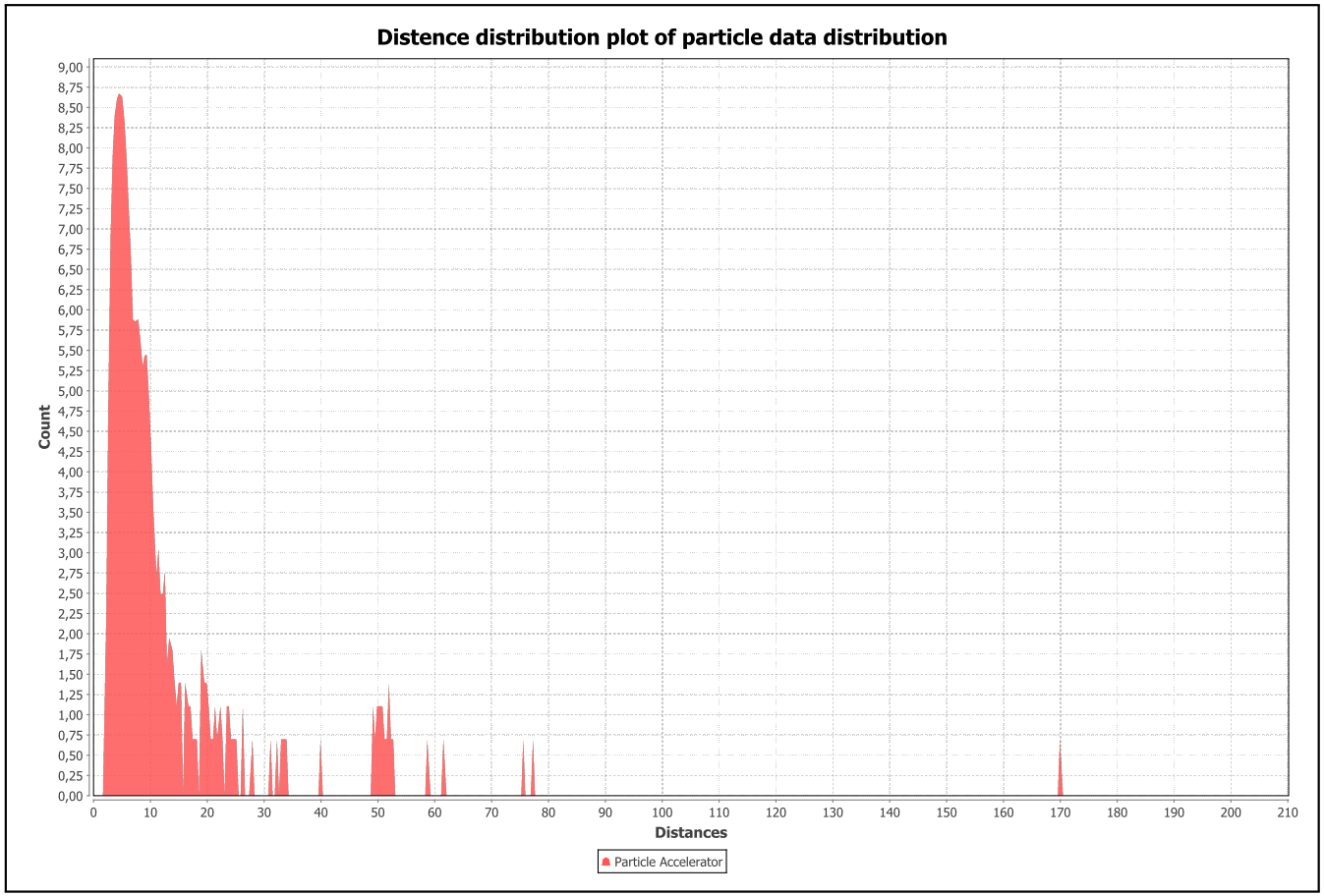
Pairwise distance diagram of uniform distributed Gaussian distributions



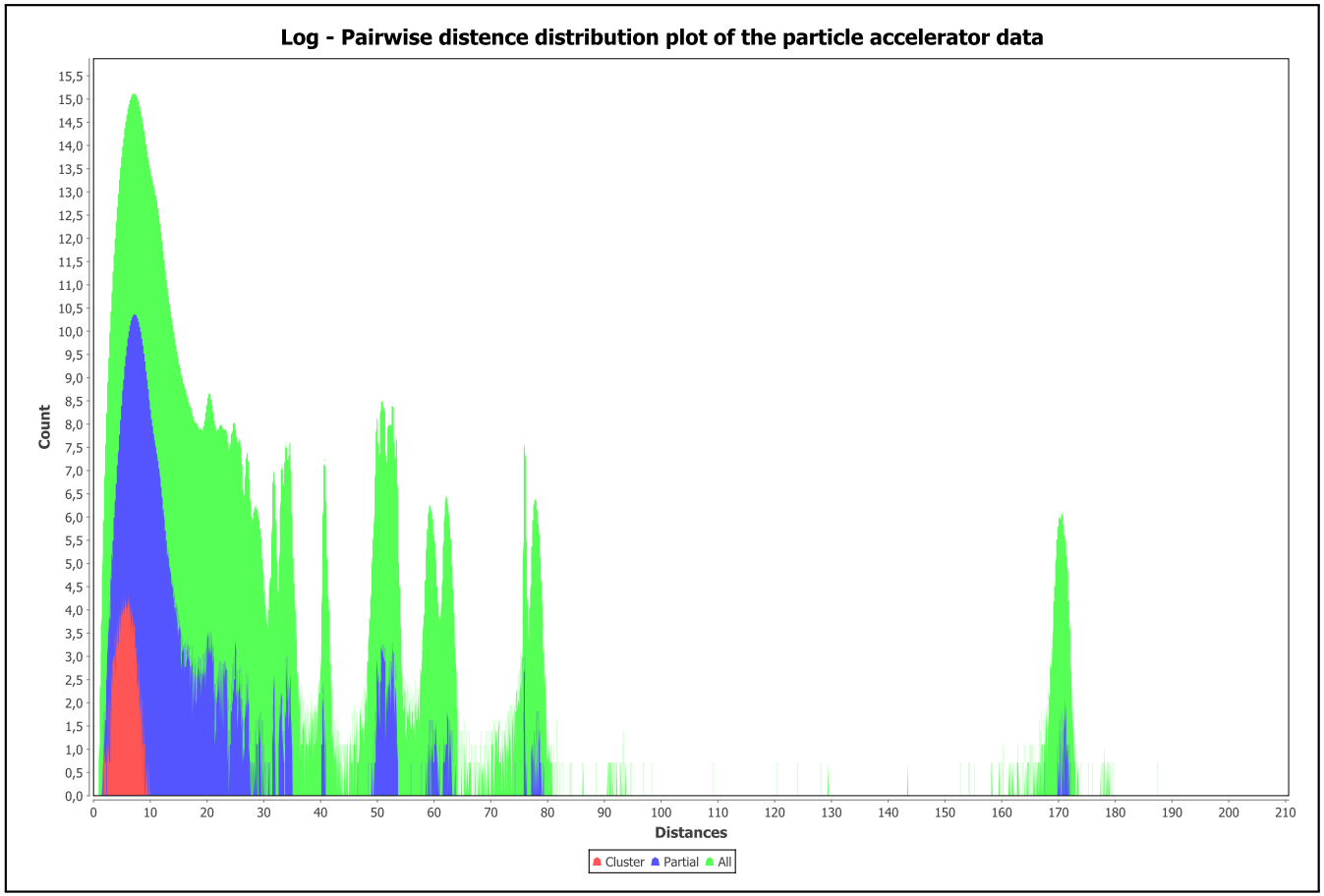
Example data set from particle accelerator

- Monte-Carlo simulation of a particle decay in a particle accelerator
- Data consists of roughly 85 parameter that can be reduced to 33 parameter
- Extremely unbalanced, 36000 objects per data set with around $100 - 150$ 'good' data objects and rest noise: (0.27% - 0.42% not noise)
- For analysis purposes, data of the 33 remaining dimensions is normalised to mean 0, and variance 1

Distance diagram of the PA-data set

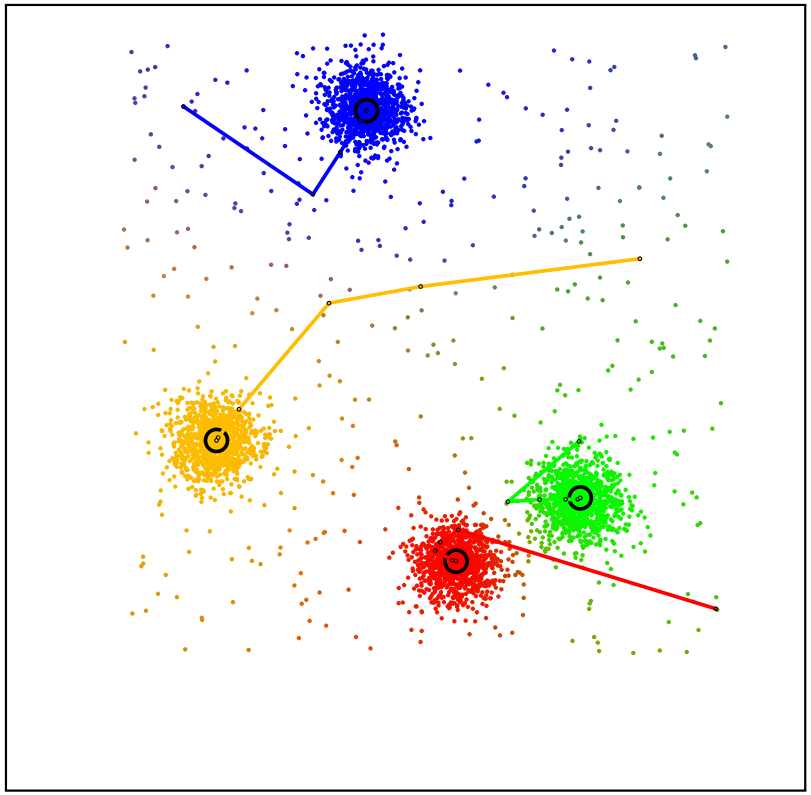


Pairwise Distance diagram of the PA-data set

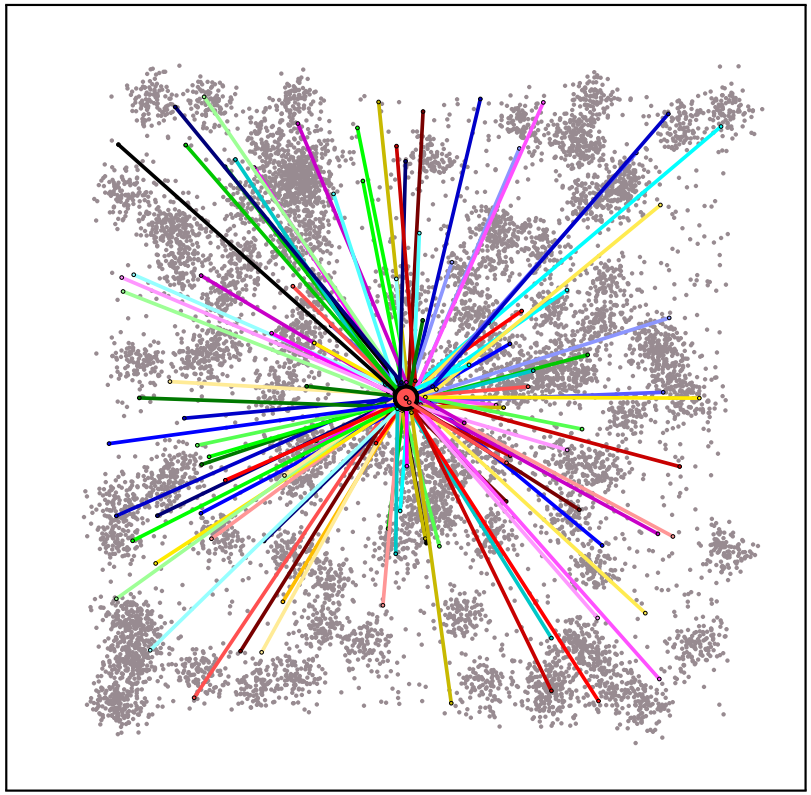


Current Section

Example: FCM



Data set 1: 2 Dim



Data set 2: 50 Dim

FCM

- Membership value update equation:

$$u_{ij}^{t+1} = \frac{\left(\frac{1}{d_{ij}^t}\right)^{\frac{2}{\omega-1}}}{\sum_{k=1}^{c}\left(\frac{1}{d_{kj}^t}\right)^{\frac{2}{\omega-1}}}$$

- Prototype update equation:

$$y_i^{t+1} = \frac{\sum_{j=1}^{n}\left(u_{ij}^t\right)^{\omega} x_j}{\sum_{j=1}^{n}\left(u_{ij}^t\right)^{\omega}}$$

FCM with distance concentration

- Suppose
  $P\left((1+\varepsilon)\cdot dist_{\min}(S^{(m)})^p > dist_{\max}(S^{(m)})^p\right) = 1$, than
  $P(d_ij^t - d^* < \varepsilon) = 1$
- Therefore $d_ij^t \approx d^*$
- Membership value update equation:

$$u_{ij}^{t+1} \approx \frac{\left(\frac{1}{d^*}\right)^{\frac{2}{\omega-1}}}{\sum_{k=1}^{c}\left(\frac{1}{d^*}\right)^{\frac{2}{\omega-1}}} = \frac{\left(\frac{1}{d^*}\right)^{\frac{2}{\omega-1}}}{c\cdot\left(\frac{1}{d^*}\right)^{\frac{2}{\omega-1}}} = \frac{1}{c}$$

FCM with distance concentration

- With $u_{ij}^t \approx \frac{1}{c}$

$$y_i^{t+1} \approx \frac{\sum_{j=1}^{n} \left(\frac{1}{c}\right)^{\omega} x_j}{\sum_{j=1}^{n} \left(\frac{1}{c}\right)^{\omega}} = \frac{\sum_{j=1}^{n} \left(\frac{1}{c}\right)^{\omega} x_j}{n \cdot \left(\frac{1}{c}\right)^{\omega}} = \frac{\sum_{j=1}^{n} x_j}{n}$$

- The new location of the prototype is approximately the centre of the data set

A collection of clustering algorithms

Clustering algorithm	Distance problem	Effect
Hierarchical	distance comparison	arbitrary result
Hard k-means	distance comparison	initialization problem
Fuzzy c-means	harmonic mean	broken
Density based	k-nearest neighbour	arbitrary result
EM	harmonic mean	broken (?)
LVQ	distance comparison	arbitrary result (?)
Fuzzy LVQ	harmonic mean	broken (?)
Kernal based	promising Algorithm (needs testing)	

Bibliography

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999).
When is nearest neighbor meaningful?
In *Database Theory - ICDT'99*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235. Springer Berlin / Heidelberg.

Durrant, R. J. and Kabán, A. (2009).
When is 'nearest neighbour' meaningful: A converse theorem and implications.
Journal of Complexity, 25(4):385 – 397.