Manifolds, a new framework for warped curves analysis

Séminaires de l'unité BIA (INRA/MIA Toulouse)

Elie Maza

INP-ENSAT – INRA (laboratoire GBF)

9 décembre 2011

Plan

1. Microarray data analysis
2. Issue
3. Some solutions
    - Shift model
    - Phase model
    - Quantile normalization
4. New framework
    - Idea
    - Approximation of δ
    - Shift model
    - Application

Plan

1. Microarray data analysis
2. Issue
3. Some solutions
   - Shift model
   - Phase model
   - Quantile normalization
4. New framework
   - Idea
   - Approximation of δ
   - Shift model
   - Application

Central Dogma of Molecular Biology (F. Crick, 1970, *Nature*)

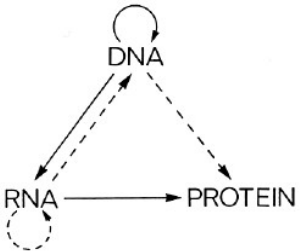
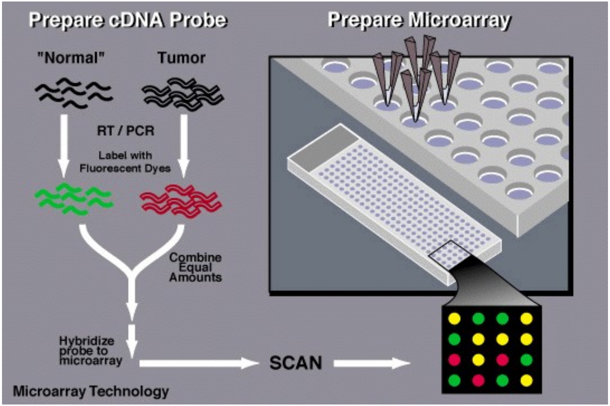


Fig. 3. A tentative classification for the present day. Solid arrows show general transfers; dotted arrows show special transfers. Again, the absent arrows are the undetected transfers specified by the central dogma.

⇝ The study of the genetic information contained in any organism:

1. Finding coding sequences in the DNA.
2. Measuring the abundance of RNAs.
3. Studing the diversity of Proteins.

DNA Microarray : A "new" technology for transcriptome studies.



National Human Genome Research Institute, `http://www.genome.gov/`

The ratio of red and green fluorescence intensities for each spot is intended to be indicative of the relative abundance of the transcripts in the test condition compared to the reference condition:

$$M = \log_2\left(\frac{R}{G}\right) = \log_2(R) - \log_2(G)$$

$$A = \frac{1}{2}\left(\log_2(R) + \log_2(G)\right) = \log_2\left(\sqrt{RG}\right)$$

Example

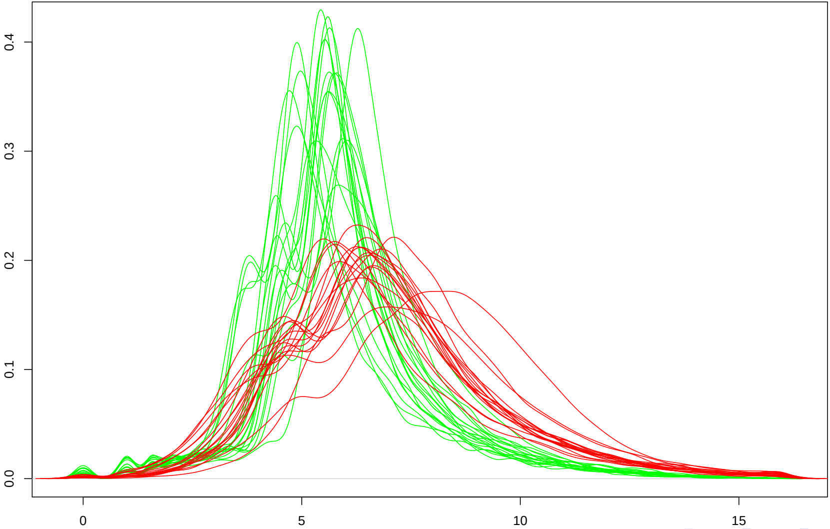
- Genomic and Biotechnology of the Fruit laboratory (GBF)
- tomato microarrays
- Data:
  - 18 conditions
  - 13056×2 measures per microarray


z−range 5.7 to 6.6 (saturation 5.7, 6.6)
z−range 5.1 to 6.1 (saturation 5.1, 6.1)
z−range 5.7 to 6.6 (saturation 5.7, 6.6)
z−range 5.1 to 6.3 (saturation 5.1, 6.3)
z−range 5.6 to 6.6 (saturation 5.6, 6.6)
z−range 5.1 to 6 (saturation 5.1, 6)
z−range 6.2 to 7.4 (saturation 6.2, 7.4)
z−range 5.5 to 6.9 (saturation 5.5, 6.9)
z−range 5.6 to 8.1 (saturation 5.6, 8.1)
z−range 5.2 to 6.5 (saturation 5.2, 6.5)
z−range 5.7 to 7.2 (saturation 5.7, 7.2)
z−range 5.1 to 5.8 (saturation 5.1, 5.8)
z−range 5.1 to 8.4 (saturation 5.1, 8.4)
z−range 4.7 to 5.3 (saturation 4.7, 5.3)
z−range 5.2 to 6.4 (saturation 5.2, 6.4)
z−range 4.9 to 5.7 (saturation 4.9, 5.7)
z−range 5.6 to 6.4 (saturation 5.6, 6.4)
z−range 5 to 5.7 (saturation 5, 5.7)
z−range 5.8 to 8.4 (saturation 5.8, 8.4)
z−range 5.1 to 7 (saturation 5.1, 7)
z−range 5.8 to 7.7 (saturation 5.8, 7.7)
z−range 5.1 to 7.9 (saturation 5.1, 7.9)
z−range 6 to 8 (saturation 6, 8)
z−range 5.2 to 6.5 (saturation 5.2, 6.5)
z−range 5.2 to 7 (saturation 5.2, 7)
z−range 4.8 to 5.7 (saturation 4.8, 5.7)
z−range 5.3 to 7.1 (saturation 5.3, 7.1)
z−range 5 to 5.9 (saturation 5, 5.9)
z−range 5.2 to 6.5 (saturation 5.2, 6.5)
z−range 4.9 to 7.6 (saturation 4.9, 7.6)
z−range 5.6 to 8.1 (saturation 5.6, 8.1)
z−range 5.2 to 9.6 (saturation 5.2, 9.6)
z−range 5.9 to 8 (saturation 5.9, 8)
z−range 5.2 to 6.1 (saturation 5.2, 6.1)
z−range 5.8 to 6.8 (saturation 5.8, 6.8)
z−range 5.2 to 6 (saturation 5.2, 6)

High variability of microarray data:

- RNA extraction
- hybridization conditions (temperature, humidity, ...)
- image acquisition
- heat and light sensitivities for Cy3 (Green) and Cy5 (Red)
- ...

$\rightsquigarrow$ Normalization procedures!

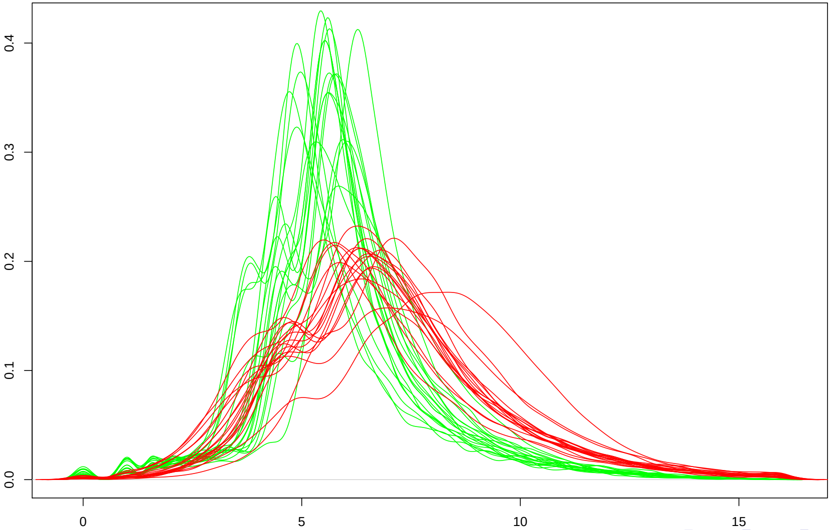

0.0
0.1
0.2
0.3
0.4
0
5
10
15

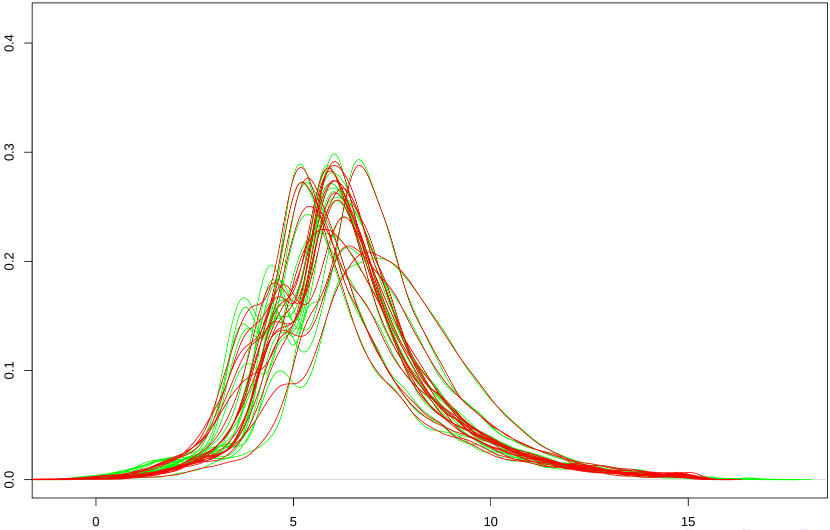
Two "standard" normalization procedures:

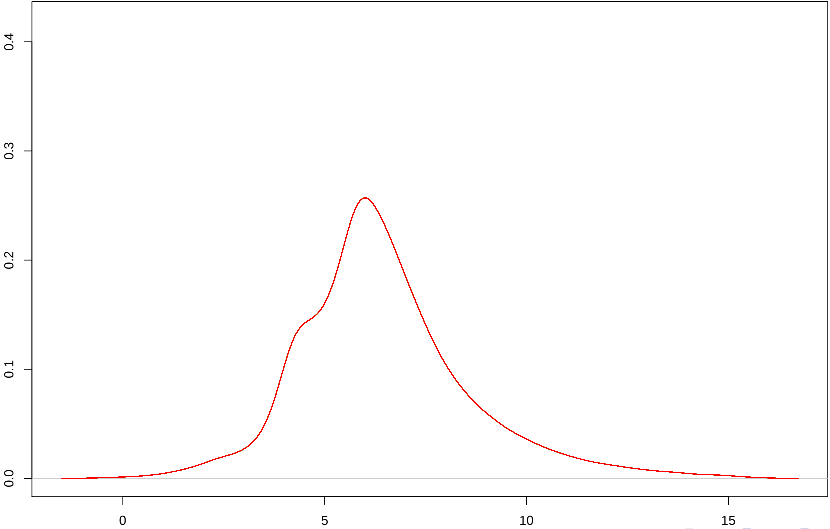
- within arrays (print-tip loess)
- between arrays (quantile normalization)

Software package `limma`:

- Smyth, G. K. (2005). Limma: linear models for microarray data. In: 'Bioinformatics and Computational Biology Solutions using R and Bioconductor'. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York, pages 397–420.


0.0
0.1
0.2
0.3
0.4
0
5
10
15


0.0
0.1
0.2
0.3
0.4
0
5
10
15


0.0
0.1
0.2
0.3
0.4
0
5
10
15

Plan

1. Microarray data analysis
2. Issue
3. Some solutions
   - Shift model
   - Phase model
   - Quantile normalization
4. New framework
   - Idea
   - Approximation of δ
   - Shift model
   - Application

Warping model:

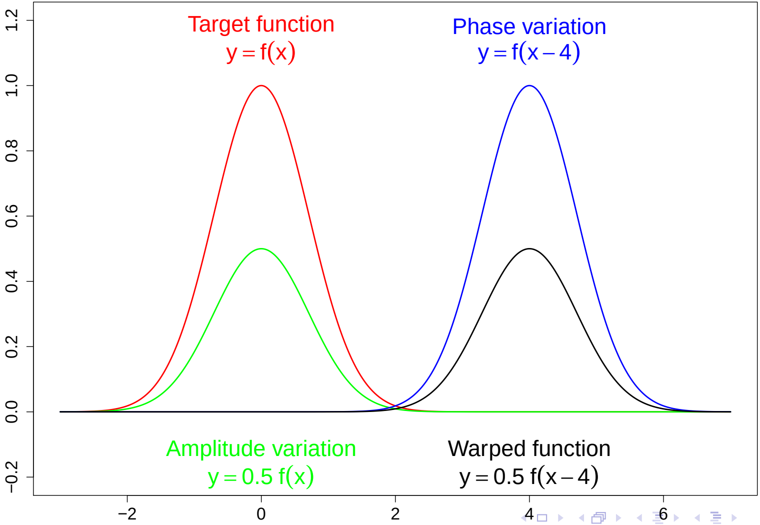
$$G = F \circ H^{-1}$$

where

- F is a stochastic function such as $\mathbb{E}(F) = f$ (amplitude variation)
- H is a strictly increasing stochastic function such as $\mathbb{E}(H) = \phi$ (phase variation)

The structural expectation:

$$\boxed{f_{\mathrm{ES}} \stackrel{\mathrm{def}}{=} f \circ \phi^{-1}}$$

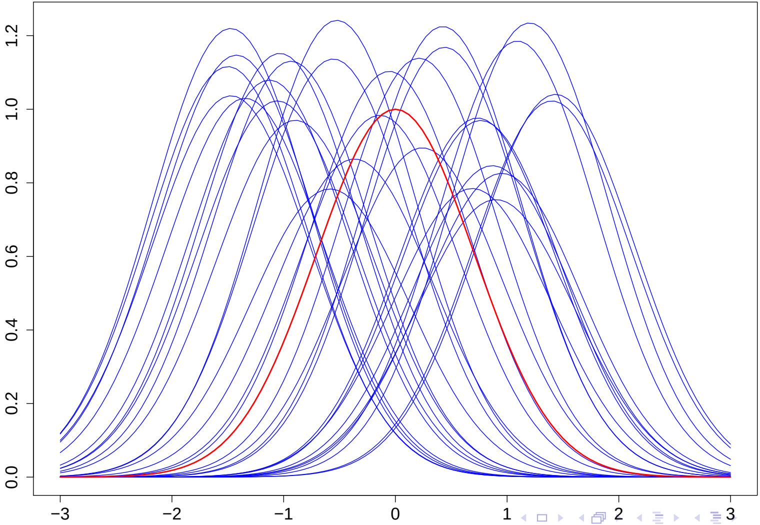

Target function
$y = f(x)$
Phase variation
$y = f(x - 4)$
Amplitude variation
$y = 0.5\,f(x)$
Warped function
$y = 0.5\,f(x - 4)$
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−2
0
2
4
6

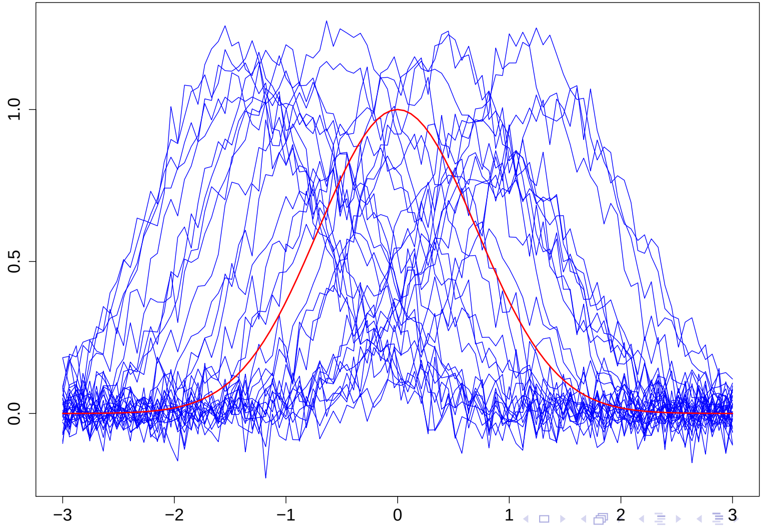
Data:

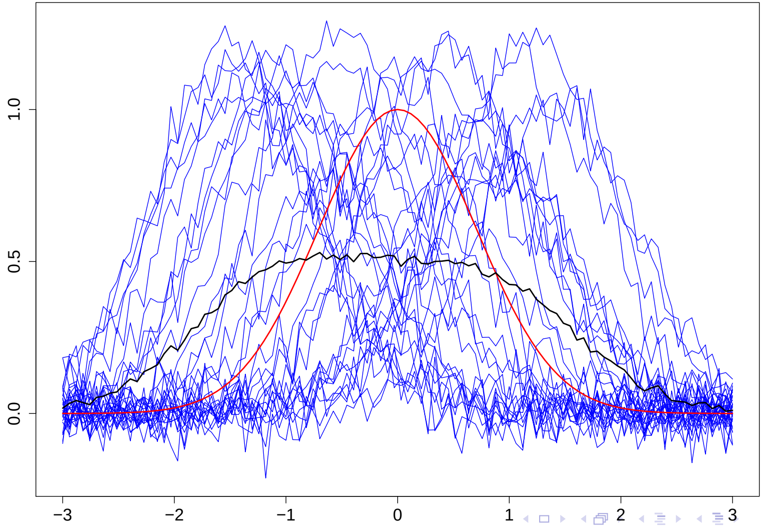
$$Y_{ij} = F_i \circ H_i^{-1}(t_{ij}) + \epsilon_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, m$$

where

- $F_i \sim F$ are *iid*
- $H_i \sim H$ are *iid*
- $t_{ij} \in [a, b] \subset \mathbb{R}$
- ϵ_{ij} are *iid* with mean 0 and variance σ^2


1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3


0.0
0.5
1.0
−3
−2
−1
0
1
2
3


1.0
0.5
0.0
−3
−2
−1
0
1
2
3

Solution:

1. Alignement of curves.
2. Estimation of mean of aligned curves.

Plan

Plan

1. Microarray data analysis
2. Issue
3. Some solutions
   - Shift model
   - Phase model
   - Quantile normalization
4. New framework
   - Idea
   - Approximation of δ
   - Shift model
   - Application

Model:

$$Y_{ij} = f\left(t_j - \theta_i^*\right) + \epsilon_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, m,$$

with

- $f : \mathbb{R} \longrightarrow \mathbb{R}$ an unknown T-periodic function
- $\theta = \left(\theta_i^*\right)_{i=1,\ldots,n}$ an unknown shift parameter
- $\forall j = 1, \ldots, m,\ t_j = \frac{j-1}{m} T \subset [0, T[$
- $\forall i = 1, \ldots, n,\ \left(\epsilon_{ij}\right)_{j=1,\ldots,m}$ are *iid* and $\mathcal{N}(0, 1)$

Other model:

$$d_{il} = e^{-il\alpha_i^*} c_l(f) + w_{il}, \; i = 1, \dots, n, \; l = -(m-1)/2, \dots, (m-1)/2,$$

with

- $(c_l(f))_{l \in \mathbb{Z}}$ the Fourier coefficients of f:

$$\forall l \in \mathbb{Z}, \; c_l(f) = \frac{1}{T} \int_0^T f(t) e^{-i2\pi \frac{tl}{T}} \mathrm{d}t$$

- $\alpha^* = (\alpha_i^*)_{i=1,\dots,n} = \left(\frac{2\pi}{T}\theta_i^*\right)_{i=1,\dots,n}$ the normalized shift parameter
- $\forall i \in \{1, \dots, n\}, \; (w_{il})_{l=-(m-1)/2,\dots,(m-1)/2}$ are complex *iid* variables with mean 0 and variance $\frac{1}{n}$

Contraste:

$$M_n(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=-\frac{m-1}{2}}^{l=\frac{m-1}{2}} \delta_l^2 \left| \tilde{c}_{il}(\alpha) - \hat{c}_l(\alpha) \right|^2$$

with

- $\tilde{c}_{il}(\alpha) = e^{il\alpha_i} d_{il}$ are the scaled Fourier coefficients
- $\hat{c}_l(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \tilde{c}_{il}(\alpha)$ are the mean coefficients
- $(\delta_l)_{l\in\mathbb{Z}}$ is such that $\sum_{l\in\mathbb{Z}} \delta_l^2 < +\infty$

Remark: $\tilde{c}_{il}(\alpha^*) = c_l(f) + e^{il\alpha_i^*} w_{il}$ and $\hat{c}_l(\alpha^*) = c_l(f) + \frac{1}{n} \sum_{i=1}^{n} e^{il\alpha_i^*} w_{il}$

Semi-parametric Estimation of Shifts. F. Gamboa, J-M. Loubes and E. Maza. *Electronic Journal of Statistics*, Vol. 1, 2007, 616–640.

Plan

1. Microarray data analysis
2. Issue
3. Some solutions
   - Shift model
   - Phase model
   - Quantile normalization
4. New framework
   - Idea
   - Approximation of δ
   - Shift model
   - Application

Model:

$$Y_{ij} = f \circ H_i^{-1}(t_j),\ i = 1, \ldots, n,\ j = 1, \ldots, m,$$

with

- $f : [a, b] \rightarrow \mathbb{R}$ a continuous function
- $\forall j = 1, \ldots, m,\ t_j = a + (j-1)\frac{b-a}{m-1} \subset [a, b]$
- $H_i \sim H$ are *iid*, continuous and strictly increasing

Idea: if f is strictly increasing, then

$$\left(f \circ H_i^{-1}\right)^{-1} = H_i \circ f^{-1}$$

hence

$$\mathbb{E}\left(\left(f \circ H_i^{-1}\right)^{-1}\right) = \mathbb{E}\left(H_i\right) \circ f^{-1} = \phi \circ f^{-1} = f_{\mathrm{ES}}^{-1}$$

Plan

Bolstad et al. (2003)

- Let $X \sim F$ a random variable.
- Let $X^1, \ldots, X^i, \ldots, X^m$ *iid* such that $F_i = F$.
- Let, for all $i \in \{1, \ldots, m\}$, $X_1^i, \ldots, X_j^i, \ldots, X_n^i$ *iid* such that $X_j^i \sim X^i$.
- Let, for all $i \in \{1, \ldots, m\}$, $X_{(1)}^i, \ldots, X_{(j)}^i, \ldots, X_{(n)}^i$ the order statistics.

The quantile normalization is defined, for all $j \in \{1, \ldots, n\}$, by

$$\hat{X}_{(j)} = \frac{1}{m} \sum_{i=1}^{m} X_{(j)}^i$$

Remark

For all $i \in \{1, \ldots, m\}$, assuming

$$X^i = H_i(X)$$

we have

$$F_i = F \circ H_i^{-1}$$

Remark: Bolstad et al. (2003) assume that $H = Id$.

Proof

1. **Non parametric estimation of the structural expectation of a stochastic increasing function.** J.-F. Dupuy, J.-M. Loubes and E. Maza, *Statistics and Computing*, 2011.

$$\hat{F}^{-1} = \frac{1}{m}\sum_{i=1}^{m} \hat{F}_i^{-1} \xrightarrow[m \to +\infty]{\mathbf{P}} F^{-1}$$

2. **Statistical properties of the quantile normalization method for DNA microarray analysis.** S. Gallón, J.-M. Loubes and E. Maza, preprint.

$$\hat{X}_{(j)} = \frac{1}{m}\sum_{i=1}^{m} X_{(j)}^{i} \xrightarrow[m \to +\infty]{\mathbf{P}} X_{(j)}$$

Example: for $i \in \{1, 2\}$, let

$$f_i = P_{1i} \times \mathcal{N}\left(M_{1i}, S_{1i}^2\right) + P_{2i} \times \mathcal{N}\left(M_{2i}, S_{2i}^2\right)$$

with $M_{1i} \sim \mathcal{U}[-1.8, -1.2]$, $M_{2i} \sim \mathcal{U}[1.2, 1.8]$, $n = 1000$ and

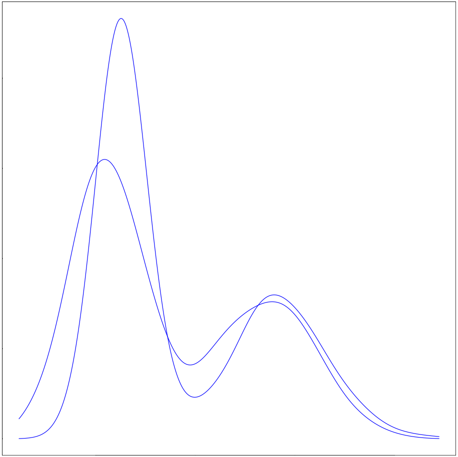
Simulation 1:

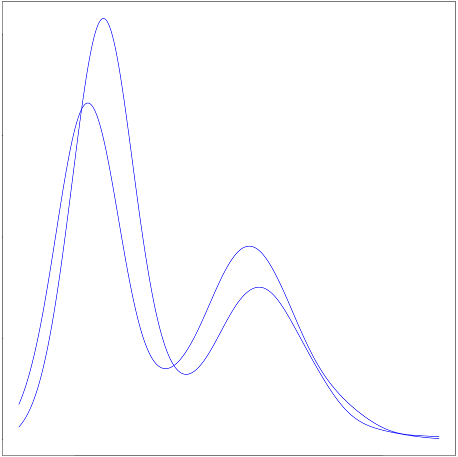
- $P_{11} = P_{12} = 0.6$
- $P_{21} = P_{22} = 0.4$
- $S_{1i} \sim \mathcal{U}[0.2, 0.8]$
- $S_{2i} \sim \mathcal{U}[0.7, 1.3]$

Simulation 2:

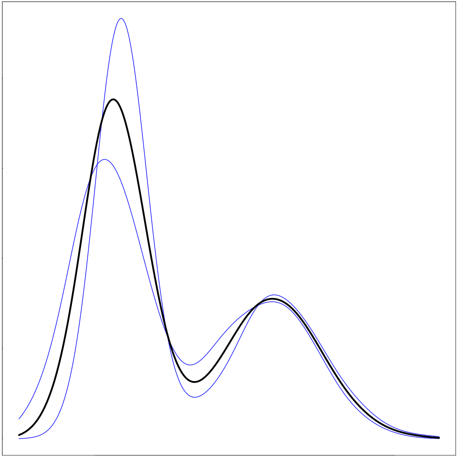
- $P_{1i} \sim \mathcal{U}[0.45, 0.75]$
- $P_{2i} = 1 - P_{1i}$
- $S_{11} = S_{12} = 0.5$
- $S_{21} = S_{22} = 1$

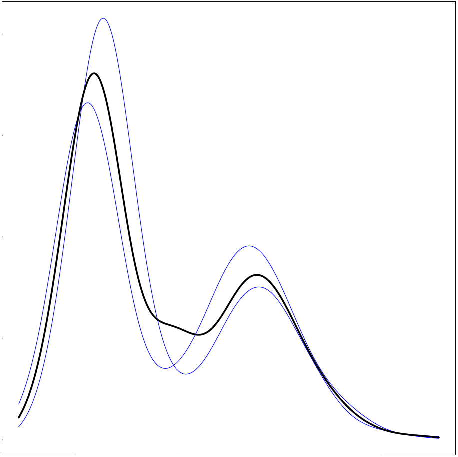
Quantile normalization



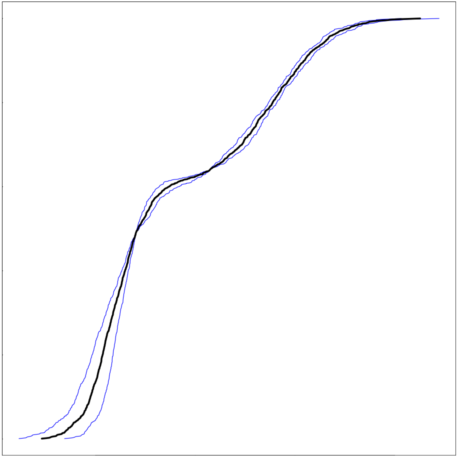


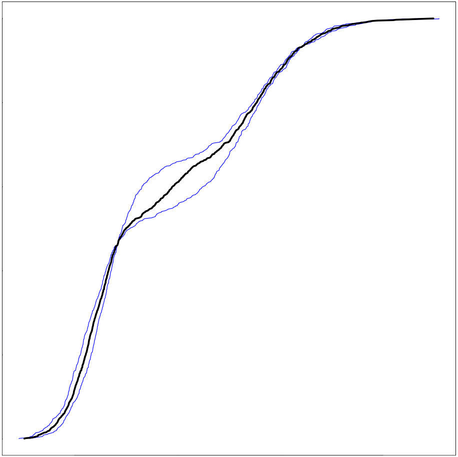
Quantile normalization





Quantile normalization





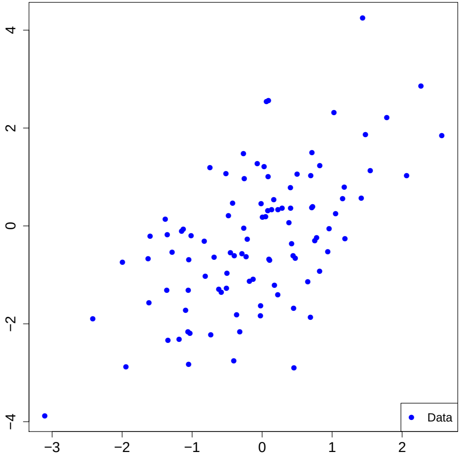
Limites

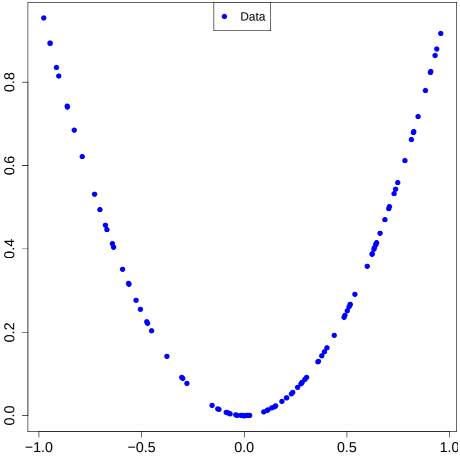
- Simulation 1:
$$F_i \approx F \circ H_i^{-1}$$
- Simulation 2:
$$f_i \approx f \circ H_i^{-1}$$

Plan

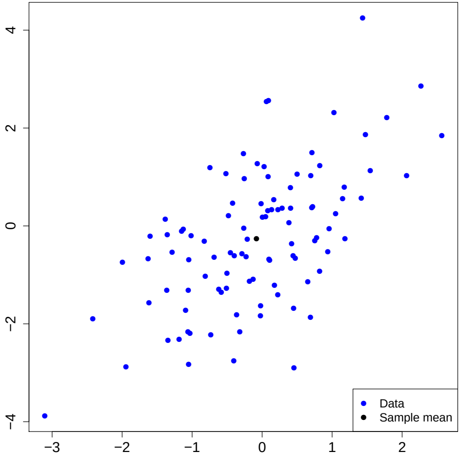
Plan

Idea

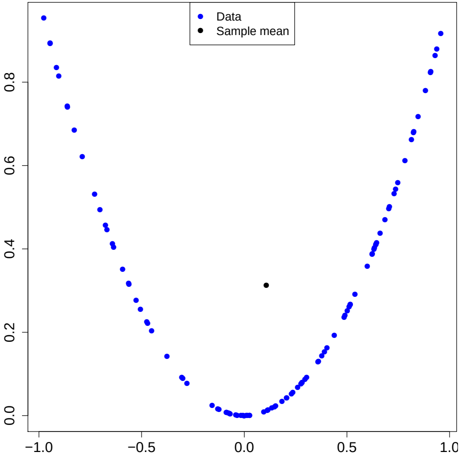




Idea

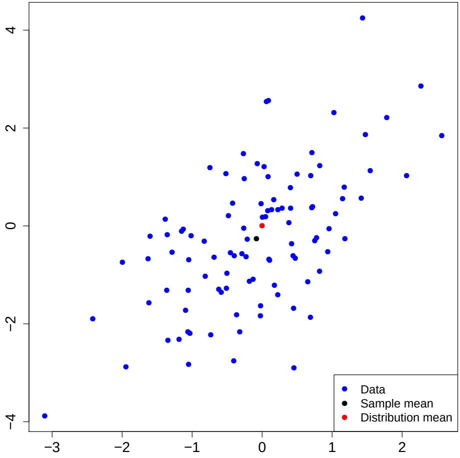


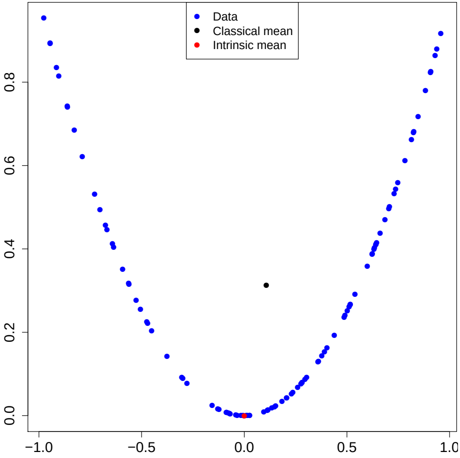
Data
Sample mean
4
2
0
−2
−4
−3
−2
−1
0
1
2



Data
Sample mean
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0

Idea


Data
Sample mean
Distribution mean
4
2
0
−2
−4
−3
−2
−1
0
1
2


Data
Classical mean
Intrinsic mean
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0

Euclidean mean $\mu \in \mathbb{R}^n$:

$$\hat{\mu} = \arg\min_{\mu \in \mathbb{R}^n} \sum_{i=1}^{m} \left[\mathrm{d}\left(X^i, \mu\right)\right]^2$$

Euclidean mean $\mu \in \mathbb{R}^n$:

$$\hat{\mu} = \arg\min_{\mu \in \mathbb{R}^n} \sum_{i=1}^{m} \left[\mathrm{d}\left(X^i, \mu\right) \right]^2$$

Intrinsic mean $\eta \in \mathcal{M}$:

$$\hat{\eta} = \arg\min_{\eta \in \mathcal{M}} \sum_{i=1}^{m} \left[\delta\left(X^i, \eta\right) \right]^2$$

Euclidean mean $\mu \in \mathbb{R}^n$:

$$\hat{\mu} = \arg \min_{\mu \in \mathbb{R}^n} \sum_{i=1}^{m} \left[\mathrm{d}\left(X^i, \mu\right)\right]^2$$

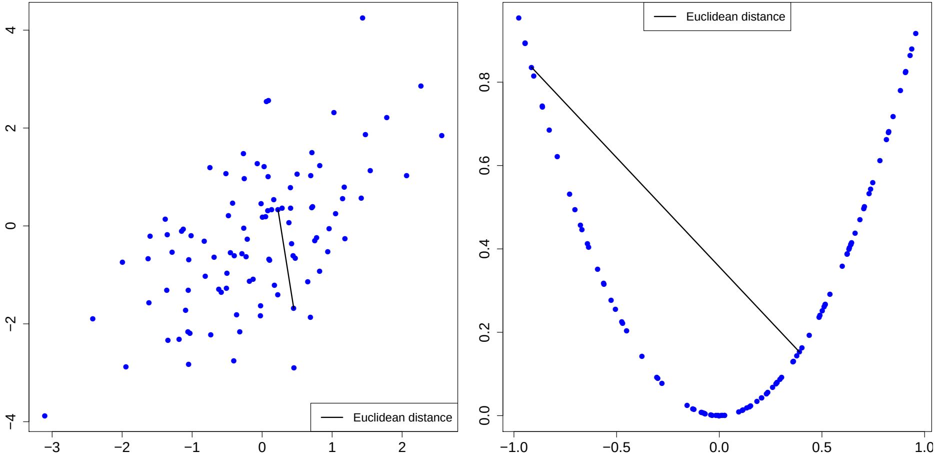
Intrinsic mean $\eta \in \mathcal{M}$:

$$\hat{\eta} = \arg \min_{\eta \in \mathcal{M}} \sum_{i=1}^{m} \left[\delta\left(X^i, \eta\right)\right]^2$$

Intrinsic mean $\eta \in \mathcal{M}$:

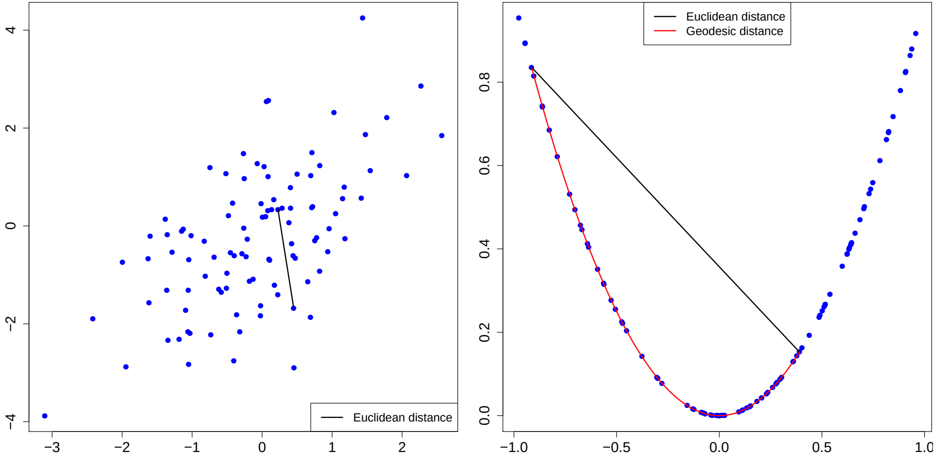
$$\hat{\eta} = \arg \min_{\eta \in \{X^1, \ldots, X^m\}} \sum_{i=1}^{m} \left[\hat{\delta}\left(X^i, \eta\right)\right]^2$$

Idea



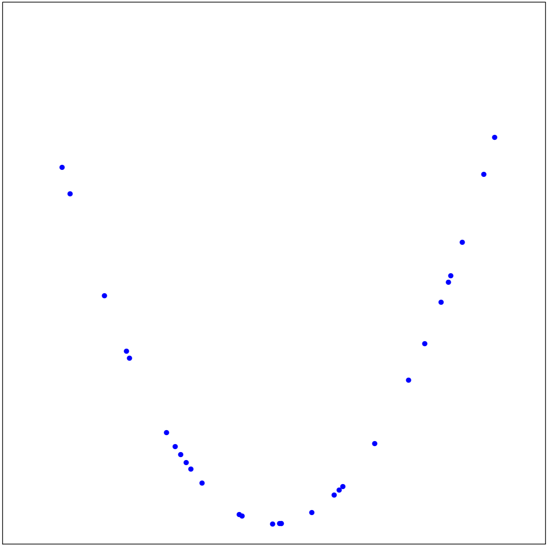
4
2
0
−2
−4
−3
−2
−1
0
1
2
Euclidean distance
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
Euclidean distance

Idea


Euclidean distance
Euclidean distance
Geodesic distance
−3
−2
−1
0
1
2
−4
−2
0
2
4
−1.0
−0.5
0.0
0.5
1.0
0.0
0.2
0.4
0.6
0.8

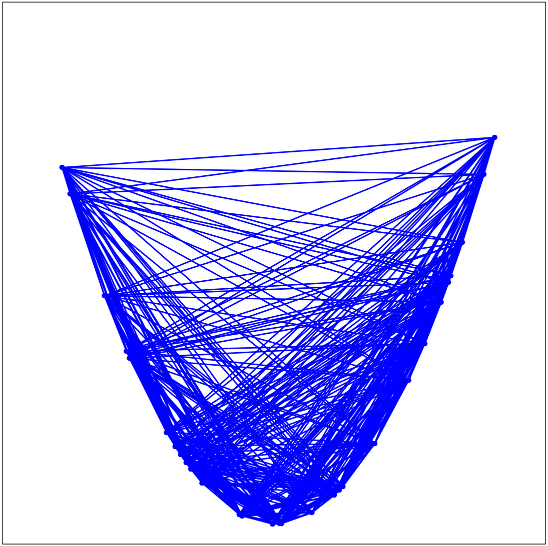
Plan

Approximation of δ



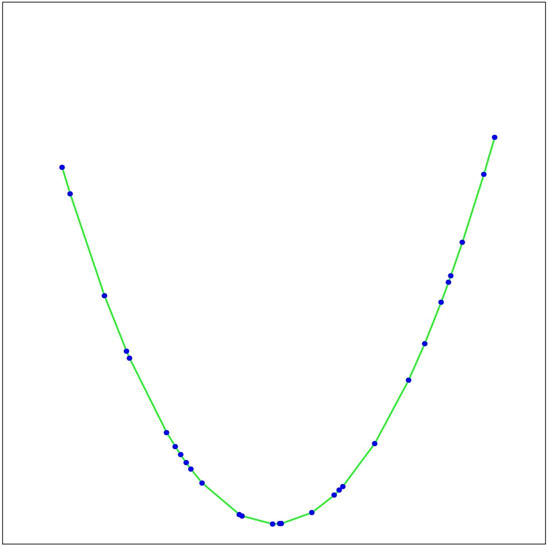
$n = 30$ points

Approximation of δ



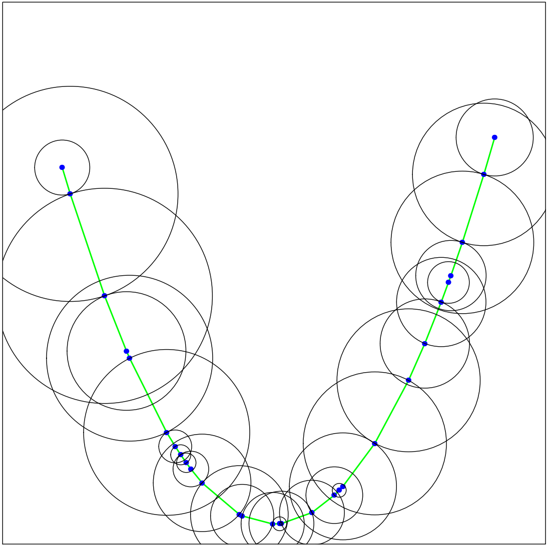
Euclidean complete graph

Approximation of δ



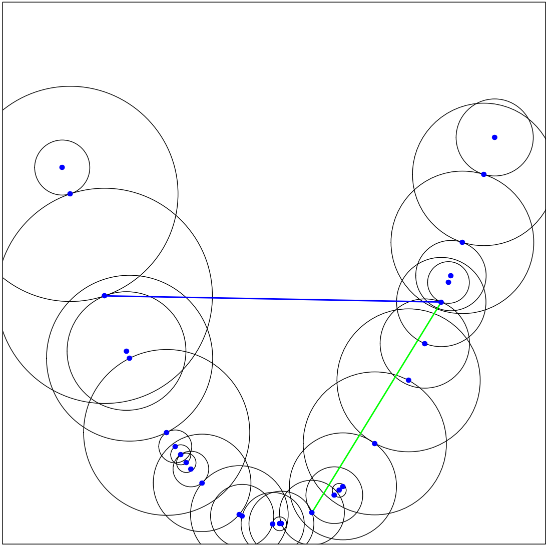
Minimum Spanning Tree (MST)

Approximation of δ



$$\cup_{i=1}^{n} \mathcal{B}\left(X_i, r_i\right)$$

Approximation of δ

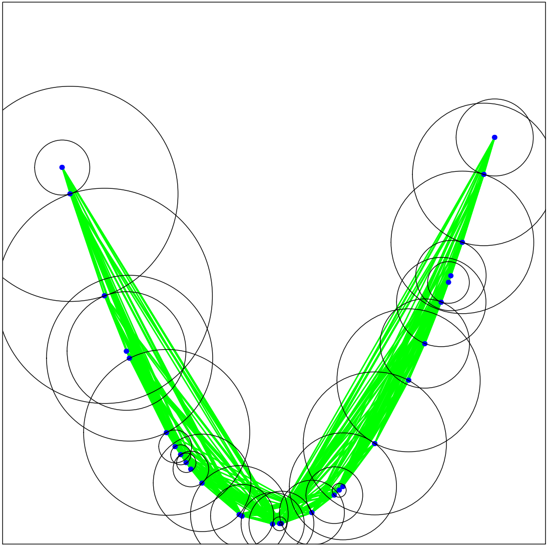


$$\overline{X_i X_j} \in \cup_{k=1}^{n} \mathcal{B}\left(X_k, r_k\right)$$

or

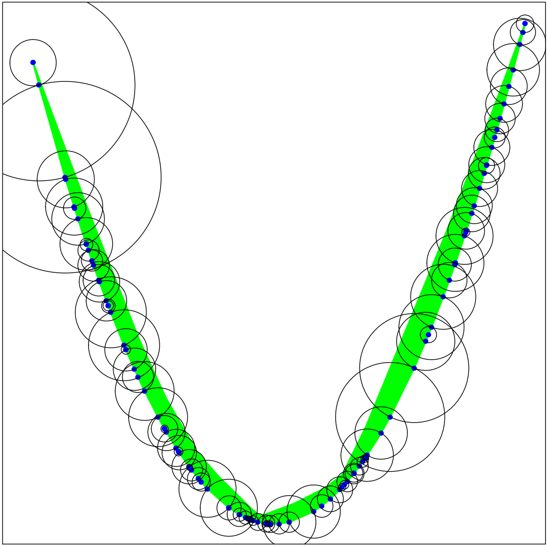
$$\overline{X_i X_j} \notin \cup_{k=1}^{n} \mathcal{B}\left(X_k, r_k\right)$$

Approximation of δ



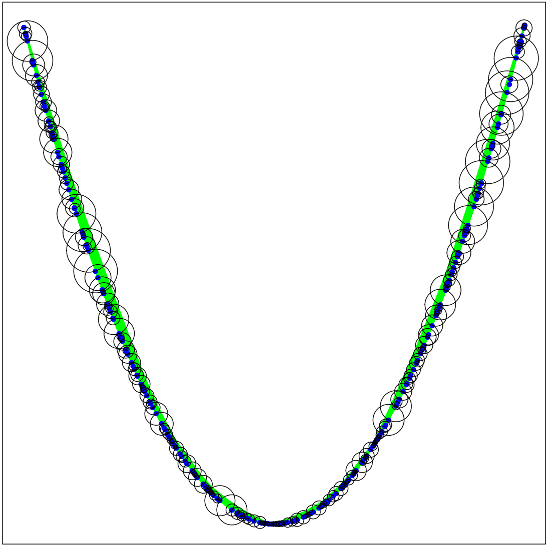
$n = 30$ observed points

Approximation of δ



$n = 100$ observed points

Approximation of δ



$n = 300$ observed points

Plan

1. Microarray data analysis
2. Issue
3. Some solutions
   - Shift model
   - Phase model
   - Quantile normalization
4. New framework
   - Idea
   - Approximation of δ
   - Shift model
   - Application

Model:

$$X_i^j = f\left(t_j - A_i\right),\ i \in \{1, \dots, n\},\ j \in \{1, \dots, m\}$$

with

- $f : \mathbb{R} \rightarrow \mathbb{R}$ an unknown function
- A an unknown real valued variable
- $A_i \sim A$ *iid*
- $t_j \in \mathbb{R}$

We define the structural median of the stochastic function $f(\cdot - A)$ by

$$f_{\mathrm{SM}} = f\left(\cdot - \mathrm{med}(A)\right)$$

A natural estimator of f_{SM} is

$$\widehat{f}_{\mathrm{SM}} = \left(f\left(t_1 - \widehat{\mathrm{med}}(A)\right), f\left(t_2 - \widehat{\mathrm{med}}(A)\right), \ldots, f\left(t_m - \widehat{\mathrm{med}}(A)\right)\right)$$

Let

$$\begin{array}{rcl} X:\mathbb{R} & \longrightarrow & \mathbb{R}^m \\ a & \longmapsto & X(a) = \left(f\left(t_1 - a\right), f\left(t_2 - a\right), \ldots, f\left(t_m - a\right)\right) \end{array}$$

<u>Lemma:</u> The set

$$\mathcal{C} = \{X(a) \in \mathbb{R}^m,\ a \in \mathbb{R}\}$$

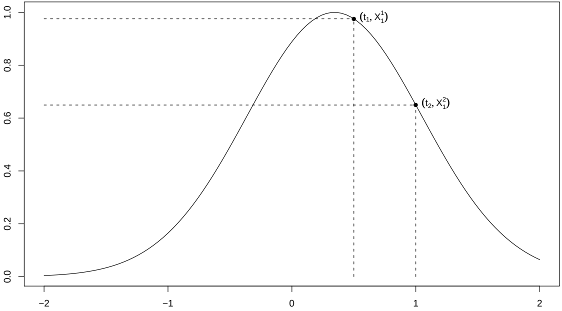
is a 1-dimensional embedded manifold, with distance

$$\delta\left(X_1, X_2\right) = \left| \int_{a_1}^{a_2} \left\| X'(a) \right\| \mathrm{d}a \right|$$

where $X_1 = X\left(a_1\right)$ and $X_2 = X\left(a_2\right)$.

Theorem: we have $\widehat{\mu}_I^1 = \widehat{f}_{\mathrm{SM}}$.

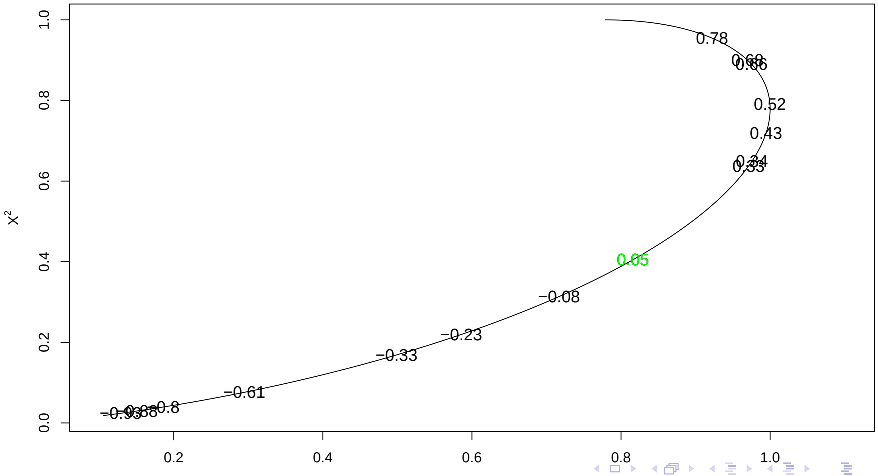
Shift model



Example:

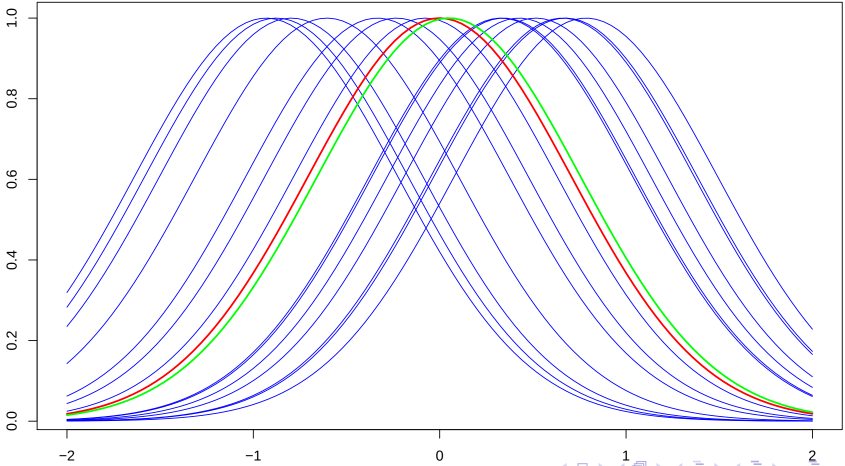
- $f(t) = \exp\left(-t^2\right)$
- $A \sim \mathcal{U}\left(]-1,1[\right)$
- $n = 30,\ m = 2$
- $t_1 = 0.5,\ t_2 = 1$

Shift model



0.78
0.68
0.66
0.52
0.43
0.34
0.33
0.05
−0.08
−0.23
−0.33
−0.61
−0.8
−0.93
−0.88
X2
0.0
0.2
0.4
0.6
0.8
1.0

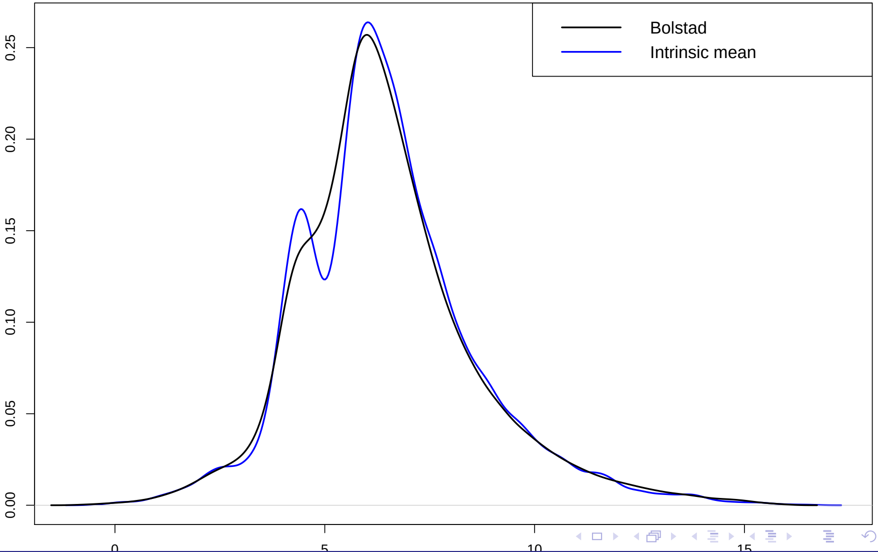
Shift model



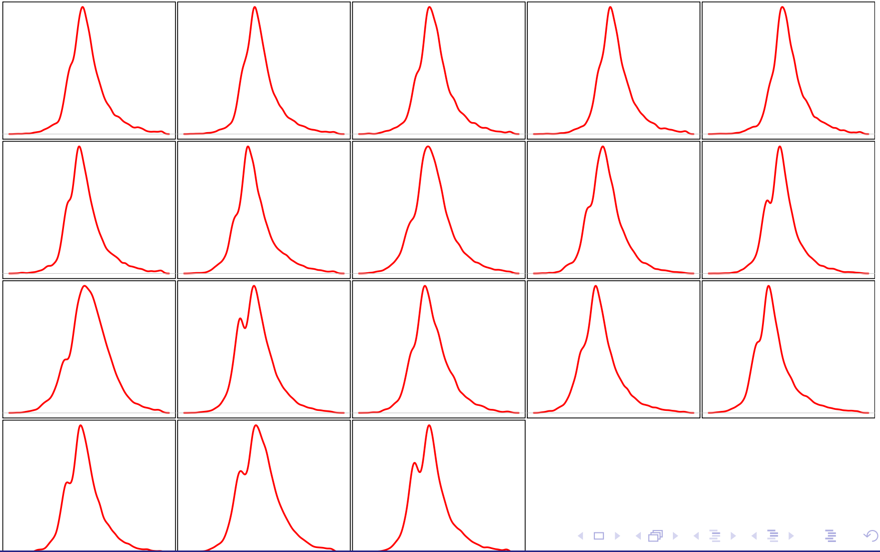
Plan

1. Microarray data analysis
2. Issue
3. Some solutions
   - Shift model
   - Phase model
   - Quantile normalization
4. New framework
   - Idea
   - Approximation of δ
   - Shift model
   - Application

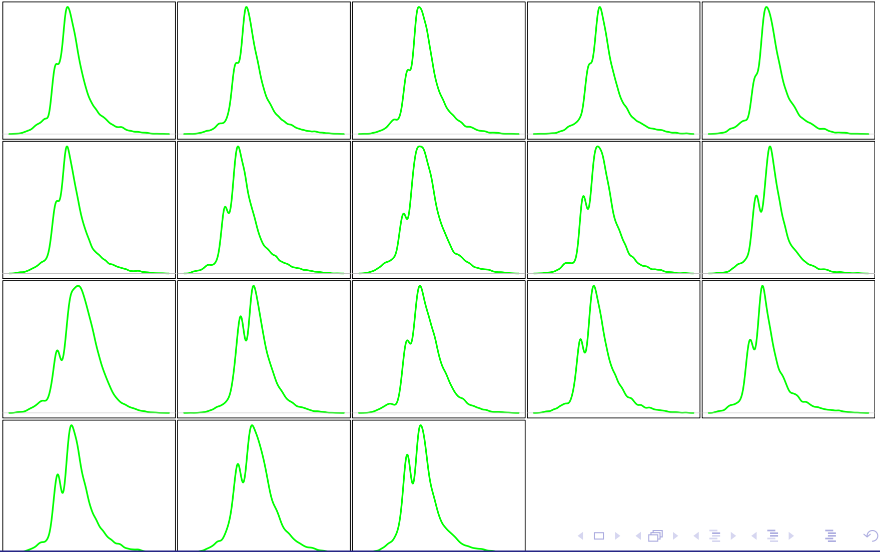
Application



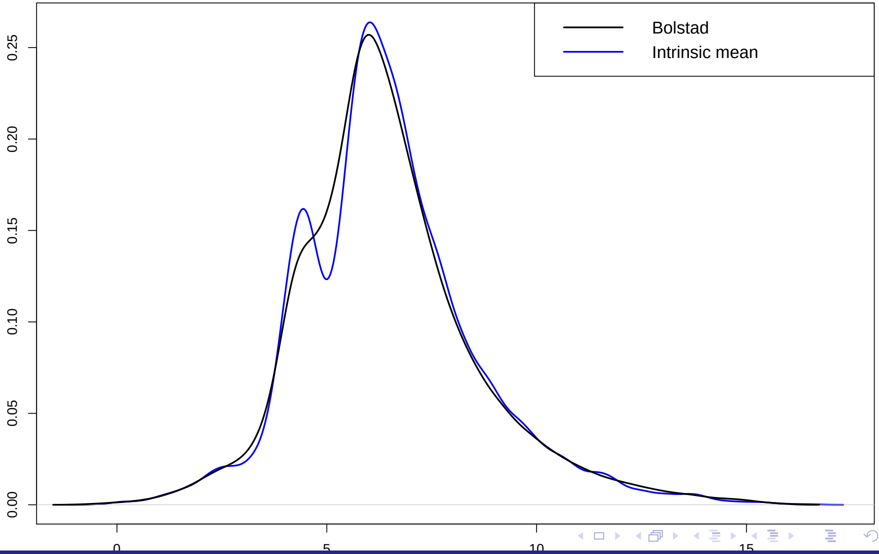
Application



Application



Application



Thank you for your attention!