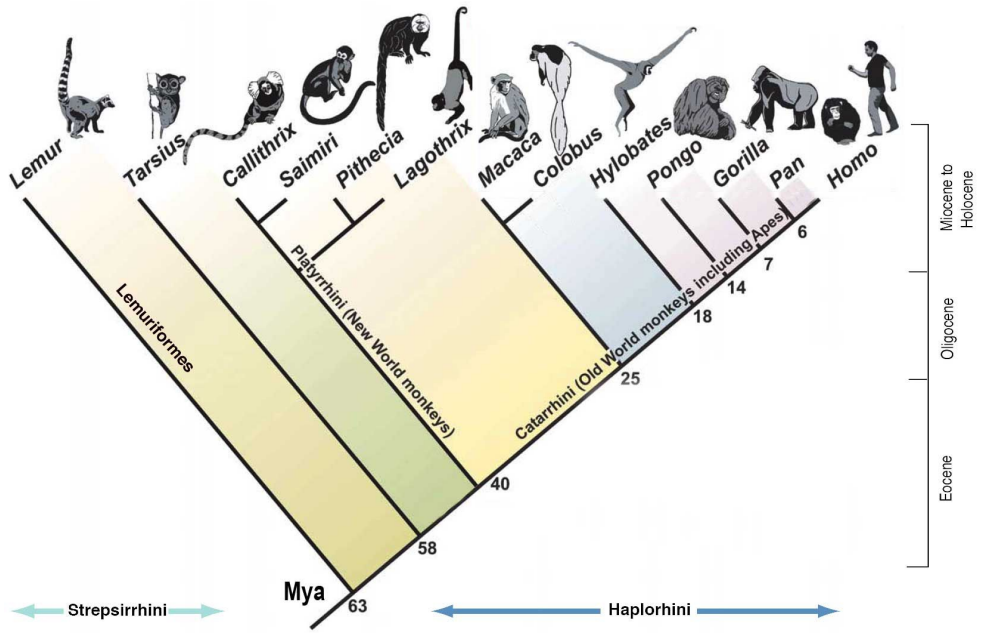
Reconstructing hidden manifolds: a challenge for mathematical community (?)

Eugene Stepanov

Steklov Research Institute of Mathematics, Russian Academy of Sciences

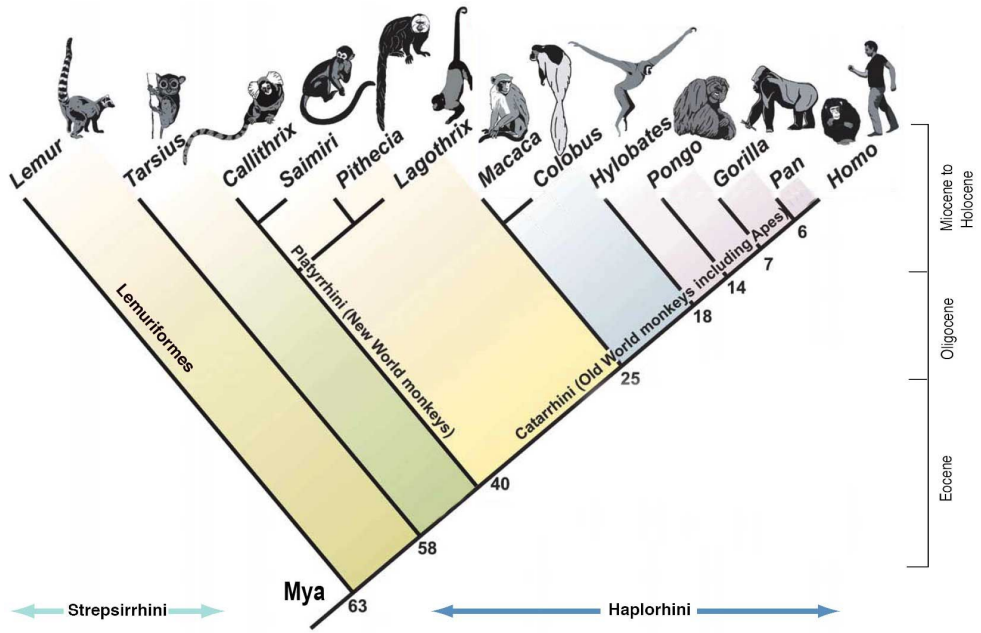
How to study evolution

- Before 1960s: trees based on expert opinion and on morphological data (types of eyes, number of legs etx); informal methods;
- Later: more objective data including molecular biology/genomic data; formal methods;
- Since mid 1990s-2000: using information on evolutionary distances



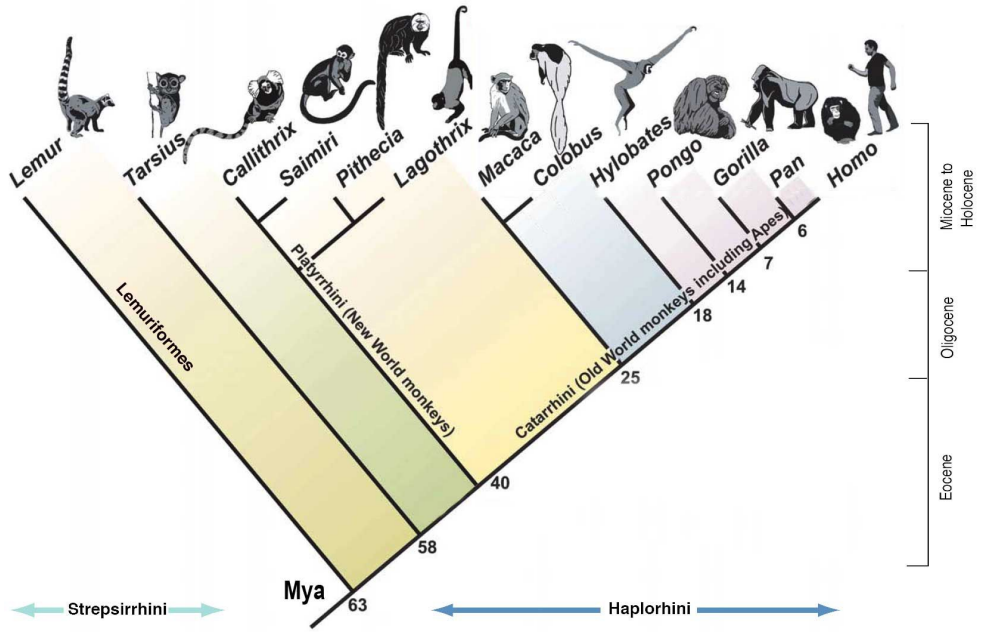
How to study evolution

- Before 1960s: trees based on expert opinion and on morphological data (types of eyes, number of legs etx); informal methods;
- Later: more objective data including molecular biology/genomic data; formal methods;
- Since mid 1990s-2000: using information on evolutionary distances

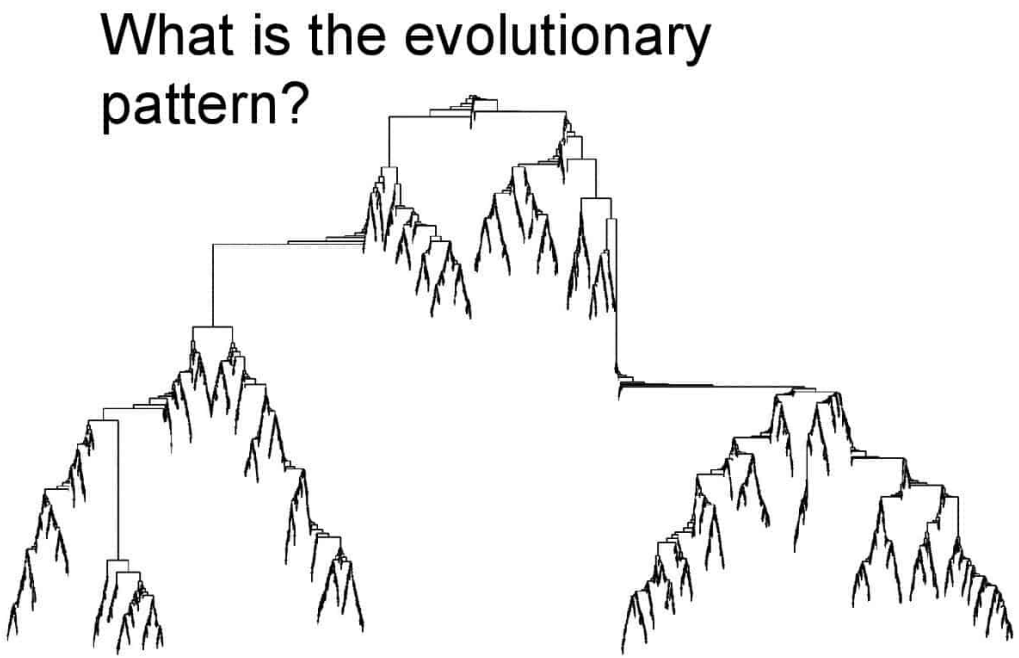


How to study evolution

- Before 1960s: trees based on expert opinion and on morphological data (types of eyes, number of legs etx); informal methods;
- Later: more objective data including molecular biology/genomic data; formal methods;
- Since mid 1990s-2000: using information on evolutionary distances

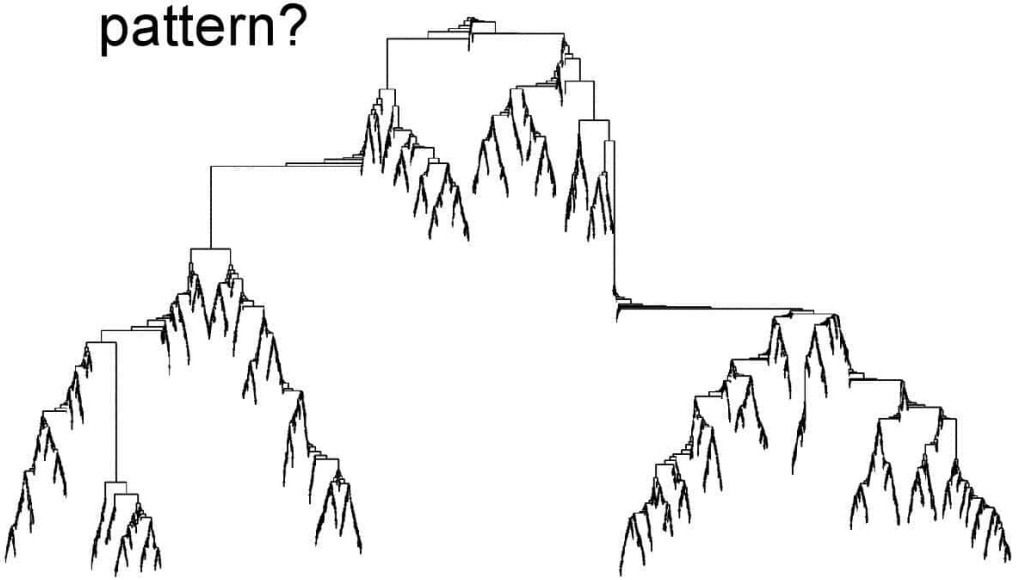


Evolution pattern: a phylogenetic tree?



Evolution pattern: an embedding?

What is the evolutionary pattern?



Classic phylogeny does not operate with empty places

How to read this evolutionary message?

Linnaeus classification system principles

The natural system, unlike the catalogue list "by itself indicates even missed plants…"

Carl Linnaeus "Philosophia botanica"

Some authors look at it (the Natural System) merely as a scheme for arranging together those living objects which are most alike, and for separating those which are most unlike.

Charles Darwin "Origin of species"


CAROLI LINNAEI

EQ. AUR. DE STELLA POLARI ARCHIATRI REGII, MED. ET BOTAN. PROFESS. UPSAL. ACAD. PAR. HOLM. PETROPOL. IMPERIAL. BEROLIN. LONDIN. ANGLICO-EDIMB. MONTPEL. TOLOS. FLORENT. SOC.

PHILOSOPHIA BOTANICA

IN QUA EXPLICANTUR FUNDAMENTA BOTANICA CUM DEFINITIONIBUS PARTIUM, EXEMPLIS TERMINORUM, OBSERVATIONIBUS RARIORUM.

ADIECTIS FIGURIS AENEIS.

EDITIO SECUNDA.

LABORE ET FAVORE.

VIENNÆ,

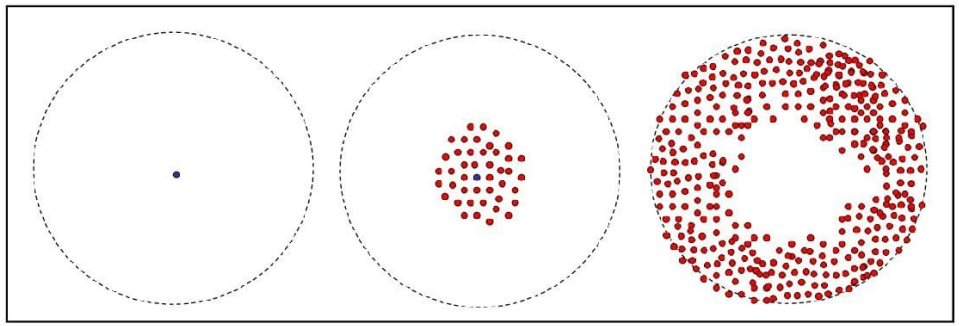
TYPIS JOANNIS THOMAE NOBIL. DE TRATTNERN, CÆS. REG. MAJ. AULÆ TYPOGR. ET BIBLIOPOLÆ.

1 7 8 3.

Evolution: basic hypotheses

The evolutionary process in the Evolutionary space probably looks like (expectations):

Irrevesible (Dollo's law)
Radial (in case of there was no HGT caused 16S rRNA gene mosaicism)
Extention form the LCA point (if the LCA hypothesis is true for Bacteria)

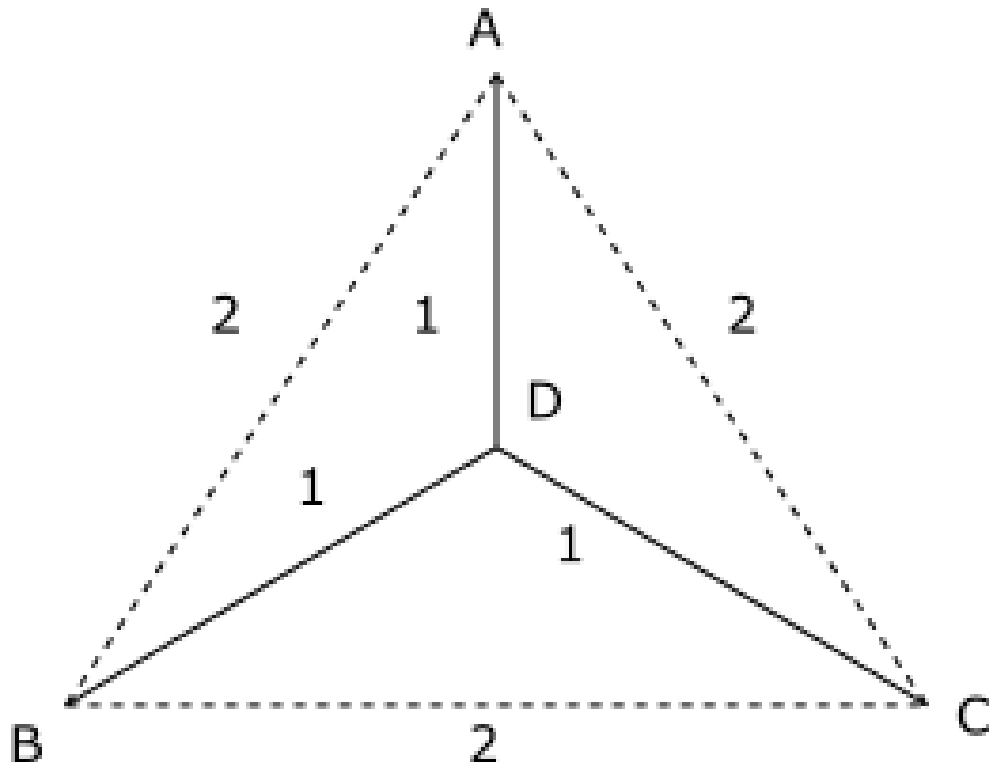


Isometric embedding in a Euclidean space?

(E, d) – finite metric space. Find $f : E \to \mathbb{R}^n$ *isometry*, i.e.

$$|f(x) - f(y)| = d(x, y).$$

3 points embed isometrically in $\mathbb{R}^2$, but already 4 may not embed in any $\mathbb{R}^n$...

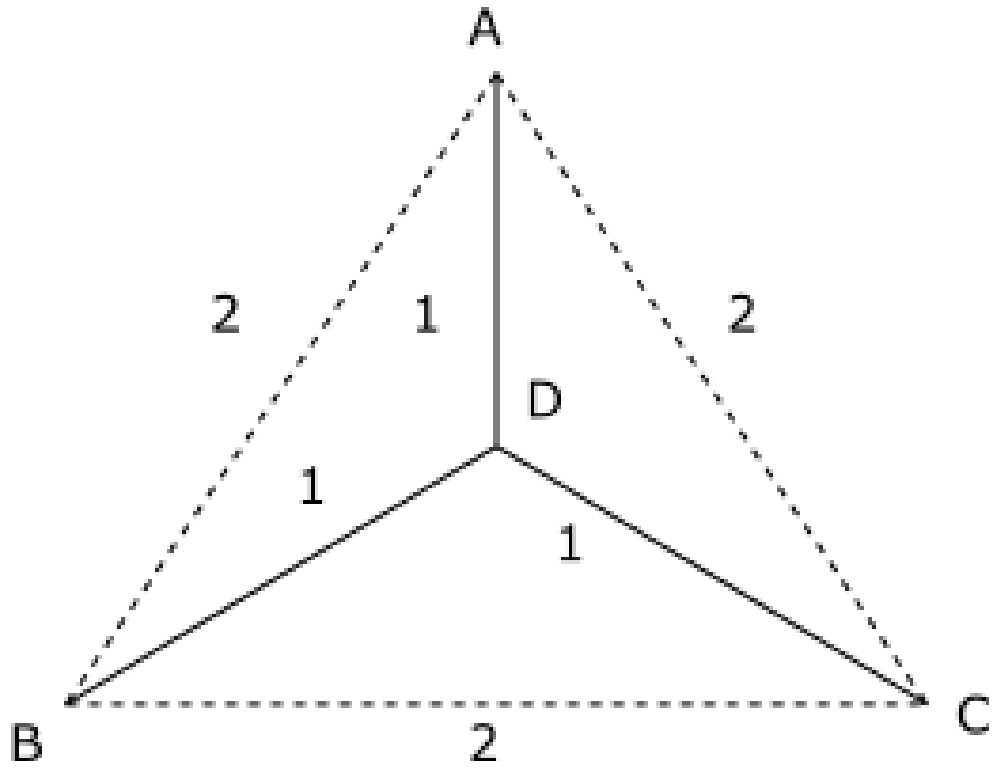


Isometric embedding in a Euclidean space?

(E, d) – finite metric space. Find $f : E \to \mathbb{R}^n$ *isometry*, i.e.

$$|f(x) - f(y)| = d(x, y).$$

3 points embed isometrically in $\mathbb{R}^2$, but already 4 may not embed in any $\mathbb{R}^n$...



Isometric embedding in a Euclidean space

Schoenberg theorem: criterium based on the matrix $\{g_{ij}\}$,

$$g_{ij} := \frac{1}{2}(d_{1i}^2 + d_{1j}^2 - d_{ij}^2) \quad \text{where } d_{ij} := d(x_i, x_j) :$$

(E, d) is isometrically embeddable in $\mathbb{R}^n$, iff $\{g_{ij}\}$ is positive semidefinite.

Equivalently, Blumenthal (Cayley-Menger) theorem: criterium based on Cayley-Menger determinants (express volumes of euclidean simplices):

$$\begin{vmatrix} 0 & d_{12}^2 & d_{13}^2 & \dots & d_{1k}^2 & 1 \\ d_{21}^2 & 0 & d_{23}^2 & \dots & d_{2k}^2 & 1 \\ d_{31}^2 & d_{32}^2 & 0 & \dots & d_{3k}^2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{k1}^2 & d_{k2}^2 & d_{k3}^2 & \dots & 0 & 1 \\ 1 & 1 & 1 & \dots & 1 & 0 \end{vmatrix},$$

Algorithm: multidimensional scaling (MDS).

Isometric embedding in a Euclidean space

Schoenberg theorem: criterium based on the matrix $\{g_{ij}\}$,

$$g_{ij} := \frac{1}{2}(d_{1i}^2 + d_{1j}^2 - d_{ij}^2) \quad \text{where } d_{ij} := d(x_i, x_j) :$$

(E, d) is isometrically embeddable in $\mathbb{R}^n$, iff $\{g_{ij}\}$ is positive semidefinite.

Equivalently, Blumenthal (Cayley-Menger) theorem: criterium based on Cayley-Menger determinants (express volumes of euclidean simplices):

$$\begin{vmatrix} 0 & d_{12}^2 & d_{13}^2 & \dots & d_{1k}^2 & 1 \\ d_{21}^2 & 0 & d_{23}^2 & \dots & d_{2k}^2 & 1 \\ d_{31}^2 & d_{32}^2 & 0 & \dots & d_{3k}^2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{k1}^2 & d_{k2}^2 & d_{k3}^2 & \dots & 0 & 1 \\ 1 & 1 & 1 & \dots & 1 & 0 \end{vmatrix},$$

Algorithm: multidimensional scaling (MDS).

Isometric embedding in a Euclidean space

Schoenberg theorem: criterium based on the matrix $\{g_{ij}\}$,

$$g_{ij} := \frac{1}{2}(d_{1i}^2 + d_{1j}^2 - d_{ij}^2) \quad \text{where } d_{ij} := d(x_i, x_j) :$$

(E, d) is isometrically embeddable in $\mathbb{R}^n$, iff $\{g_{ij}\}$ is positive semidefinite.

Equivalently, Blumenthal (Cayley-Menger) theorem: criterium based on Cayley-Menger determinants (express volumes of euclidean simplices):

$$\begin{vmatrix} 0 & d_{12}^2 & d_{13}^2 & \dots & d_{1k}^2 & 1 \\ d_{21}^2 & 0 & d_{23}^2 & \dots & d_{2k}^2 & 1 \\ d_{31}^2 & d_{32}^2 & 0 & \dots & d_{3k}^2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{k1}^2 & d_{k2}^2 & d_{k3}^2 & \dots & 0 & 1 \\ 1 & 1 & 1 & \dots & 1 & 0 \end{vmatrix},$$

Algorithm: multidimensional scaling (MDS).

Evolutionary space

SILVA database: datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (Bacteria, Archaea and Eukarya).
Currently more than 9 mln rRNA sequences.

Isometric embedding in a Euclidean space

Distance geometry: applications in

- telecommunication networks (e.g. GPS): identify the positions of objects (ships, sensors) known distances between them;
- (bio)chemistry: reconstruct the 3D structure of a protein molecule from the distances between atoms (NMR data).

Isometric embedding in a Euclidean space

Distance geometry: applications in

- telecommunication networks (e.g. GPS): identify the positions of objects (ships, sensors) known distances between them;
- (bio)chemistry: reconstruct the 3D structure of a protein molecule from the distances between atoms (NMR data).

BiLipschitz embedding in a Euclidean space

Theorem (J. Bourgain 1985)

Let $\#E = N$. Then (E, d) can be embedded biLipschitz into $\mathbb{R}^n$ with $n = O(\log^2 N)$ and distortion $O(\log N)$, i.e. there is an $f : E \to \mathbb{R}^n$ with

$$d(x, y) \leq |f(x) - f(y)| \leq Cd(x, y)$$

with $C = O(\log N)$.

Very similar to Johnson-Lindenstrauss lemmma: just to compare,

Theorem (Johnson-Lindenstrauss 1984)

Let $\#E = N$, $E \subset \mathbb{R}^m$. Then $(E, |\cdot|)$ can be for every $\varepsilon > 0$ "compressed" into $\mathbb{R}^n$ with $n = O(\log N/\varepsilon^2)$ and distortion $1 + \varepsilon$, i.e. there is a biLipschitz (almost isometric) embdeeding $f : E \to \mathbb{R}^n$ with

$$d(x, y) \leq |f(x) - f(y)| \leq (1 + \varepsilon)d(x, y).$$

BiLipschitz embedding in a Euclidean space

Theorem (J. Bourgain 1985)

Let $\#E = N$. Then (E, d) can be embedded biLipschitz into $\mathbb{R}^n$ with $n = O(\log^2 N)$ and distortion $O(\log N)$, i.e. there is an $f : E \to \mathbb{R}^n$ with

$$d(x, y) \leq |f(x) - f(y)| \leq Cd(x, y)$$

with $C = O(\log N)$.

Very similar to Johnson-Lindenstrauss lemmma: just to compare,

Theorem (Johnson-Lindenstrauss 1984)

Let $\#E = N$, $E \subset \mathbb{R}^m$. Then $(E, |\cdot|)$ can be for every $\varepsilon > 0$ "compressed" into $\mathbb{R}^n$ with $n = O(\log N/\varepsilon^2)$ and distortion $1 + \varepsilon$, i.e. there is a biLipschitz (almost isometric) embdeeding $f : E \to \mathbb{R}^n$ with

$$d(x, y) \leq |f(x) - f(y)| \leq (1 + \varepsilon)d(x, y).$$

Reconstructing a manifold from intrinsic distances #1

M unknown (compact) smooth Riemannian manifold, $\{y_j\} \subset M$ (say, a dense set) (or, in probablistic setting, i.i.d. random points with uniform law)

$d_{ij} := d_M(y_i, y_j)$ known

Can one reconstruct an M abstract manifold from intrinsic distances?

Reconstructing a manifold from intrinsic distances #1

M unknown (compact) smooth Riemannian manifold, $\{y_j\} \subset M$ (say, a dense set) (or, in probablistic setting, i.i.d. random points with uniform law)

$d_{ij} := d_M(y_i, y_j)$ known

Can one reconstruct an M abstract manifold from intrinsic distances?

Reconstructing a manifold from intrinsic distances #1

M unknown (compact) smooth Riemannian manifold, $\{y_j\} \subset M$ (say, a dense set) (or, in probablistic setting, i.i.d. random points with uniform law)

$d_{ij} := d_M(y_i, y_j)$ known

Can one reconstruct an M abstract manifold from intrinsic distances?

Reconstructing a manifold from intrinsic distances #2

Goal: $M \mapsto \Sigma$, $y_j \mapsto x_j$,

$$d_\Sigma(x_i, x_j) = d_M(y_i, y_j)$$

(isometric embedding), or

$$(1-\varepsilon)d_M(y_i, y_j) \le d_\Sigma(x_i, x_j) \le (1+\varepsilon)d_M(y_i, y_j),$$

$\varepsilon > 0$ small.
(almost isometric embedding)
Algorithms

- MDS (Multidimensional scaling).Most frequently used (in one of numerous modifications). Originally designed for euclidean distances only! It is genuinely believed that nevertheless, it reconstructs intrinsic distances. See e.g.
  Jianzhong Wang. *Geometric structure of high-dimensional data and dimensionality reduction*. Springer.
- MVU (maximum variance unfolding).

Reconstructing a manifold from intrinsic distances #2

Goal: $M \mapsto \Sigma$, $y_j \mapsto x_j$,

$$d_\Sigma(x_i, x_j) = d_M(y_i, y_j)$$

(isometric embedding), or

$$(1-\varepsilon)d_M(y_i, y_j) \leq d_\Sigma(x_i, x_j) \leq (1+\varepsilon)d_M(y_i, y_j),$$

$\varepsilon > 0$ small.
(almost isometric embedding)

Algorithms

- MDS (Multidimensional scaling).Most frequently used (in one of numerous modifications). Originally designed for euclidean distances only! It is genuinely believed that nevertheless, it reconstructs intrinsic distances. See e.g.
  Jianzhong Wang. *Geometric structure of high-dimensional data and dimensionality reduction*. Springer.
- MVU (maximum variance unfolding).

Reconstructing a manifold from intrinsic distances #2

Goal: $M \mapsto \Sigma$, $y_j \mapsto x_j$,

$$d_\Sigma(x_i, x_j) = d_M(y_i, y_j)$$

(isometric embedding), or

$$(1-\varepsilon)d_M(y_i, y_j) \leq d_\Sigma(x_i, x_j) \leq (1+\varepsilon)d_M(y_i, y_j),$$

$\varepsilon > 0$ small.
(almost isometric embedding)
Algorithms

- MDS (Multidimensional scaling). Most frequently used (in one of numerous modifications). Originally designed for euclidean distances only! It is genuinely believed that nevertheless, it reconstructs intrinsic distances. See e.g.
  Jianzhong Wang. *Geometric structure of high-dimensional data and dimensionality reduction*. Springer.
- MVU (maximum variance unfolding).

Reconstructing a manifold from intrinsic distances #2

Goal: $M \mapsto \Sigma$, $y_j \mapsto x_j$,

$$d_\Sigma(x_i, x_j) = d_M(y_i, y_j)$$

(isometric embedding), or

$$(1-\varepsilon)d_M(y_i, y_j) \leq d_\Sigma(x_i, x_j) \leq (1+\varepsilon)d_M(y_i, y_j),$$

$\varepsilon > 0$ small.
(almost isometric embedding)
Algorithms

- MDS (Multidimensional scaling).Most frequently used (in one of numerous modifications). Originally designed for euclidean distances only! It is genuinely believed that nevertheless, it reconstructs intrinsic distances. See e.g.
  Jianzhong Wang. *Geometric structure of high-dimensional data and dimensionality reduction*. Springer.
- MVU (maximum variance unfolding).

Reconstructing a manifold from intrinsic distances #2

Goal: $M \mapsto \Sigma$, $y_j \mapsto x_j$,

$$d_\Sigma(x_i, x_j) = d_M(y_i, y_j)$$

(isometric embedding), or

$$(1-\varepsilon)d_M(y_i, y_j) \leq d_\Sigma(x_i, x_j) \leq (1+\varepsilon)d_M(y_i, y_j),$$

$\varepsilon > 0$ small.
(almost isometric embedding)
Algorithms

- MDS (Multidimensional scaling).Most frequently used (in one of numerous modifications). Originally designed for euclidean distances only! It is genuinely believed that nevertheless, it reconstructs intrinsic distances. See e.g.
  Jianzhong Wang. *Geometric structure of high-dimensional data and dimensionality reduction*. Springer.
- MVU (maximum variance unfolding).

Reconstructing a manifold from intrinsic distances #2

Goal: $M \mapsto \Sigma$, $y_j \mapsto x_j$,

$$d_\Sigma(x_i, x_j) = d_M(y_i, y_j)$$

(isometric embedding), or

$$(1-\varepsilon)d_M(y_i, y_j) \le d_\Sigma(x_i, x_j) \le (1+\varepsilon)d_M(y_i, y_j),$$

$\varepsilon > 0$ small.
(almost isometric embedding)
Algorithms

- MDS (Multidimensional scaling).Most frequently used (in one of numerous modifications). Originally designed for euclidean distances only! It is genuinely believed that nevertheless, it reconstructs intrinsic distances. See e.g.
  Jianzhong Wang. *Geometric structure of high-dimensional data and dimensionality reduction*. Springer.
- MVU (maximum variance unfolding).

Reconstructing a manifold from intrinsic distances #2

Goal: $M \mapsto \Sigma$, $y_j \mapsto x_j$,

$$d_\Sigma(x_i, x_j) = d_M(y_i, y_j)$$

(isometric embedding), or

$$(1-\varepsilon)d_M(y_i, y_j) \le d_\Sigma(x_i, x_j) \le (1+\varepsilon)d_M(y_i, y_j),$$

$\varepsilon > 0$ small.
(almost isometric embedding)
Algorithms

- MDS (Multidimensional scaling).Most frequently used (in one of numerous modifications). Originally designed for euclidean distances only! It is genuinely believed that nevertheless, it reconstructs intrinsic distances. See e.g.
  Jianzhong Wang. *Geometric structure of high-dimensional data and dimensionality reduction*. Springer.
- MVU (maximum variance unfolding).

Reconstructing a manifold from intrinsic distances #2

Goal: $M \mapsto \Sigma$, $y_j \mapsto x_j$,

$$d_\Sigma(x_i, x_j) = d_M(y_i, y_j)$$

(isometric embedding), or

$$(1-\varepsilon)d_M(y_i, y_j) \leq d_\Sigma(x_i, x_j) \leq (1+\varepsilon)d_M(y_i, y_j),$$

$\varepsilon > 0$ small.
(almost isometric embedding)
Algorithms

- MDS (Multidimensional scaling).Most frequently used (in one of numerous modifications). Originally designed for euclidean distances only! It is genuinely believed that nevertheless, it reconstructs intrinsic distances. See e.g.
  Jianzhong Wang. *Geometric structure of high-dimensional data and dimensionality reduction*. Springer.
- MVU (maximum variance unfolding).

Reconstructing a manifold from intrinsic distances #3

- C. Fefferman, S. Ivanov, Ya. Kurylev, M. Lassas, H. Narayaman. Reconstruction and Interpolation of Manifolds I: The geometric Whitney problem.
  `https://arxiv.org/pdf/1508.00674.pdf`
- C. Fefferman, S. Ivanov, M. Lassas, H. Narayaman. Reconstruction of a Riemannian manifold from noisy intrinsic distances. `https://arxiv.org/pdf/1905.07182.pdf`

One can reconstruct an abstract manifold from intrinsic distances without changing too much sectional curvatures

Reconstructing a manifold from intrinsic distances #3

- C. Fefferman, S. Ivanov, Ya. Kurylev, M. Lassas, H. Narayaman. Reconstruction and Interpolation of Manifolds I: The geometric Whitney problem. `https://arxiv.org/pdf/1508.00674.pdf`
- C. Fefferman, S. Ivanov, M. Lassas, H. Narayaman. Reconstruction of a Riemannian manifold from noisy intrinsic distances. `https://arxiv.org/pdf/1905.07182.pdf`

One can reconstruct an abstract manifold from intrinsic distances without changing too much sectional curvatures

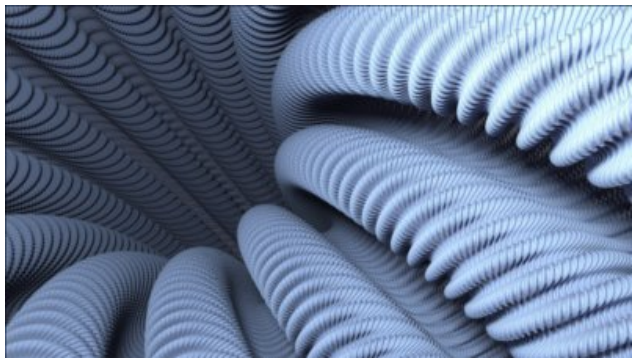
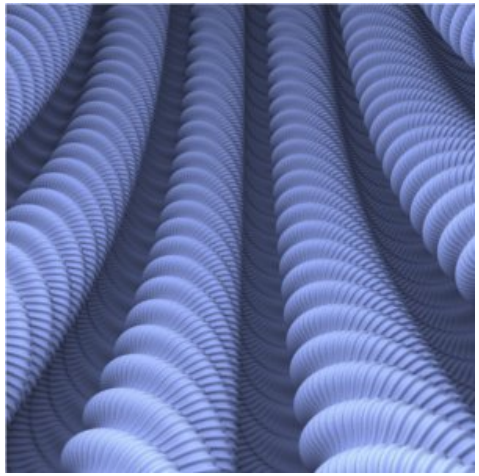
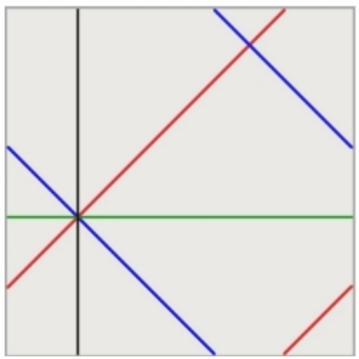
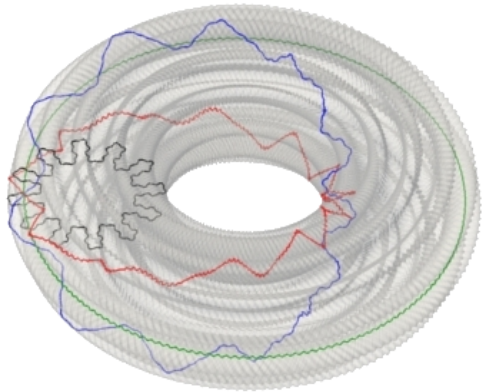
Reconstructing a manifold from intrinsic distances #3

- C. Fefferman, S. Ivanov, Ya. Kurylev, M. Lassas, H. Narayaman. Reconstruction and Interpolation of Manifolds I: The geometric Whitney problem. `https://arxiv.org/pdf/1508.00674.pdf`
- C. Fefferman, S. Ivanov, M. Lassas, H. Narayaman. Reconstruction of a Riemannian manifold from noisy intrinsic distances. `https://arxiv.org/pdf/1905.07182.pdf`

One can reconstruct an abstract manifold from intrinsic distances without changing too much sectional curvatures

How to reconstruct the embedding: curvature/reach conditions are essential

(Square) flat torus isometric embedding in $\mathbb{R}^3$ by Nash-Kuiper theorem (only C^1) via Gromov convex integration construction:

Hevea project: `http://hevea-project.fr/`

Most popular approach: MDS #1

Lara Kassab. Multidimensional scaling: Infinite metric measure spaces.
arXiv:1904.07763, 2019.
Reconstructing S^1 (with intrinsic distances) from N uniformly distributed points

- produces the closed curve $\gamma_N \colon [0, 2\pi] \to \mathbb{R}^n$ defined by

  $$\gamma_N(t) := (a_1^N \cos(t), a_1^N \sin(t), \ldots, a_{2k+1}^N \cos((2k+1)t), a_{2k+1}^N \sin((2k+1)t), \ldots)$$

  where $\lim_N a_j^N = a_j := \sqrt{2}/j$ (with j odd).
- optimal dimension infinite (ℓ^2 instead of $\mathbb{R}^n$),
- one can show

  $$|\gamma(t) - \gamma(s)| = 2\sqrt{\pi}|t-s|^{1/2}.$$

  snowflake instead of a circumference; NO ISOMETRY! But still a homeomorphism...

Most popular approach: MDS #1

Lara Kassab. Multidimensional scaling: Infinite metric measure spaces.
arXiv:1904.07763, 2019.
Reconstructing S^1 (with intrinsic distances) from N uniformly distributed points

- produces the closed curve $\gamma_N \colon [0, 2\pi] \to \mathbb{R}^n$ defined by

  $$\gamma_N(t) := (a_1^N \cos(t), a_1^N \sin(t), \ldots, a_{2k+1}^N \cos((2k+1)t), a_{2k+1}^N \sin((2k+1)t), \ldots)$$

  where $\lim_N a_j^N = a_j := \sqrt{2}/j$ (with j odd).
- optimal dimension infinite (ℓ^2 instead of $\mathbb{R}^n$),
- one can show

  $$|\gamma(t) - \gamma(s)| = 2\sqrt{\pi}|t-s|^{1/2}.$$

  snowflake instead of a circumference; NO ISOMETRY! But still a homeomorphism...

Most popular approach: MDS #1

Lara Kassab. Multidimensional scaling: Infinite metric measure spaces.
arXiv:1904.07763, 2019.
Reconstructing S^1 (with intrinsic distances) from N uniformly distributed points

- produces the closed curve $\gamma_N \colon [0, 2\pi] \to \mathbb{R}^n$ defined by

  $$\gamma_N(t) := (a_1^N \cos(t), a_1^N \sin(t), \ldots, a_{2k+1}^N \cos((2k+1)t), a_{2k+1}^N \sin((2k+1)t), \ldots)$$

  where $\lim_N a_j^N = a_j := \sqrt{2}/j$ (with j odd).
- optimal dimension infinite (ℓ^2 instead of $\mathbb{R}^n$),
- one can show

  $$|\gamma(t) - \gamma(s)| = 2\sqrt{\pi}|t-s|^{1/2}.$$

  snowflake instead of a circumference; NO ISOMETRY! But still a homeomorphism...

Most popular approach: MDS #1

Lara Kassab. Multidimensional scaling: Infinite metric measure spaces.
arXiv:1904.07763, 2019.
Reconstructing S^1 (with intrinsic distances) from N uniformly distributed points

- produces the closed curve $\gamma_N \colon [0, 2\pi] \to \mathbb{R}^n$ defined by

  $$\gamma_N(t) := (a_1^N \cos(t), a_1^N \sin(t), \ldots, a_{2k+1}^N \cos((2k+1)t), a_{2k+1}^N \sin((2k+1)t), \ldots)$$

  where $\lim_N a_j^N = a_j := \sqrt{2}/j$ (with j odd).
- optimal dimension infinite (ℓ^2 instead of $\mathbb{R}^n$),
- one can show

  $$|\gamma(t) - \gamma(s)| = 2\sqrt{\pi}|t - s|^{1/2}.$$

  snowflake instead of a circumference; NO ISOMETRY! But still a homeomorphism...

Most popular approach: MDS #1

Lara Kassab. Multidimensional scaling: Infinite metric measure spaces.
arXiv:1904.07763, 2019.
Reconstructing S^1 (with intrinsic distances) from N uniformly distributed points

- produces the closed curve $\gamma_N \colon [0, 2\pi] \to \mathbb{R}^n$ defined by

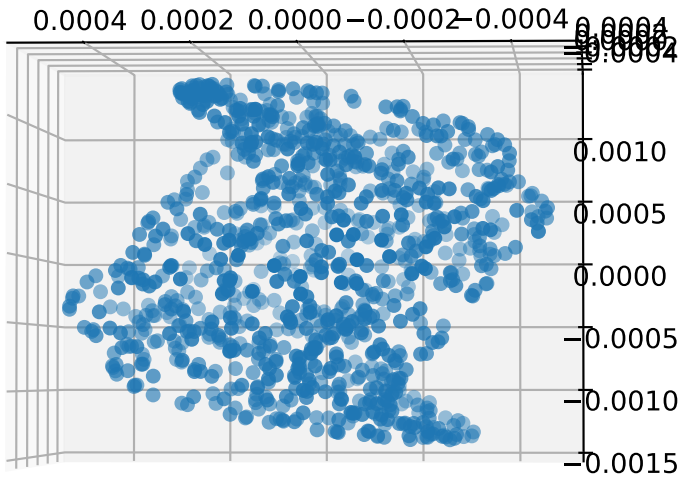
  $$\gamma_N(t) := (a_1^N \cos(t), a_1^N \sin(t), \ldots, a_{2k+1}^N \cos((2k+1)t), a_{2k+1}^N \sin((2k+1)t), \ldots)$$

  where $\lim_N a_j^N = a_j := \sqrt{2}/j$ (with j odd).
- optimal dimension infinite (ℓ^2 instead of $\mathbb{R}^n$),
- one can show

  $$|\gamma(t) - \gamma(s)| = 2\sqrt{\pi}|t - s|^{1/2}.$$

  snowflake instead of a circumference; NO ISOMETRY! But still a homeomorphism...

Most popular approach: MDS #2



The two-dimensional sphere S^2 reconstructed by MDS.

Variational approach for manifold reconstruction #1

M compact smooth submanifold in $\mathbb{R}^n$: $M \subset \mathbb{R}^n$,

$$\operatorname{Reach} M := \{\rho > 0 \colon \text{every } x \in (M)_\rho \text{ has a unique projection on } M\} > 0.$$

(normal injectivity radius from M)
$\Sigma_k \subset M$ closed sets ("data points"), $\Sigma_k \to M$ in Hausdorff distance.
Fixed an $\varepsilon > 0$ and a $k \in \mathbb{N}$, define the functionals

$$F_{\varepsilon,k} \colon C(M;\mathbb{R}^n) \to \mathbb{R}, F_\varepsilon \colon C(M;\mathbb{R}^n) \to \mathbb{R}$$

by the formula

$$F_{\varepsilon,k}(f) := \sup\left\{\left|\frac{|f(x)-f(y)|^2}{d_M^2(x,y)} - 1\right| : \{x,y\} \subset \Sigma_k, 0 < d_M(x,y) \le \varepsilon\right\},$$

$$F_\varepsilon(f) := \sup\left\{\left|\frac{|f(x)-f(y)|^2}{d_M^2(x,y)} - 1\right| : 0 < d_M(x,y) \le \varepsilon\right\}.$$

Variational approach for manifold reconstruction #1

M compact smooth submanifold in $\mathbb{R}^n$: $M \subset \mathbb{R}^n$,

Reach $M :=$

$\{\rho > 0 \colon$ every $x \in (M)_\rho$ has a unique projection on $M\} > 0$.

(normal injectivity radius from M)

$\Sigma_k \subset M$ closed sets ("data points"), $\Sigma_k \to M$ in Hausdorff distance.
Fixed an $\varepsilon > 0$ and a $k \in \mathbb{N}$, define the functionals

$$F_{\varepsilon,k} \colon C(M;\mathbb{R}^n) \to \mathbb{R}, F_\varepsilon \colon C(M;\mathbb{R}^n) \to \mathbb{R}$$

by the formula

$$F_{\varepsilon,k}(f) := \sup\left\{\left|\frac{|f(x)-f(y)|^2}{d_M^2(x,y)} - 1\right| : \{x,y\} \subset \Sigma_k, 0 < d_M(x,y) \le \varepsilon\right\},$$

$$F_\varepsilon(f) := \sup\left\{\left|\frac{|f(x)-f(y)|^2}{d_M^2(x,y)} - 1\right| : 0 < d_M(x,y) \le \varepsilon\right\}.$$

Variational approach for manifold reconstruction #1

M compact smooth submanifold in $\mathbb{R}^n$: $M \subset \mathbb{R}^n$,

Reach $M :=$

$\{\rho > 0 \colon$ every $x \in (M)_\rho$ has a unique projection on $M\} > 0$.

(normal injectivity radius from M)
$\Sigma_k \subset M$ closed sets ("data points"), $\Sigma_k \to M$ in Hausdorff distance.
Fixed an $\varepsilon > 0$ and a $k \in \mathbb{N}$, define the functionals

$$F_{\varepsilon,k} \colon C(M;\mathbb{R}^n) \to \mathbb{R}, F_\varepsilon \colon C(M;\mathbb{R}^n) \to \mathbb{R}$$

by the formula

$$F_{\varepsilon,k}(f) := \sup\left\{\left|\frac{|f(x)-f(y)|^2}{d_M^2(x,y)} - 1\right| : \{x,y\} \subset \Sigma_k, 0 < d_M(x,y) \le \varepsilon\right\},$$

$$F_\varepsilon(f) := \sup\left\{\left|\frac{|f(x)-f(y)|^2}{d_M^2(x,y)} - 1\right| : 0 < d_M(x,y) \le \varepsilon\right\}.$$

Variational approach for manifold reconstruction #1

M compact smooth submanifold in $\mathbb{R}^n$: $M \subset \mathbb{R}^n$,

Reach $M :=$

$$\{\rho > 0 \colon \text{every } x \in (M)_\rho \text{ has a unique projection on } M\} > 0.$$

(normal injectivity radius from M)
$\Sigma_k \subset M$ closed sets ("data points"), $\Sigma_k \to M$ in Hausdorff distance.
Fixed an $\varepsilon > 0$ and a $k \in \mathbb{N}$, define the functionals

$$F_{\varepsilon,k} \colon C(M;\mathbb{R}^n) \to \mathbb{R}, F_\varepsilon \colon C(M;\mathbb{R}^n) \to \mathbb{R}$$

by the formula

$$F_{\varepsilon,k}(f) := \sup\left\{\left|\frac{|f(x)-f(y)|^2}{d_M^2(x,y)} - 1\right| : \{x,y\} \subset \Sigma_k, 0 < d_M(x,y) \le \varepsilon\right\},$$

$$F_\varepsilon(f) := \sup\left\{\left|\frac{|f(x)-f(y)|^2}{d_M^2(x,y)} - 1\right| : 0 < d_M(x,y) \le \varepsilon\right\}.$$

Variational approach for manifold reconstruction #1

M compact smooth submanifold in $\mathbb{R}^n$: $M \subset \mathbb{R}^n$,

Reach $M :=$

$$\{\rho > 0 \colon \text{every } x \in (M)_\rho \text{ has a unique projection on } M\} > 0.$$

(normal injectivity radius from M)
$\Sigma_k \subset M$ closed sets ("data points"), $\Sigma_k \to M$ in Hausdorff distance.
Fixed an $\varepsilon > 0$ and a $k \in \mathbb{N}$, define the functionals

$$F_{\varepsilon,k} \colon C(M;\mathbb{R}^n) \to \mathbb{R}, F_\varepsilon \colon C(M;\mathbb{R}^n) \to \mathbb{R}$$

by the formula

$$F_{\varepsilon,k}(f) := \sup\left\{ \left| \frac{|f(x)-f(y)|^2}{d_M^2(x,y)} - 1 \right| : \{x,y\} \subset \Sigma_k, 0 < d_M(x,y) \le \varepsilon \right\},$$

$$F_\varepsilon(f) := \sup\left\{ \left| \frac{|f(x)-f(y)|^2}{d_M^2(x,y)} - 1 \right| : 0 < d_M(x,y) \le \varepsilon \right\}.$$

Variational approach for manifold reconstruction #2

Theorem

There are ε_0, C_1, C_2 *(depending on* Reach$M$ *and* diam$M$*), such that the variational problems*

$$\min\{F_{\varepsilon,k}(f) : f \in \mathcal{C}\}, \quad \textit{where}$$
$$\mathcal{C} := \{f \in C(M : \mathbb{R}^n), f(x_0) = 0, |f(x) - f(y)| \geq C_2 d_M(x,y)\}, \tag{P_k}$$

have solutions for every $k \in \mathbb{N}$, $\varepsilon < \varepsilon_0$. *If* f_k *solves* (P_k), *then up to a subsequence* $\lim_k f_k = f$, *where* f *solves*

$$\min\{F_\varepsilon(f) : f \in \mathcal{C}\} \tag{P}$$

Variational approach for manifold reconstruction #2

Theorem

There are ε_0, C_1, C_2 (depending on Reach$M$ and diam$M$), such that the variational problems

$$\min\{F_{\varepsilon,k}(f) : f \in \mathcal{C}\}, \quad \text{where}$$
$$\mathcal{C} := \{f \in C(M : \mathbb{R}^n), f(x_0) = 0, |f(x) - f(y)| \geq C_2 d_M(x,y)\}, \tag{P_k}$$

have solutions for every $k \in \mathbb{N}$, $\varepsilon < \varepsilon_0$. If f_k solves (P_k), then up to a subsequence $\lim_k f_k = f$, where f solves

$$\min\{F_\varepsilon(f) : f \in \mathcal{C}\} \tag{P}$$

Variational approach for manifold reconstruction #3

Moreover,

$$d_M(x,y)(1-C_1\varepsilon) \leq |f(x)-f(y)| \leq d_M(x,y)(1+C_1\varepsilon), \tag{1}$$

if $d_M(x,y) < \varepsilon$, and

$$d_M(x,y)(1-C_1\varepsilon) \leq d_\Sigma(f(x),f(y)) \leq d_M(x,y)(1+C_1\varepsilon), \tag{2}$$

where $\Sigma := f(M)$. N.B. $f_k(M) \to \Sigma$ in Hausdorff distance.

Variational approach for manifold reconstruction #3

Moreover,

$$d_M(x,y)(1-C_1\varepsilon) \leq |f(x)-f(y)| \leq d_M(x,y)(1+C_1\varepsilon), \tag{1}$$

if $d_M(x,y) < \varepsilon$, and

$$d_M(x,y)(1-C_1\varepsilon) \leq d_\Sigma(f(x),f(y)) \leq d_M(x,y)(1+C_1\varepsilon), \tag{2}$$

where $\Sigma := f(M)$. N.B. $f_k(M) \to \Sigma$ in Hausdorff distance.

Towards an algorithm: discrete setting #1

Let $\{y_i\} \subset M$ be a dense set in M,

$$d_{ij} := d_M(y_i, y_j).$$

Fixed an $\varepsilon > 0$ and a $k \in \mathbb{N}$, define the functional $F_{\varepsilon,k} \colon (\mathbb{R}^n)^k \to \mathbb{R}$ by the formula

$$F_{\varepsilon,k}(x_1, \ldots, x_k) := \max \left\{ \left| \frac{|x_i - x_j|^2}{d_{ij}^2} - 1 \right| : i, j = 1, \ldots, k, i \neq j, d_{ij} < \varepsilon \right\}.$$

Towards an algorithm: discrete setting #1

Let $\{y_i\} \subset M$ be a dense set in M,

$$d_{ij} := d_M(y_i, y_j).$$

Fixed an $\varepsilon > 0$ and a $k \in \mathbb{N}$, define the functional $F_{\varepsilon,k} \colon (\mathbb{R}^n)^k \to \mathbb{R}$ by the formula

$$F_{\varepsilon,k}(x_1, \ldots, x_k) := \max \left\{ \left| \frac{|x_i - x_j|^2}{d_{ij}^2} - 1 \right| : i, j = 1, \ldots, k, i \neq j, d_{ij} < \varepsilon \right\}.$$

Towards an algorithm: discrete setting #2

Proposition

Let $(x_i^k)_{i=1}^k \in (\mathbb{R}^n)^k$ *be a minimizer of* $F_{\varepsilon,k}$ *with* $\varepsilon < \varepsilon_0$ *over*

$$X^k := \left\{((x_i)_{i=1}^k \in (\mathbb{R}^n)^k : |x_i - x_j| \geq C_2 d_{ij}, i,j = 1,\dots,k\right\}.$$

Then up to a subsequence $x_i^k \to x_i$ *as* $k \to \infty$, *and*

$$d_{ij}(1 - C_1\varepsilon) \leq |x_i - x_j| \leq d_{ij}(1 + C_1\varepsilon),$$

whenever $d_{ij} < \varepsilon$, *and*

$$d_{ij}(1 - C_1\varepsilon) \leq d_\Sigma(x_i, x_j) \leq d_{ij}(1 + C_1\varepsilon)$$

for all $\{i,j\} \subset \mathbb{N}$.

An algorithm: convex programming

Define $K_{ij} := x_i \cdot x_j$ (Gram matrix of a set of vectors $\{x_i\}$),

$$G_{\varepsilon,k}(x_1,\ldots,x_k) := \max\left\{\left|\frac{K_{ii}+K_{jj}-2K_{ij}}{d_{ij}^2}-1\right| : i,j=1,\ldots,k, i\neq j, d_{ij}<\varepsilon\right\}$$

Problem: minimize G over the set of positive semidefinite matrices K satisfying the set of linear constraints

$$K_{ii}+K_{jj}-2K_{ij} \geq C_2^2 d_{ij}^2, i,j=1,\ldots,k.$$

An algorithm: convex programming

Define $K_{ij} := x_i \cdot x_j$ (Gram matrix of a set of vectors $\{x_i\}$),

$$G_{\varepsilon,k}(x_1,\ldots,x_k) := \max\left\{\left|\frac{K_{ii}+K_{jj}-2K_{ij}}{d_{ij}^2}-1\right| : i,j=1,\ldots,k, i\neq j, d_{ij}<\varepsilon\right\}$$

Problem: minimize G over the set of positive semidefinite matrices K satisfying the set of linear constraints

$$K_{ii}+K_{jj}-2K_{ij} \geq C_2^2 d_{ij}^2, i,j=1,\ldots,k.$$

An algorithm: even better, semidefinite programming #1

Adding a scalar variable $t \in \mathbb{R}$:

$$\begin{aligned}
&\text{minimize} \quad t \quad \text{over the pairs} \quad (t, K) \quad \text{subject to} \\
&\qquad -td_{ij}^2 \leq K_{ii} + K_{jj} - 2K_{ij} - d_{ij}^2 \leq td_{ij}^2, \\
&\qquad\qquad \text{for all } i, j = 1, \ldots, k, i \neq j, d_{ij} < \varepsilon, \\
&\qquad K_{ii} + K_{jj} - 2K_{ij} \geq C_2^2 d_{ij}^2, \\
&\qquad\qquad \text{for all } i, j = 1, \ldots, k, i \neq j, \\
&\qquad K \quad \text{positive semidefinite } k \times k \text{ matrix.}
\end{aligned}$$

All constraints linear except the last one (which is convex cone constraint).
Seemingly (but not really!) similar to maximum variance unfolding algorithm (MVU).

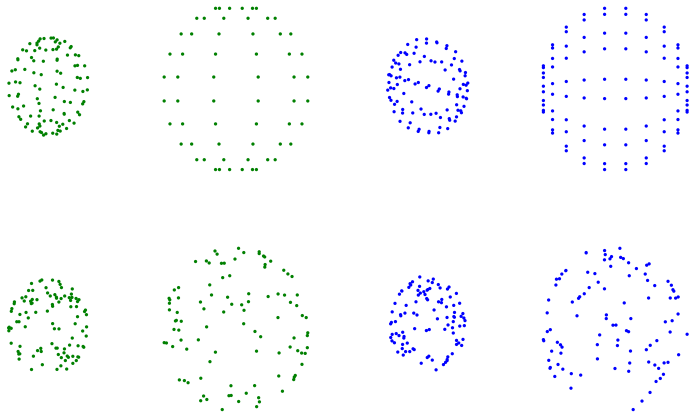
An algorithm: even better, semidefinite programming #1

Adding a scalar variable $t \in \mathbb{R}$:

$$\begin{aligned}
&\text{minimize} \quad t \quad \text{over the pairs} \quad (t, K) \quad \text{subject to} \\
&\qquad -td_{ij}^2 \leq K_{ii} + K_{jj} - 2K_{ij} - d_{ij}^2 \leq td_{ij}^2, \\
&\qquad\qquad \text{for all } i, j = 1, \ldots, k, i \neq j, d_{ij} < \varepsilon, \\
&\qquad K_{ii} + K_{jj} - 2K_{ij} \geq C_2^2 d_{ij}^2, \\
&\qquad\qquad \text{for all } i, j = 1, \ldots, k, i \neq j, \\
&\qquad K \quad \text{positive semidefinite } k \times k \text{ matrix.}
\end{aligned}$$

All constraints linear except the last one (which is convex cone constraint).
Seemingly (but not really!) similar to maximum variance unfolding algorithm (MVU).

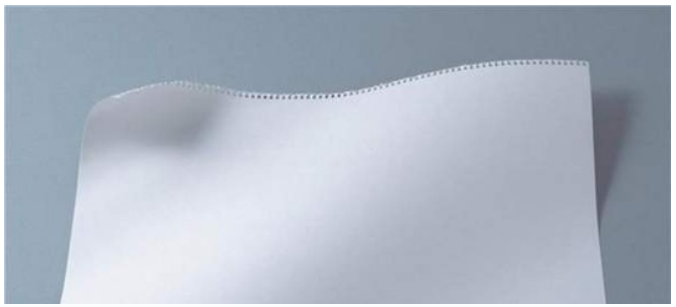
An algorithm: even better, semidefinite programming #2



Reconstruction of a unit sphere from pairwise distances. Top line, columns 1 and 2: points on a grid on the sphere. Top line, columns 3 and 4: the recovered points of the unit sphere. Bottom line, columns 1 and 2: points drawn from the uniform distribution on unit sphere. Bottom line, columns 3 and 4: the recovered points from approximate geodesic distances.

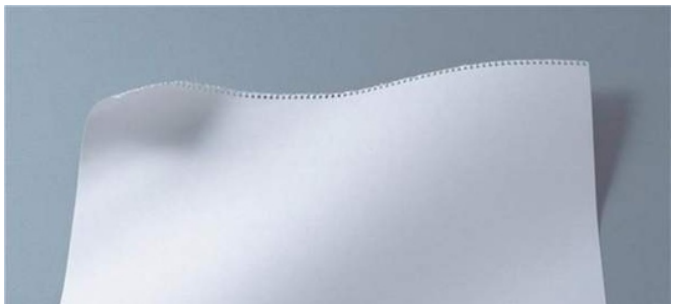
Problems and challenges

- **Slow** algorithm (number of unknowns $O(N^2)$, where N is the number of data points). Can one do, e.g. $O(N \log N)$? Fast algorithms for topological data analysis?
- Hypotheses on curvature/reach essential, but **cannot** be deduced from the data!
  Problems with biological/phylogeny applications? Natural e.g. for 3D molecule reconstruction from NMR data.
- Too many solutions! E.g. a bent sheet of paper is indistinguishable from the flat one.



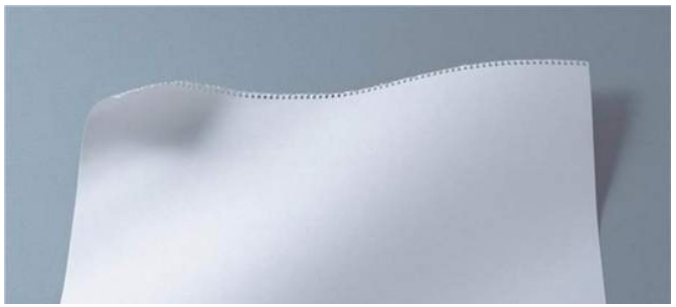
Problems and challenges

- **Slow** algorithm (number of unknowns $O(N^2)$, where N is the number of data points). Can one do, e.g. $O(N \log N)$? Fast algorithms for topological data analysis?
- Hypotheses on curvature/reach essential, but **cannot** be deduced from the data!
  Problems with biological/phylogeny applications? Natural e.g. for 3D molecule reconstruction from NMR data.
- Too many solutions! E.g. a bent sheet of paper is indistinguishable from the flat one.



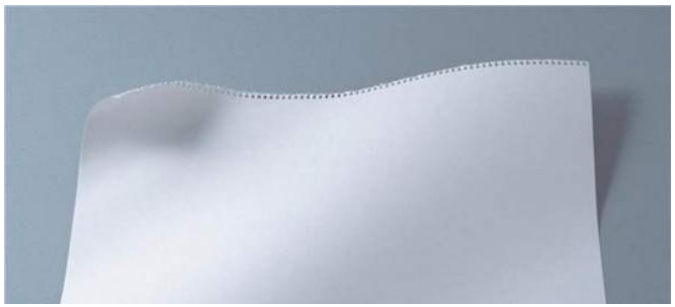
Problems and challenges

- **Slow** algorithm (number of unknowns $O(N^2)$, where N is the number of data points). Can one do, e.g. $O(N \log N)$? Fast algorithms for topological data analysis?
- Hypotheses on curvature/reach essential, but **cannot** be deduced from the data!
  Problems with biological/phylogeny applications? Natural e.g. for 3D molecule reconstruction from NMR data.
- Too many solutions! E.g. a bent sheet of paper is indistinguishable from the flat one.



Problems and challenges

- **Slow** algorithm (number of unknowns $O(N^2)$, where N is the number of data points). Can one do, e.g. $O(N \log N)$? Fast algorithms for topological data analysis?
- Hypotheses on curvature/reach essential, but **cannot** be deduced from the data!
  Problems with biological/phylogeny applications? Natural e.g. for 3D molecule reconstruction from NMR data.
- Too many solutions! E.g. a bent sheet of paper is indistinguishable from the flat one.



Topological data analysis: computing Čech cohomologies #1

M and $\Sigma := f(M)$ are homeomorphic (even biLipschitz equivalent)
Let

- $\Lambda \subset \mathbb{N}$ be a finite set of indices such that $\{Y_\lambda\}$ is a finite δ-net of of M (equipped with d_M),
- $C_\Sigma(r)$ the Čech complex built on the euclidean balls $B_r(X_\lambda)$, $X_\lambda := f(Y_\lambda)$,
- $C_M(r)$ the Čech complex built on the euclidean balls $B_r(Y_\lambda)$.

The vertices of these complexes may be considered the same (namely, the set of vertices of all them may be identified with the index set Λ).

Topological data analysis: computing Čech cohomologies #2

Proposition

Let σ and δ be small enough (below some precise threshold depending on Reach M), Then $H^*(C_\Sigma(\sigma); \mathbb{R}) \simeq H^*(M; \mathbb{R})$, H^* standing for the Čech cohomology.

Topological data analysis: computing Čech cohomologies #3

Remark

One may take Y_λ to be drawn by sampling M in i.i.d. way according to the volume measure on M. In fact if $\#\Lambda > n(M, \rho, p)$, then

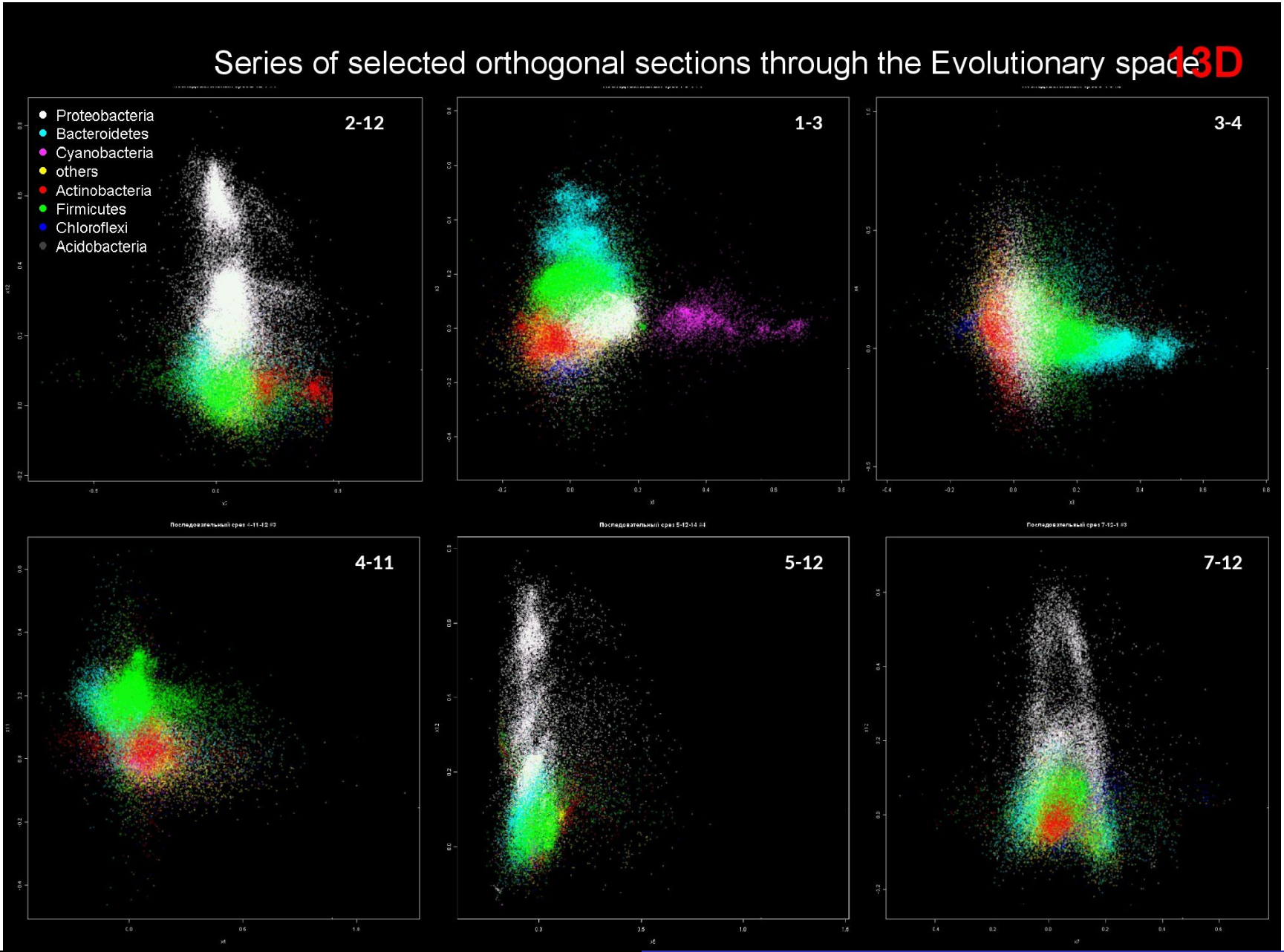
$$M \subset \bigcup_\lambda B_\rho(\bar{Y}_\lambda)$$

with probability at least $1 - p$ and the number $n(M, \rho, p)$ depends explicitly, besides ρ and p, also on the total volume and the dimension of M.

Evolutionary space

Just to recall: SILVA database: datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (Bacteria, Archaea and Eukarya).
Currently more than 9 mln rRNA sequences.

Structure of evolutionary space? #1

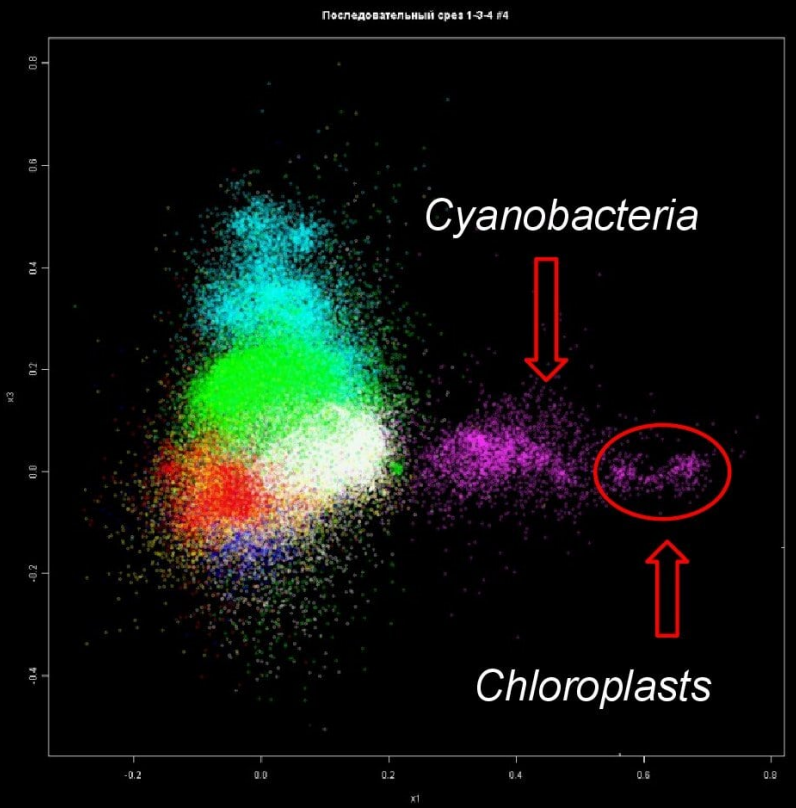


Structure of evolutionary space? #2

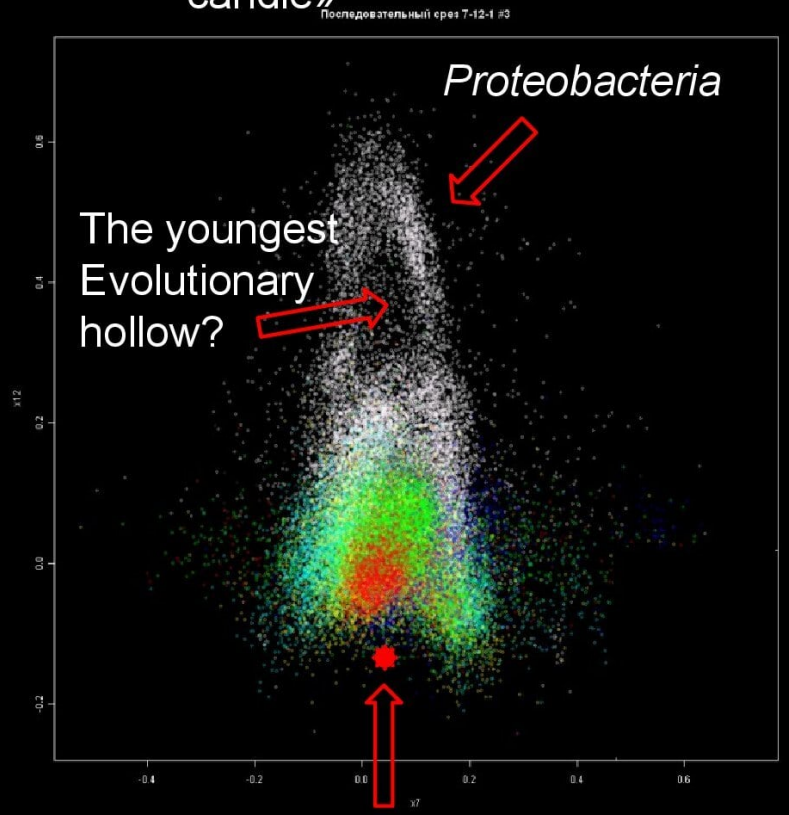
Sections, not projections !

13D

«The evolutionary gun»



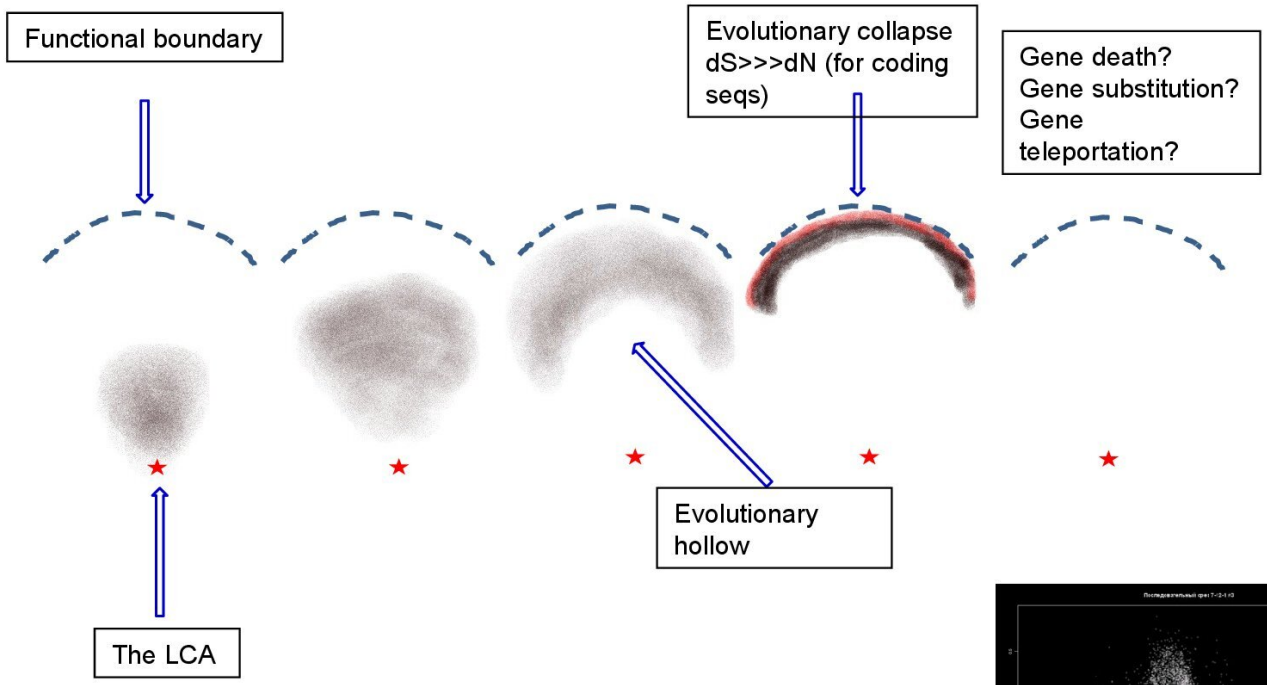
«The evolutionary candle»



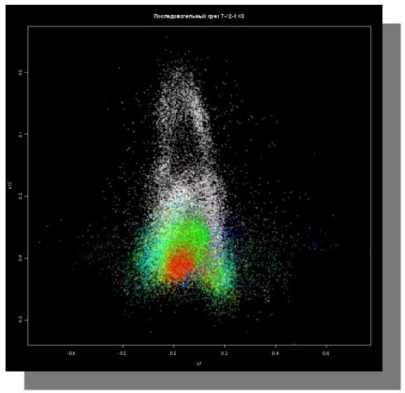
The last common ancestor of Bacteria?
The oldest Evolutionary hollow?

Structure of evolutionary space? #3

Evolutionary events on the late stages of evolution near the functional boundary of the gene



We can try to find these events in “rapid” genes, “rapid” taxa, “rapid” niches



Many thanks to...

Many thanks to
Eugene Andronov, ARRIAM, St. Petersburg,
Roberto Grossi, UniPi, Pisa,
Yuri Porozov, Sechenov Medical Univ., Moscow,
Nikita Puchkin, HSE, Moscow,
Vladimir Spokoyny, WIAS, Berlin,
Gaik Tamazian, Dobzhanky Lab., St. Petersburg,
Valentina Tozzini, NEST/SNS, Pisa,
Dario Trevisan, UniPi, Pisa