

Advances in multivariate statistics and its applications

Elena Tuzhilina

Stanford University, Department of Statistics

elenatuz@stanford.edu



Trevor Hastie

List of projects

① Conformation reconstruction



Department of
Epidemiology &
Biostatistics

Real-life
challenge



② Weighted low-rank matrix approximation

③ Structured canonical correlation analysis



Precision Psychiatry and Translational
Neuroscience Lab (PanLab)

Statistical
modeling



④ COVID-19 Forecasting

Carnegie Mellon University

DELPHI GROUP

Optimization



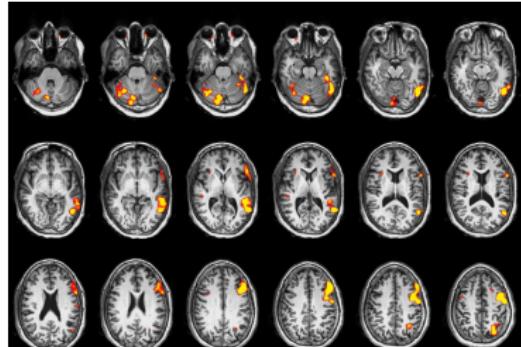
Efficient
implementation

Structured canonical correlation analysis

Goal: How emotional disorders influence the brain activity?

Why is it important? To understand better the depression and anxiety phenomenon and design the personalized treatment.

Why is it challenging? The brain data is extremely high-dimensional and has the structure imposed by the brain geometry.



Brain activation

References:

- ① "Canonical Correlation Analysis in high dimensions with structured regularization", **E.Tuzhilina, L.Tozzi, T.Hastie, Statistical Modelling, SAGE, 2021**
- ② "Relating whole-brain functional connectivity to self-reported negative emotion in a large sample of young adults using group regularized canonical correlation analysis", **L.Tozzi, E.Tuzhilina, M.Glasser, T.Hastie, L.Williams, NeuroImage, Vol. 237, pp. 118-137, 2021**

Weighted low-rank matrix approximation

Goal: Given some $M \in \mathbb{R}^{n \times p}$ and weights $W \in [0, 1]^{n \times p}$ identify the “best” way to approximate M with a rank- k matrix X

$$\text{minimize } \|\sqrt{W} * (M - X)\|_F^2 \text{ w.r.t. } X \in \mathbb{R}^{n \times p} \text{ subject to } \text{rk}(X) \leq k$$

Well-studied cases:

- Explicit solution when $W_{ij} = 1$.
Applications: PCA, CCA, LDA.
- Soft-impute when $W_{ij} \in \{0, 1\}$.
Applications: recommendation systems.

	3	1	5	1	?	1
	2	?	?	2	5	1
	1	3	1	4	2	5
	?	3	1	?	4	3
	2	2	1	3	1	4

Netflix Prize

Why is it important? LRMA is the core of many ML techniques: data compression, dimension reduction and de-noising.

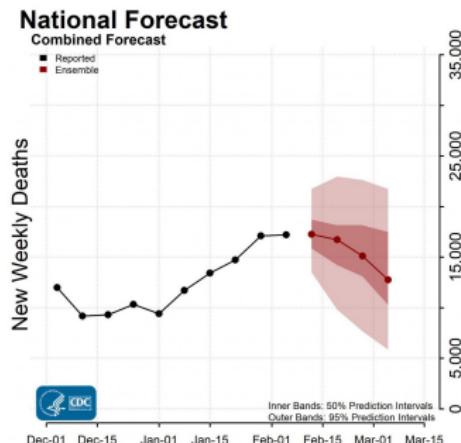
Reference: “Weighted Low Rank Matrix Approximation and acceleration”, E.Tuzhilina, T.Hastie, arXiv, 2021

COVID-19 forecasting

Goal: Predict the trajectory of the COVID-19 pandemic by forecasting multiple ahead values of the signal.

Why is it important? Helps inform public health decision making by projecting the likely impact of the COVID-19 pandemic.

Data: *COVIDCast API* containing cases, deaths, hospitalizations, as well as many unique indicators related to mobility, healthcare and survey.



COVID trajectory

Reference: "*Smooth multi-period forecasting with application to prediction of COVID-19 cases*", E.Tuzhilina, T.Hastie, R.Tibshirani, K.Tay, D.McDonald, arXiv, 2022

Conformation reconstruction

Plan

- ① Background and previous work
- ② Methodology and data application
- ③ Reconstruction validation
- ④ Extensions



Mark Segal



Trevor Hastie

References:

- ① “Principal curve approaches for inferring 3D chromatin architecture”, **E.Tuzhilina, T.Hastie, M.Segal**, *Biostatistics*, 2020
- ② “Statistical curve models for inferring 3D chromatin architecture”, **E.Tuzhilina, T.Hastie, M.Segal**, *bioRxiv*, 2022
- ③ R package PoisMS available at GitHub

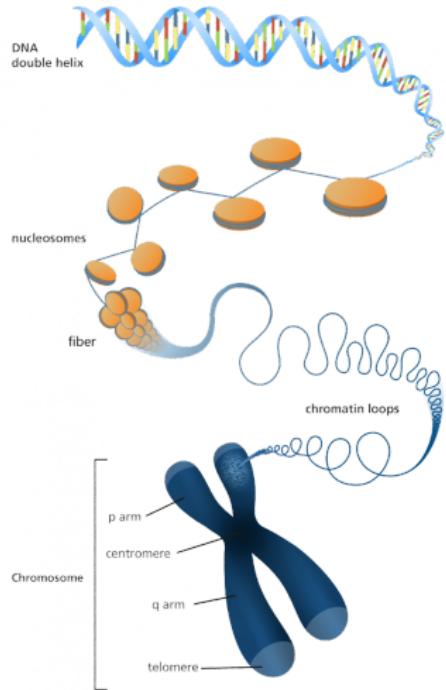
Conformation reconstruction

Plan

- ① Background and previous work
- ② Methodology and data application
- ③ Reconstruction validation
- ④ Extensions

What defines chromatin conformation?

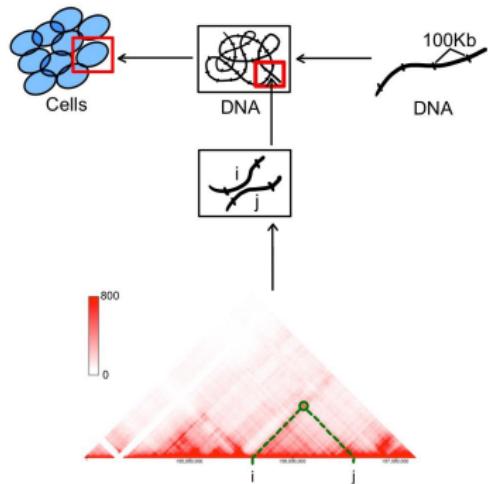
Chromatin



Chromatin = DNA + nucleosomes

- ① DNA
- ② 'Beads-on-a-string'
- ③ Chromatin fiber
- ④ Chromatin loop
- ⑤ Chromosome

Contact matrix



Terminology

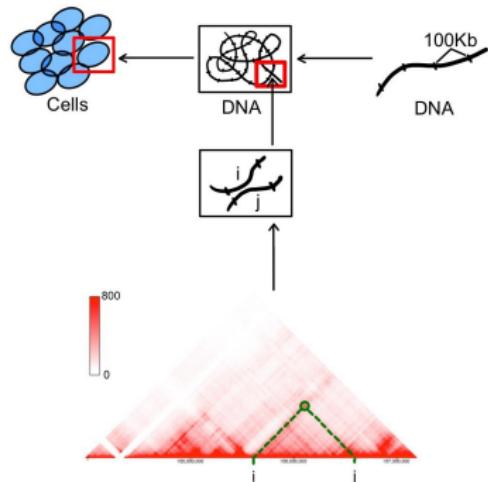
- genomic locus = 'piece'
- resolution = 'size of a piece'
- contact (formaldehyde + cross-linking + sequencing)

Notations

- $n = \#$ genomic loci
- $C_{ij} = \#$ contacts between loci i and j

Contact matrix $C = [C_{ij}] \in \mathbb{Z}_+^{n \times n}$

Contact matrix



Terminology

- genomic locus = 'piece'
- resolution = 'size of a piece'
- contact (formaldehyde + cross-linking + sequencing)

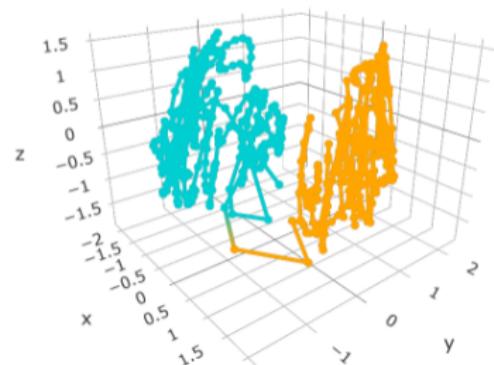
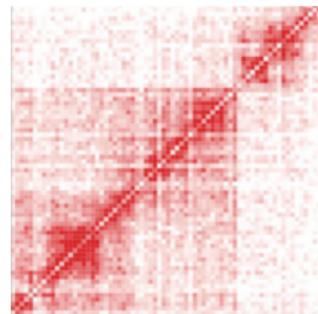
Notations

- $n = \#$ genomic loci
- $C_{ij} = \#$ contacts between loci i and j

Contact matrix $C = [C_{ij}] \in \mathbb{Z}_+^{n \times n}$

Chromatin reconstruction problem

Goal: Use the information contained in C to reconstruct the locus spatial coordinates $x_1, \dots, x_n \in \mathbb{R}^3$.



Previous work

Main ingredients

- loss function $\ell(x_1, \dots, x_n)$
- optimization problem minimizing $\ell(x_1, \dots, x_n)$ w.r.t. $x_1, \dots, x_n \in \mathbb{R}^3$
- iterative optimization algorithm

Previous work

Main ingredients

- loss function $\ell(x_1, \dots, x_n)$
- optimization problem minimizing $\ell(x_1, \dots, x_n)$ w.r.t. $x_1, \dots, x_n \in \mathbb{R}^3$
- iterative optimization algorithm

Example (deterministic model, metric MDS)

- ① Convert C to a distance matrix D , e.g. $D_{ij} = \begin{cases} (C_{ij})^{-\alpha} & \text{if } C_{ij} > 0 \\ \infty & \text{if } C_{ij} = 0 \end{cases}$
- ② Minimize Stress objective

$$\ell(x_1, \dots, x_n) = \sum_{i=1}^n \sum_{j=1}^n (D_{ij} - \|x_i - x_j\|)^2$$

w.r.t. $x_1, \dots, x_n \in \mathbb{R}^3$

Previous work

Main ingredients

- loss function $\ell(x_1, \dots, x_n)$
- optimization problem minimizing $\ell(x_1, \dots, x_n)$ w.r.t. $x_1, \dots, x_n \in \mathbb{R}^3$
- iterative optimization algorithm

Example (distribution-based model, Poisson)

- ① $C_{ij} \sim \text{Pois}(\lambda_{ij})$, where $\lambda_{ij} = \lambda_{ij}(x_1, \dots, x_n) = \beta \|x_i - x_j\|^\alpha$
- ② Minimize negative log-likelihood

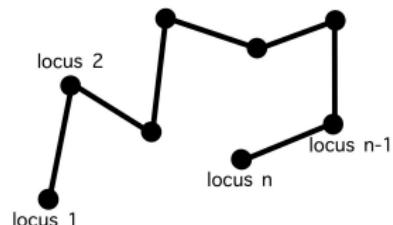
$$\ell(x_1, \dots, x_n) = \sum_{i=1}^n \sum_{j=1}^n \beta \|x_i - x_j\|^\alpha - C_{ij} \log (\beta \|x_i - x_j\|^\alpha)$$

w.r.t. $x_1, \dots, x_n \in \mathbb{R}^3$

Controlling reconstruction smoothness

Previous approaches: model chromatin by
a polygonal chain

$$\begin{aligned} & \text{minimize } \ell(x_1, \dots, x_n) + \lambda h(x_1, \dots, x_n) \\ & \text{w.r.t. } x_1, \dots, x_n \in \mathbb{R}^3 \end{aligned}$$

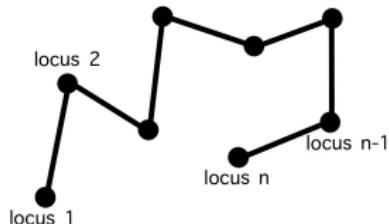


Controlling reconstruction smoothness

Previous approaches: model chromatin by
a polygonal chain

$$\text{minimize } \ell(x_1, \dots, x_n) + \lambda h(x_1, \dots, x_n)$$

w.r.t. $x_1, \dots, x_n \in \mathbb{R}^3$



Example (two penalties)

$$h_1(x_1, \dots, x_n) = (n - 1) \frac{\sum_{i=1}^{n-1} \|x_{i+1} - x_i\|^2}{\left(\sum_{i=1}^{n-1} \|x_{i+1} - x_i\|\right)^2} - 1$$

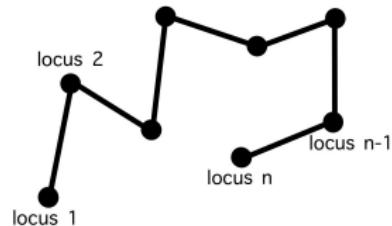


$$h_2(x_1, \dots, x_n) = \frac{1}{n - 2} \sum_{i=2}^{n-1} \frac{\langle x_{i-1} - x_i, x_{i+1} - x_i \rangle}{\|x_{i-1} - x_i\| \|x_{i+1} - x_i\|}$$

Controlling reconstruction smoothness

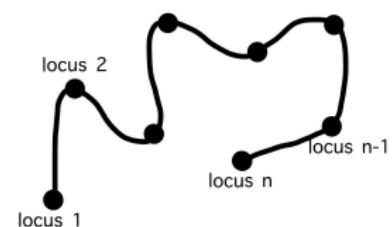
Previous approaches: model chromatin by
a polygonal chain

$$\begin{aligned} & \text{minimize } \ell(x_1, \dots, x_n) + \lambda h(x_1, \dots, x_n) \\ & \text{w.r.t. } x_1, \dots, x_n \in \mathbb{R}^3 \end{aligned}$$



Our approach: model chromatin by
a smooth curve

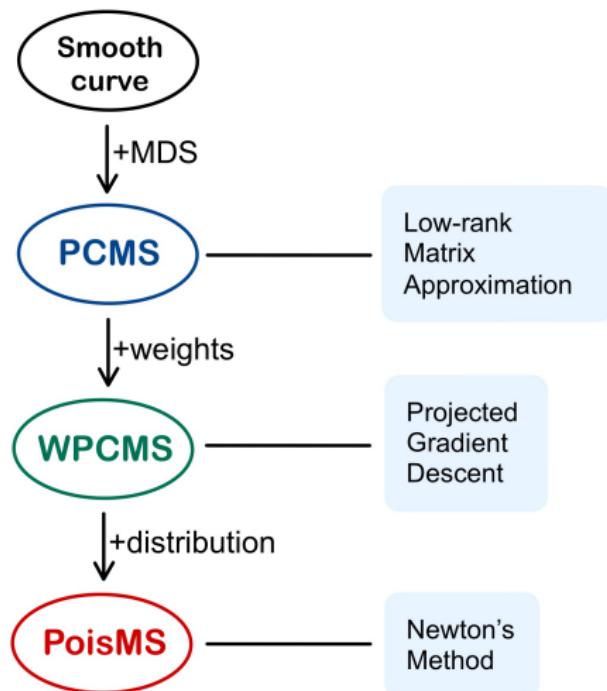
$$\begin{aligned} & \text{minimize } \ell(x_1, \dots, x_n) \text{ w.r.t. } x_1, \dots, x_n \in \mathbb{R}^3 \\ & x_1, \dots, x_n \in \text{smooth one-dimensional curve} \end{aligned}$$



Conformation reconstruction

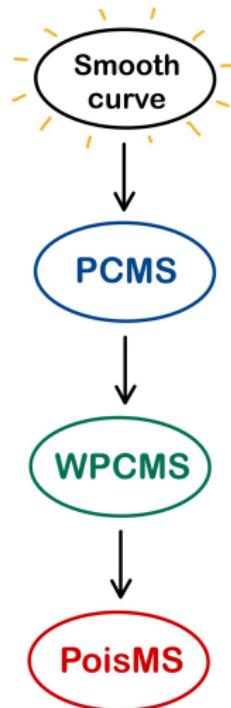
Plan

- ① Background and previous work
- ② Methodology and data application
- ③ Tuning and validation
- ④ Extensions



Smooth curve constraint

Idea: use cubic splines



Smooth curve constraint

① $x_1, \dots, x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$

② $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^k \Theta_{\ell j} h_\ell(t)$

③ $x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^k \Theta_{\ell 1} h_\ell(i) \\ \sum_{\ell=1}^k \Theta_{\ell 2} h_\ell(i) \\ \sum_{\ell=1}^k \Theta_{\ell 3} h_\ell(i) \end{pmatrix}$

Smooth curve constraint

- ① $x_1, \dots, x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
- ② $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^k \Theta_{\ell j} h_\ell(t)$

hyperparameter k = spline degrees-of-freedom (df)

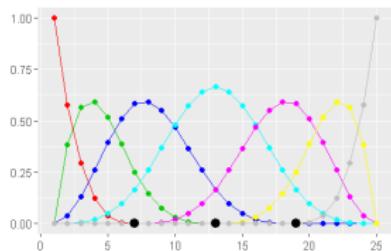
- ③ $x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^k \Theta_{\ell 1} h_\ell(i) \\ \sum_{\ell=1}^k \Theta_{\ell 2} h_\ell(i) \\ \sum_{\ell=1}^k \Theta_{\ell 3} h_\ell(i) \end{pmatrix}$

Smooth curve constraint

- ① $x_1, \dots, x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
- ② $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^k \Theta_{\ell j} h_\ell(t)$
- ③ $x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^k \Theta_{\ell 1} h_\ell(i) \\ \sum_{\ell=1}^k \Theta_{\ell 2} h_\ell(i) \\ \sum_{\ell=1}^k \Theta_{\ell 3} h_\ell(i) \end{pmatrix}$

Smooth curve constraint

- ① $x_1, \dots, x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
- ② $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^k \Theta_{\ell j} h_\ell(t)$
- ③ $x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^k \Theta_{\ell 1} h_\ell(i) \\ \sum_{\ell=1}^k \Theta_{\ell 2} h_\ell(i) \\ \sum_{\ell=1}^k \Theta_{\ell 3} h_\ell(i) \end{pmatrix}$



$$\longrightarrow H = \begin{pmatrix} | & & | \\ h_1 & \dots & h_k \\ | & & | \end{pmatrix} \in \mathbb{R}^{n \times k}$$

Smooth curve constraint

- ① $x_1, \dots, x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
- ② $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^k \Theta_{\ell j} h_\ell(t)$

$$\textcircled{3} \quad x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^k \Theta_{\ell 1} h_\ell(i) \\ \sum_{\ell=1}^k \Theta_{\ell 2} h_\ell(i) \\ \sum_{\ell=1}^k \Theta_{\ell 3} h_\ell(i) \end{pmatrix}$$

$$X = \begin{pmatrix} -x_1^T - \\ \dots \\ -x_n^T - \end{pmatrix} = \begin{pmatrix} | & | & | \\ \gamma_1 & \gamma_2 & \gamma_3 \\ | & | & | \end{pmatrix} \in \mathbb{R}^{n \times 3} \quad H = \begin{pmatrix} | & & | \\ h_1 & \dots & h_k \\ | & & | \end{pmatrix} \in \mathbb{R}^{n \times k}$$

Smooth curve constraint

- ① $x_1, \dots, x_n \in \gamma(t)$, where $\gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$
- ② $h_1(t), \dots, h_k(t)$ – cubic spline basis functions, $\gamma_j(t) = \sum_{\ell=1}^k \Theta_{\ell j} h_\ell(t)$

$$\textcircled{3} \quad x_i = \gamma(i) = \begin{pmatrix} \sum_{\ell=1}^k \Theta_{\ell 1} h_\ell(i) \\ \sum_{\ell=1}^k \Theta_{\ell 2} h_\ell(i) \\ \sum_{\ell=1}^k \Theta_{\ell 3} h_\ell(i) \end{pmatrix}$$

$$X = \begin{pmatrix} -x_1^T - \\ \dots \\ -x_n^T - \end{pmatrix} = \begin{pmatrix} | & | & | \\ \gamma_1 & \gamma_2 & \gamma_3 \\ | & | & | \end{pmatrix} \in \mathbb{R}^{n \times 3} \quad H = \begin{pmatrix} | & & | \\ h_1 & \dots & h_k \\ | & & | \end{pmatrix} \in \mathbb{R}^{n \times k}$$

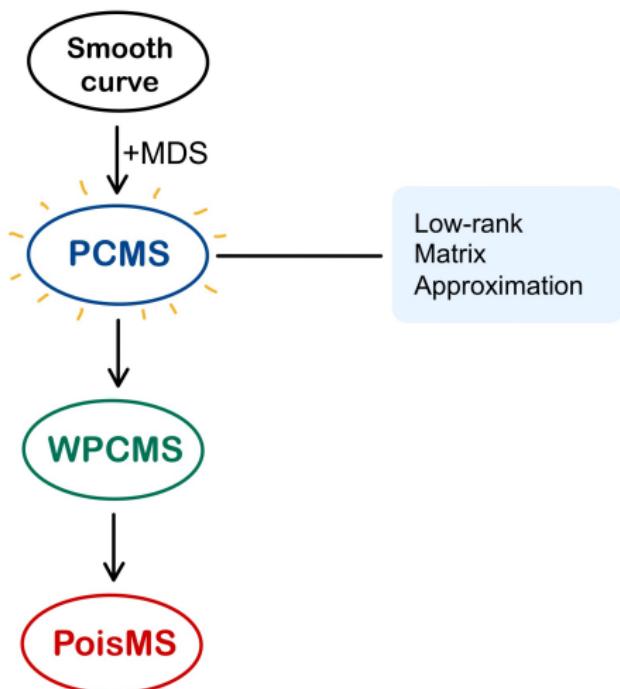
$x_1, \dots, x_n \in \text{smooth one-dimensional curve} \iff$

$\exists \Theta \in \mathbb{R}^{k \times 3}$ such that $X = H\Theta$

Principal Curve Metric Scaling

$$X = H\Theta$$

Idea: smooth curve constraint implies low-rank structure



Principal Curve Metric Scaling

PCMS = Classical MDS + Smooth curve constraint

Classical MDS

$$\ell(x_1, \dots, x_n) = \sum_{i=1}^n \sum_{j=1}^n (Z_{ij} - \langle x_i, x_j \rangle)^2 \iff \ell(X) = \|Z - S(X)\|_F^2$$

Smooth curve constraint $X = H\Theta$

PCMS optimization problem

$$\text{minimize } \ell_{PCMS}(\Theta) = \|Z - S(H\Theta)\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k \times 3}$$

Solution via Eigen Decomposition of $H^T Z H \in \mathbb{R}^{k \times k}$

Principal Curve Metric Scaling

PCMS = Classical MDS + Smooth curve constraint

Classical MDS

$$\ell(x_1, \dots, x_n) = \sum_{i=1}^n \sum_{j=1}^n (Z_{ij} - \langle x_i, x_j \rangle)^2 \iff \ell(X) = \|Z - S(X)\|_F^2$$

Smooth curve constraint $X = H\Theta$

PCMS optimization problem

$$\text{minimize } \ell_{PCMS}(\Theta) = \|Z - S(H\Theta)\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k \times 3}$$

Solution via Eigen Decomposition of $H^T Z H \in \mathbb{R}^{k \times k}$

PCMS examples

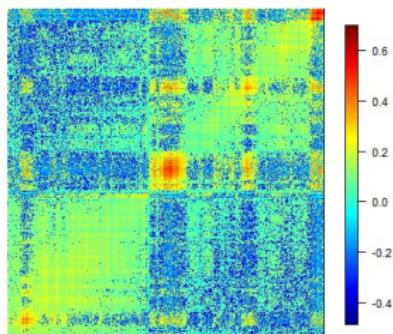
Data Hi-C data for IMR90 cells from the Gene Expression Omnibus.
Chromosome 20, probe resolution 100kb, $n = 625$.

Transformation $Z = J \left(-\frac{\mathcal{D}^2}{2} \right) J$ where $\mathcal{D} = \frac{1}{c+1}$ and $J = I - \frac{\mathbf{1}\mathbf{1}^T}{n}$

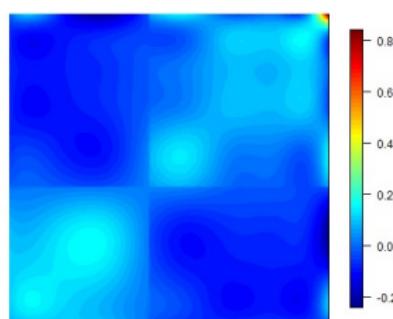
PCMS examples ($df = 10$)

Data Hi-C data for IMR90 cells from the Gene Expression Omnibus.
Chromosome 20, probe resolution 100kb, $n = 625$.

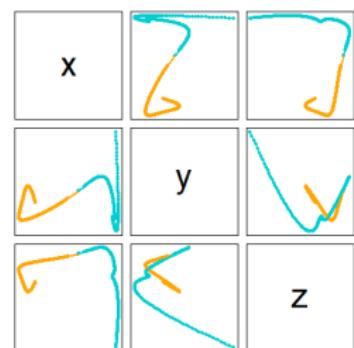
Transformation $Z = J \left(-\frac{\mathcal{D}^2}{2} \right) J$ where $\mathcal{D} = \frac{1}{c+1}$ and $J = I - \frac{11^T}{n}$



Transformed contact
matrix Z



Approximation $S(X)$

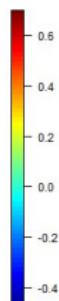
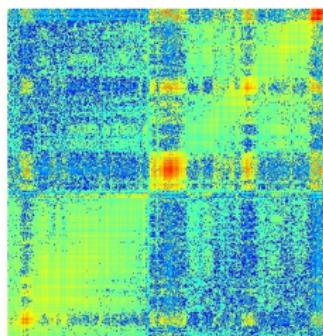


3D conformation X

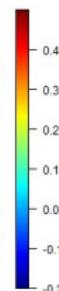
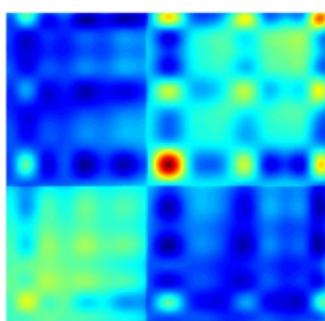
PCMS examples ($df = 20$)

Data Hi-C data for IMR90 cells from the Gene Expression Omnibus.
Chromosome 20, probe resolution 100kb, $n = 625$.

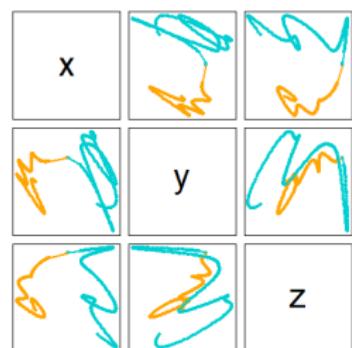
Transformation $Z = J \left(-\frac{\mathcal{D}^2}{2} \right) J$ where $\mathcal{D} = \frac{1}{c+1}$ and $J = I - \frac{11^T}{n}$



Transformed contact
matrix Z



Approximation $S(X)$

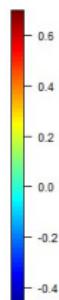
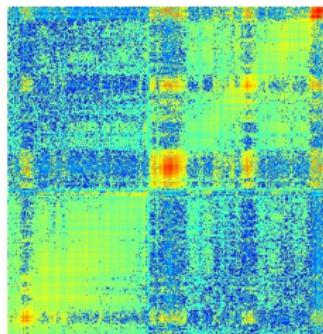


3D conformation X

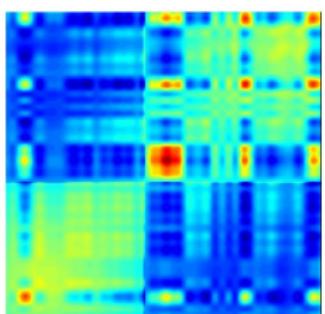
PCMS examples ($df = 50$)

Data Hi-C data for IMR90 cells from the Gene Expression Omnibus.
Chromosome 20, probe resolution 100kb, $n = 625$.

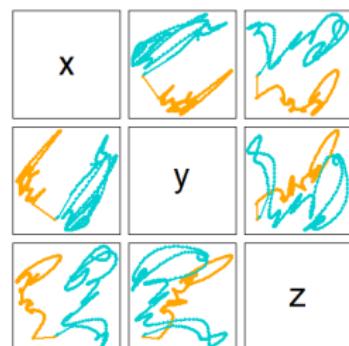
Transformation $Z = J \left(-\frac{\mathcal{D}^2}{2} \right) J$ where $\mathcal{D} = \frac{1}{c+1}$ and $J = I - \frac{11^T}{n}$



Transformed contact
matrix Z



Approximation $S(X)$

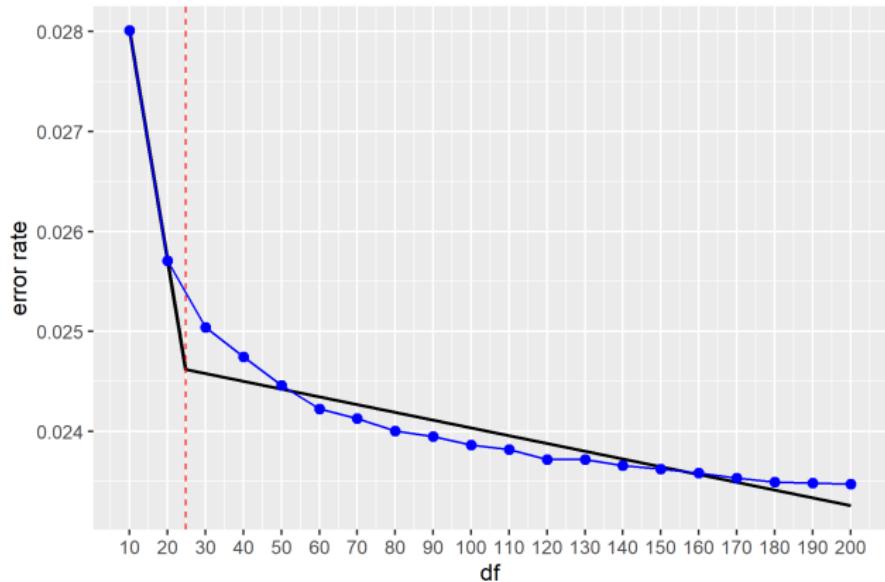


3D conformation X

Select degrees-of-freedom



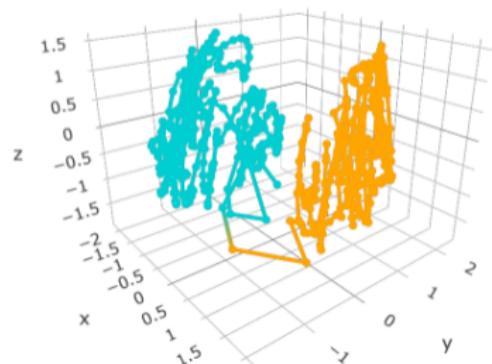
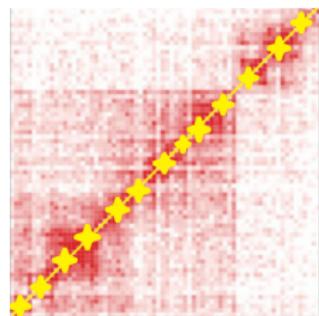
Prof. Greenacre
performs elbow
detection.



$$err(X) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (Z_{ij} - \langle x_i, x_j \rangle)^2$$

Artifacts of the contact matrix

Problem: contact counts are not equally informative for the chromatin reconstruction. Contact matrix is *sparse and diagonal dominant*.



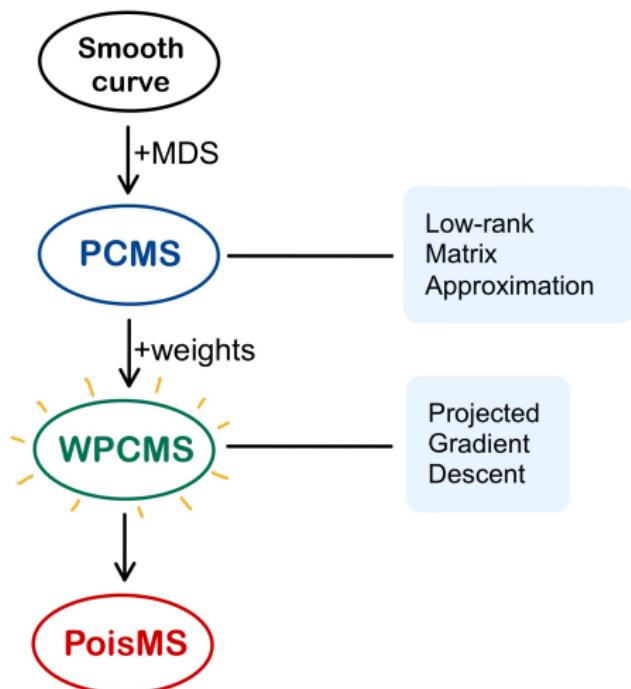
Weighted Principal Curve Metric Scaling

$$X = H\Theta$$

$$\|Z - S(X)\|_F^2$$

$S(X)$ – inner products

Idea: use PCMS for the projection step



Weighted Principal Curve Metric Scaling

WPCMS = PCMS + Weights

Motivation control the impact of some elements

Weighted Strain

$$\ell(x_1, \dots, x_n) = \sum_{i=1}^n \sum_{j=1}^n W_{ij} (Z_{ij} - \langle x_i, x_j \rangle)^2 \iff \ell(X) = \|\sqrt{W} * (Z - S(X))\|_F^2$$

Weighted Principal Curve Metric Scaling

WPCMS = PCMS + Weights + Distances

Motivation escape from double centering

Weighted Stress

$$\ell(x_1, \dots, x_n) = \sum_{i=1}^n \sum_{j=1}^n W_{ij} (Z_{ij} - \|x_i - x_j\|^2)^2 \iff \ell(X) = \|\sqrt{W} * (Z - D^2(X))\|_F^2$$

Weighted Principal Curve Metric Scaling

WPCMS = PCMS + Weights + Distances

Motivation escape from double centering

Weighted Stress

$$\ell(x_1, \dots, x_n) = \sum_{i=1}^n \sum_{j=1}^n W_{ij} (Z_{ij} - \|x_i - x_j\|^2)^2 \iff \ell(X) = \|\sqrt{W} * (Z - D^2(X))\|_F^2$$

Weighted Principal Curve Metric Scaling

$$\text{WPCMS} = \text{PCMS} + \text{Weights} + \text{Distances}$$

Motivation escape from double centering

Weighted Stress

$$\ell(x_1, \dots, x_n) = \sum_{i=1}^n \sum_{j=1}^n \textcolor{green}{W}_{ij} (Z_{ij} - \|x_i - x_j\|^2)^2 \iff \ell(X) = \|\sqrt{\textcolor{green}{W}} * (Z - \textcolor{green}{D}^2(X))\|_F^2$$

Smooth curve constraint $X = H\Theta$

WPCMS optimization problem

$$\text{minimize } \ell_{WPCMS}(\Theta) = \|\sqrt{\textcolor{green}{W}} * (Z - \textcolor{green}{D}^2(H\Theta))\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k \times 3}$$

Solution use PCMS as a building block for the iterative algorithm

Projected gradient descent

minimize $\ell(x)$ w.r.t. $x \in \mathcal{M}$

[Gradient] $x := x - \nabla_x \ell(x)$ **[Projection]** $x := \text{proj}_{\mathcal{M}}(x)$

Idea: PCMS(S) vs WPCMS(D^2) \Rightarrow link distances and inner products

$$D^2 = \text{diag}(S) \cdot \mathbf{1}^T + \mathbf{1}^T \cdot \text{diag}(S) - 2S.$$

WPCMS problem (inner products view)

$$\text{minimize } \ell_{WPCMS}(S) = \left\| \sqrt{W} * \left(Z - \text{diag}(S)\mathbf{1}^T - \mathbf{1}^T \text{diag}(S) + 2S \right) \right\|_F^2$$

$$\text{w.r.t. } S \in \mathcal{M} = \left\{ S(H\Theta) : \Theta \in \mathbb{R}^{k \times 3} \right\}$$

Projected gradient descent

minimize $\ell(x)$ w.r.t. $x \in \mathcal{M}$

[Gradient] $x := x - \nabla_x \ell(x)$ [Projection] $x := \text{proj}_{\mathcal{M}}(x)$

Idea: PCMS(S) vs WPCMS(D^2) \implies link distances and inner products

$$D^2 = \text{diag}(S) \cdot \mathbf{1}^T + \mathbf{1}^T \cdot \text{diag}(S) - 2S.$$

WPCMS problem (inner products view)

$$\text{minimize } \ell_{WPCMS}(S) = \left\| \sqrt{W} * \left(Z - \text{diag}(S)\mathbf{1}^T - \mathbf{1}^T \text{diag}(S) + 2S \right) \right\|_F^2$$

$$\text{w.r.t. } S \in \mathcal{M} = \left\{ S(H\Theta) : \Theta \in \mathbb{R}^{k \times 3} \right\}$$

Projected gradient descent

WPCMS problem (inner products view)

$$\text{minimize } \ell_{WPCMS}(S) \text{ w.r.t. } S \in \mathcal{M} = \left\{ S(H\Theta) : \Theta \in \mathbb{R}^{k \times 3} \right\}$$

PGD

- ➊ [Gradient] $S := S - \nabla \ell_{WPCMS}(S)$

$$\nabla \ell_{WPCMS}(S) = G - G_+ \text{ with } G = W * (Z - D^2) \text{ and } G_+ = \text{diag}(G \cdot \mathbf{1})$$

- ➋ [Projection] $S := \text{proj}_{\mathcal{M}}(S)$

$$\text{minimize } \|S - S(H\Theta)\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k \times 3}$$

Projected gradient descent

WPCMS problem (inner products view)

minimize $\ell_{WPCMS}(S)$ w.r.t. $S \in \mathcal{M} = \left\{ S(H\Theta) : \Theta \in \mathbb{R}^{k \times 3} \right\}$

PGD

- ➊ [Gradient] $S := S - \nabla \ell_{WPCMS}(S)$

$$\nabla \ell_{WPCMS}(S) = G - G_+ \text{ with } G = W * (Z - D^2) \text{ and } G_+ = \text{diag}(G \cdot \mathbf{1})$$

- ➋ [Projection] $S := \text{proj}_{\mathcal{M}}(S)$

$$\text{minimize } \|S - S(H\Theta)\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k \times 3}$$

Projected gradient descent

WPCMS problem (inner products view)

$$\text{minimize } \ell_{WPCMS}(S) \text{ w.r.t. } S \in \mathcal{M} = \left\{ S(H\Theta) : \Theta \in \mathbb{R}^{k \times 3} \right\}$$

PGD

- ① [Gradient] $S := S - \nabla \ell_{WPCMS}(S)$

$$\nabla \ell_{WPCMS}(S) = G - G_+ \text{ with } G = W * (Z - D^2) \text{ and } G_+ = \text{diag}(G \cdot \mathbf{1})$$

- ② [Projection] $S := \text{proj}_{\mathcal{M}}(S)$

$$\text{minimize } \|S - S(H\Theta)\|_F^2 \text{ w.r.t. } \Theta \in \mathbb{R}^{k \times 3}$$

WPCMS iterative algorithm

WPCMS problem (inner products view)

$$\text{minimize } \ell_{WPCMS}(S) \text{ w.r.t. } S \in \mathcal{M}(H) = \left\{ S(H\Theta) : \Theta \in \mathbb{R}^{k \times 3} \right\}$$

- ① **[Initialize]** Generate random $\Theta \in \mathbb{R}^{k \times 3}$, set $X := H\Theta$
- ② *Repeat until convergence:*
 - **[Gradient]** $S := S - (G - G_+)$ where $G = W * (Z - D^2(X))$
 - **[Projection]** $\Theta := \text{PCMS}(S, H)$ and $X := H\Theta$

Poisson Metric Scaling

$$X = H\Theta$$

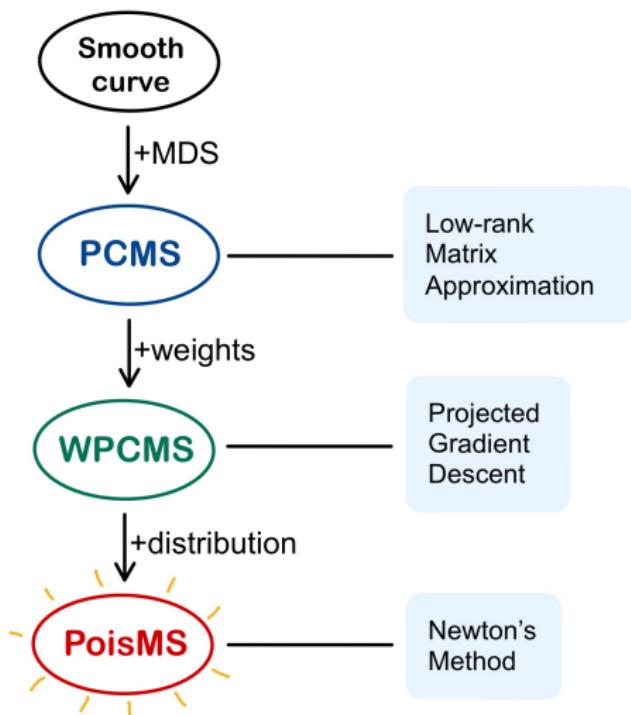
$$\|Z - S(X)\|_F^2$$

$S(X)$ – inner products

$$\|\sqrt{W} * (Z - D^2(X))\|_F^2$$

$D(X)$ – distances

Idea: use WPCMS to optimize the second order approximation of the loss



Poisson Metric Scaling

PoisMS = WPCMS + Poisson Model

Model $C_{ij} \sim Pois(\lambda_{ij})$, $\log(\lambda_{ij}) = -\|x_i - x_j\|^2 + \beta$

Negative log-likelihood

$$\ell_{PoisMS}(X, \beta) = \sum_{i=1}^n \sum_{j=1}^n \left[e^{-\|x_i - x_j\|^2 + \beta} - C_{ij} (-\|x_i - x_j\|^2 + \beta) \right]$$

Smooth curve constraint $X = H\Theta$

PoisMS optimization problem

minimize $\ell_{PoisMS}(H\Theta, \beta)$ w.r.t. $\Theta \in \mathbb{R}^{k \times 3}$ and $\beta \in \mathbb{R}$

Solution use WPCMS as a building block for the iterative algorithm

Poisson Metric Scaling

PoisMS = WPCMS + Poisson Model

Model $C_{ij} \sim Pois(\lambda_{ij})$, $\log(\lambda_{ij}) = -\|x_i - x_j\|^2 + \beta$

Negative log-likelihood

$$\ell_{PoisMS}(X, \beta) = \sum_{i=1}^n \sum_{j=1}^n \left[e^{-\|x_i - x_j\|^2 + \beta} - C_{ij} (-\|x_i - x_j\|^2 + \beta) \right]$$

Smooth curve constraint $X = H\Theta$

PoisMS optimization problem

minimize $\ell_{PoisMS}(H\Theta, \beta)$ w.r.t. $\Theta \in \mathbb{R}^{k \times 3}$ and $\beta \in \mathbb{R}$

Solution use WPCMS as a building block for the iterative algorithm

Newton's method for GLM

minimize $\ell(\eta)$ where $\eta = X\beta$

[SOA] $\ell(\eta) \approx \ell_{SOA}(\eta) = (z - \eta)^T W(z - \eta)$ [WLS] $z \sim X$

$$\ell_{PoisMS}(X, \beta) \approx \ell_{SOA}(X) = \left\| \sqrt{W} * (Z - D^2(X)) \right\|_F^2$$

$$\text{where } W = e^{-D^2(X_0) + \beta_0} \text{ and } Z = D^2(X_0) - \frac{C - W}{W}.$$

⇒ use WPCMS to minimize SOA

Newton's method for GLM

minimize $\ell(\eta)$ where $\eta = X\beta$

[SOA] $\ell(\eta) \approx \ell_{SOA}(\eta) = (z - \eta)^T W(z - \eta)$ [WLS] $z \sim X$

$$\ell_{PoisMS}(X, \beta) \approx \ell_{SOA}(X) = \left\| \sqrt{W} * (Z - D^2(X)) \right\|_F^2$$

where $W = e^{-D^2(X_0) + \beta_0}$ and $Z = D^2(X_0) - \frac{C-W}{W}$.

⇒ use WPCMS to minimize SOA

PoisMS iterative algorithm

$$\ell_{PoisMS}(X, \beta) \approx \ell_{SOA}(X) = \left\| \sqrt{W} * (Z - D^2(X)) \right\|_F^2$$

where $W = e^{-D^2(X_0) + \beta_0}$ and $Z = D^2(X_0) - \frac{C-W}{W}$.

⇒ use WPCMS to minimize SOA

① **[Initialize]** Generate random $\Theta \in \mathbb{R}^{k \times 3}$, set $X := H\Theta$

② *Repeat until convergence:*

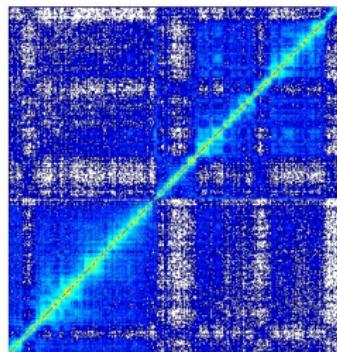
- **[SOA]** $W = e^{-D^2(X) + \beta}$ and $Z = D^2(X) - \frac{C-W}{W}$.

- **[WPCMS]** $\Theta := PCMS_W(Z, H)$ and $X := H\Theta$

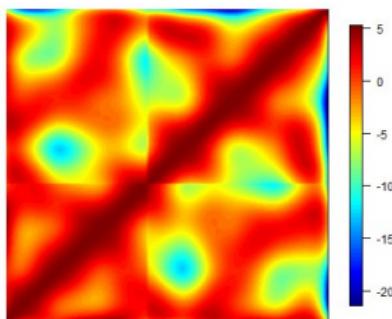
- **[Nuisance]** $\beta := \log \left(\sum_{1 \leq i, j \leq n} C_{ij} / \sum_{1 \leq i, j \leq n} e^{-\|x_i - x_j\|^2} \right)$.

PoisMS examples ($df = 10$)

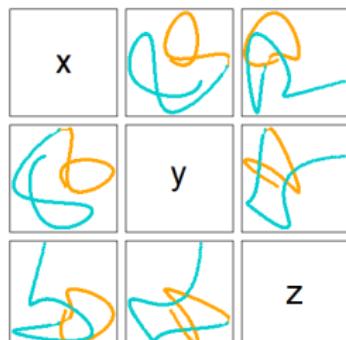
Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus.
Chromosome 20, probe resolution 100kb, $n = 625$.



Log-contact matrix
 $\log(C)$



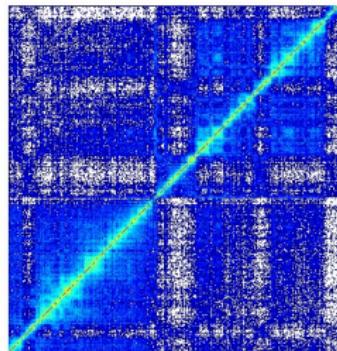
$$\text{Approximation} \\ \log(\Lambda) = -D^2(X) + \beta$$



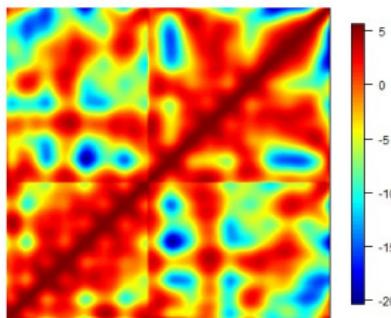
3D conformation X

PoisMS examples ($df = 20$)

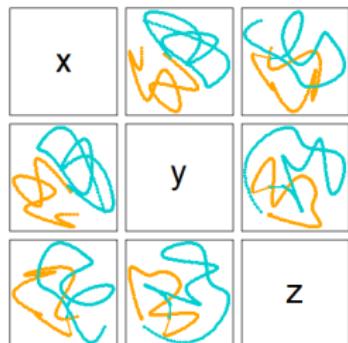
Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus.
Chromosome 20, probe resolution 100kb, $n = 625$.



Log-contact matrix
 $\log(C)$



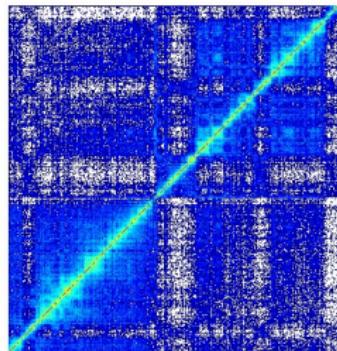
$$\text{Approximation} \\ \log(\Lambda) = -D^2(X) + \beta$$



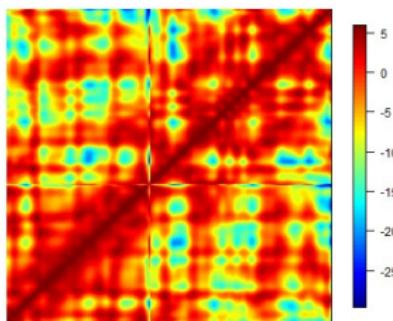
3D conformation X

PoisMS examples ($df = 50$)

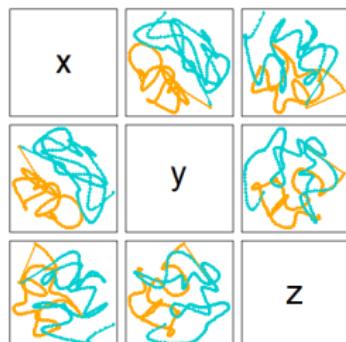
Data: Hi-C data for IMR90 cells from the Gene Expression Omnibus.
Chromosome 20, probe resolution 100kb, $n = 625$.



Log-contact matrix
 $\log(C)$

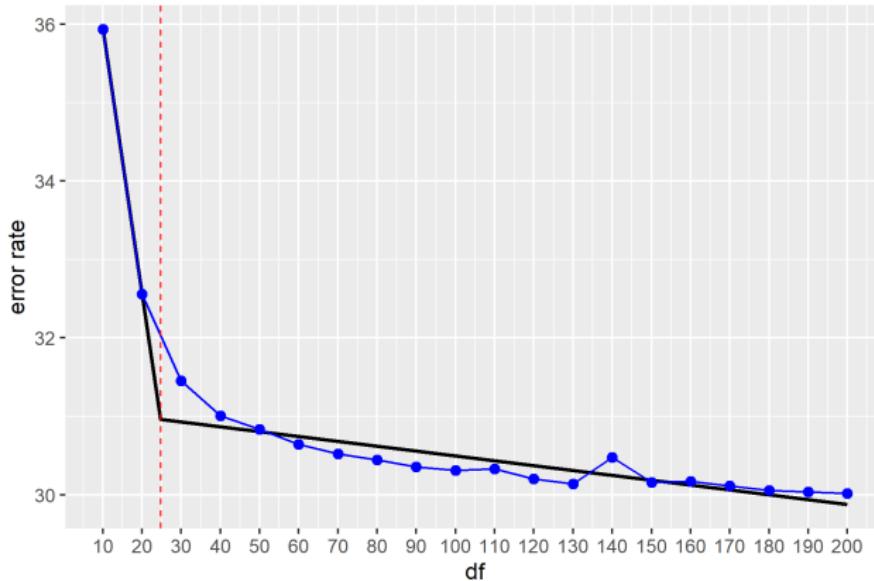


$$\text{Approximation} \\ \log(\Lambda) = -D^2(X) + \beta$$



3D conformation X

Select degrees-of-freedom



$$err(X, \beta) = \frac{2}{n^2} \sum_{j=1}^n \sum_{i=1}^n \left[C_{ij} \log \frac{C_{ij}}{\lambda_{ij}} - (C_{ij} - \lambda_{ij}) \right]$$

where $\log(\lambda_{ij}) = -\|x_i - x_j\|^2 + \beta$

Conformation reconstruction

Plan

- ① Background and previous work
- ② Methodology and data application
- ③ Reconstruction validation
- ④ Extensions

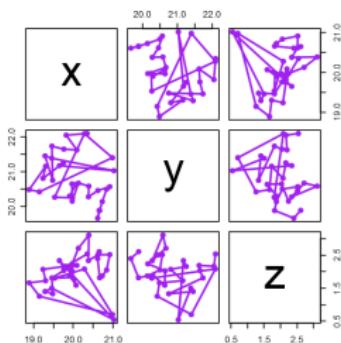
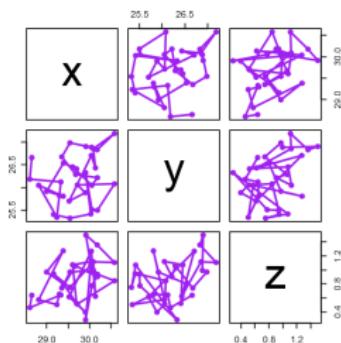
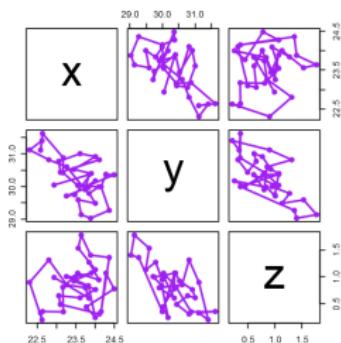
How realistic is the reconstruction?

Reconstruction validation

multiplex FISH = multiplex fluorescence in situ hybridization

- low resolution ($n_0 \approx 30$ genomic loci)
- many replicates ($N > 100$)

Example (three replicates)



Reconstruction validation

multiplex FISH = multiplex fluorescence in situ hybridization

- low resolution ($n_0 \approx 30$ genomic loci)
- many replicates ($N > 100$)

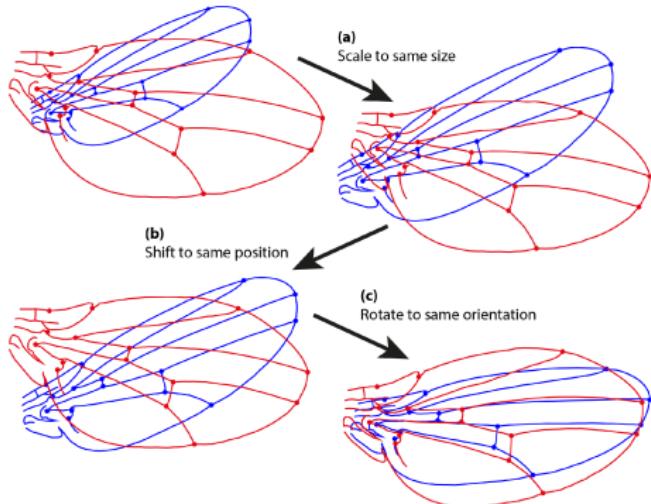
Validating the reconstruction

- ① Construct a gold standard
- ② Compute the reference distribution
- ③ Position the reconstructions

Gold standard

Notations

- replicate $M_i \in \mathbb{R}^{n_0 \times 3}$
- Procrustes distance $\rho(M_i, M_j)$



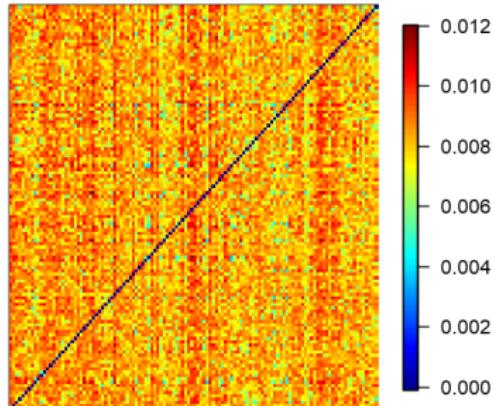
Steps

- ① Find the medoid replicate $i^* = \operatorname{argmin}_{i=1,\dots,N} \sum_{j=1}^N \rho(M_j, M_i)$
- ② Align replicates with the medoid $M_j \rightarrow M_j^{\text{rot}}$
- ③ Calculate gold standard \bar{M} as average of M_j^{rot}

Gold standard

Notations

- replicate $M_i \in \mathbb{R}^{n_0 \times 3}$
- Procrustes distance $\rho(M_i, M_j)$



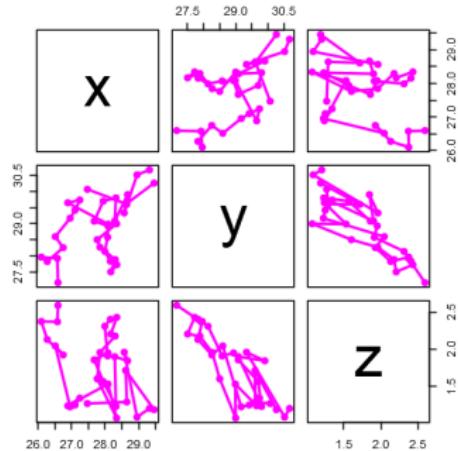
Steps

- ① Find the medoid replicate $i^* = \operatorname{argmin}_{i=1,\dots,N} \sum_{j=1}^N \rho(M_j, M_i)$
- ② Align replicates with the medoid $M_j \rightarrow M_j^{\text{rot}}$
- ③ Calculate gold standard \bar{M} as average of M_j^{rot}

Gold standard

Notations

- replicate $M_i \in \mathbb{R}^{n_0 \times 3}$
- Procrustes distance $\rho(M_i, M_j)$



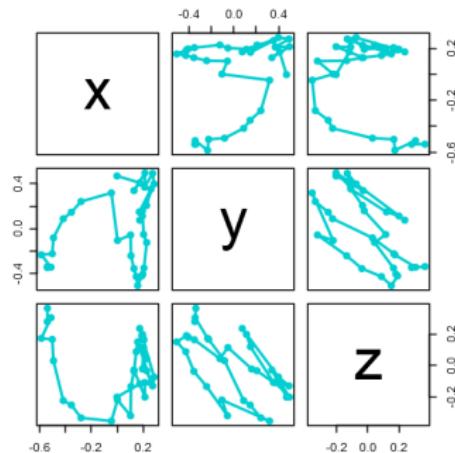
Steps

- ① Find the medoid replicate $i^* = \operatorname{argmin}_{i=1,\dots,N} \sum_{j=1}^N \rho(M_j, M_i)$
- ② Align replicates with the medoid $M_j \rightarrow M_j^{\text{rot}}$
- ③ Calculate gold standard \bar{M} as average of M_j^{rot}

Gold standard

Notations

- replicate $M_i \in \mathbb{R}^{n_0 \times 3}$
- Procrustes distance $\rho(M_i, M_j)$



Steps

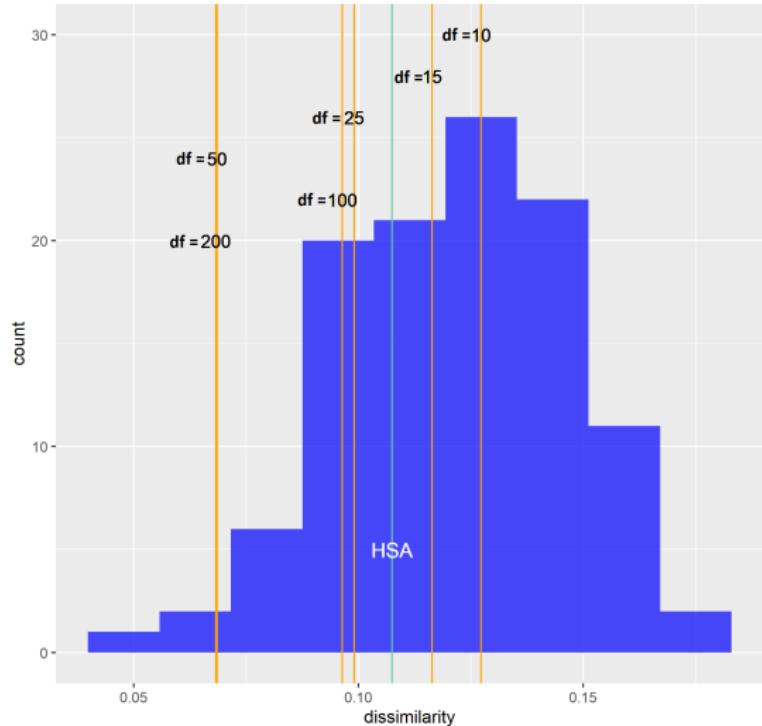
- ① Find the medoid replicate $i^* = \operatorname{argmin}_{i=1,\dots,N} \sum_{j=1}^N \rho(M_j, M_i)$
- ② Align replicates with the medoid $M_j \rightarrow M_j^{\text{rot}}$
- ③ Calculate gold standard \bar{M} as average of M_j^{rot}

Validation

Measure dissimilarities

- histogram $\rho(\bar{M}, M_i)$
- lines $\rho(\bar{M}, X)$ for PoisMS
- lines $\rho(\bar{M}, X)$ for HSA

Interpretation: resulting reconstructions lie within the range of statistical variation



Conformation reconstruction

Plan

- ① Background and previous work
- ② Methodology and data application
- ③ Reconstruction validation
- ④ Extensions

What modifications can we develop?

Discussion

Outline

- ① Set distribution of contact counts
- ② Link parameters to spatial structure
- ③ Write down the log-likelihood
- ④ Add smoothness
- ⑤ State optimization problem

Example

- ① $C_{ij} \sim Pois(\lambda_{ij})$
- ② $\log \lambda_{ij} = -\|x_i - x_j\|^2 + \beta$
- ③ $\ell_{PoisMS}(X, \beta)$
- ④ $X = H\Theta$
- ⑤ minimize $\ell(H\Theta, \beta)$

Discussion

Outline

- ① Set distribution of contact counts
- ② Link parameters to spatial structure
- ③ Write down the log-likelihood
- ④ Add smoothness
- ⑤ State optimization problem

Example

- ① $C_{ij} \sim Pois(\lambda_{ij})$
- ② $\log \lambda_{ij} = -\|x_i - x_j\|^2 + \beta$
- ③ $\ell_{PoisMS}(X, \beta)$
- ④ $X = H\Theta$
- ⑤ minimize $\ell(H\Theta, \beta)$

PCMS

Low-rank
Matrix
Approximation

Projection
step

WPCMS

Projected
Gradient
Descent

Optimize
SOA

PoisMS

Newton's
Method

Discussion

Outline

- ① Set distribution of contact counts
- ② Link parameters to spatial structure
- ③ Write down the log-likelihood
- ④ Add smoothness
- ⑤ State optimization problem

Example

- ① $C_{ij} \sim Pois(\lambda_{ij})$
- ② $\log \lambda_{ij} = -\|x_i - x_j\|^2 + \beta$
- ③ $\ell_{PoisMS}(X, \beta)$
- ④ $X = H\Theta$
- ⑤ minimize $\ell(H\Theta, \beta)$

Other distributions?

$C_{ij} \sim$ Zero-inflated Poisson, Hurdle Poisson, Negative binomial

Discussion

Outline

- ① Set distribution of contact counts
- ② Link parameters to spatial structure
- ③ Write down the log-likelihood
- ④ Add smoothness
- ⑤ State optimization problem

Example

- ① $C_{ij} \sim Pois(\lambda_{ij})$
- ② $\log \lambda_{ij} = -\|x_i - x_j\|^2 + \beta$
- ③ $\ell_{PoisMS}(X, \beta)$
- ④ $X = H\Theta$
- ⑤ minimize $\ell(H\Theta, \beta)$

Other smoothing approaches?

$$x_1, \dots, x_n \in \gamma(t), \text{ where } \gamma(t) = \begin{pmatrix} \gamma_1(t) \\ \gamma_2(t) \\ \gamma_3(t) \end{pmatrix}$$

Use roughness penalty and minimize $\ell_{PoisMS}(\gamma) + \lambda \int \|\gamma''(t)\|^2 dt$

Discussion

Outline

- ① Hyperparameter tuning
- ② Model validation

Example

- ① Elbow detection
- ② Multiplex FISH

Discussion

Outline

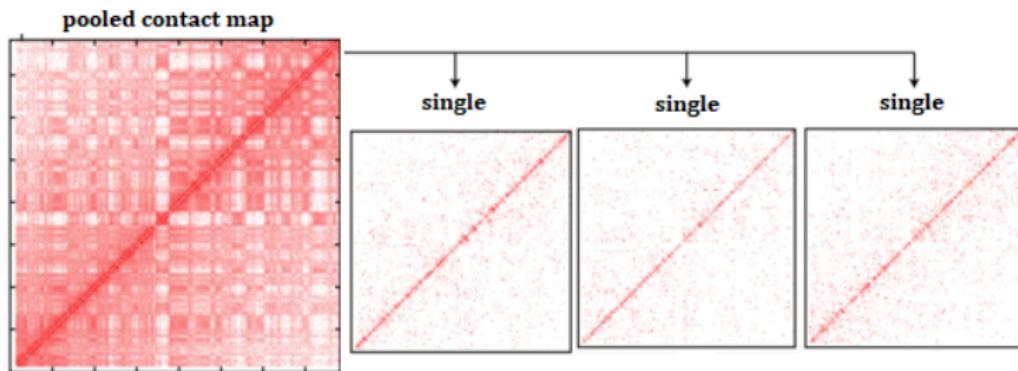
- ① Hyperparameter tuning
- ② Model validation

Example

- ① Elbow detection
- ② Multiplex FISH

Other approaches?

Use single-cell data for cross-validation



Discussion

Outline

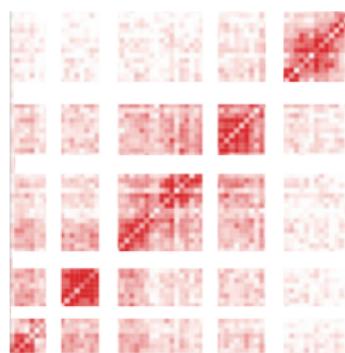
- ① Hyperparameter tuning
- ② Model validation

Example

- ① Elbow detection
- ② Multiplex FISH

Other approaches?

Run block-cross validation on the pooled data



BIG thanks to...

References

-  T. Stevens et al. *3D structures of individual mammalian genomes studied by single-cell Hi-C*. Nature, 544:59–64, 2017.
-  N. Varoquaux et al. *A statistical approach for inferring the three-dimensional structure of the genome*. Bioinformatics, 30(12):26-33, 2014.
-  Z. Zhang et al. *3D Chromosome Modeling with Semi-Definite Programming and Hi-C Data*. Journal of computational biology, 20(11):831–846, 2013.
-  C. Zou et al. *HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure*. Genome Biology, 17(40), 2016.
-  M. Rosenthal et al. *Bayesian Estimation of 3D Chromosomal Structure from Single Cell Hi-C Data*. Journal of Computational Biology, published online, 2019.
-  A. Buja et al. *Data Visualization With Multidimensional Scaling*. Journal of Computational and Graphical Statistics, 17(2):444-472, 2008.
-  R. Mazumder et al. *Spectral Regularization Algorithms for Learning Large Incomplete Matrices*. Journal of Machine Learning Research, 11: 2287-2322, 2010.
-  N. Srebro et al. *Weighted Low-Rank Approximations*. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

Thank you!

Complexity

PCMS complexity is $O(nk + k^3)$

- Compute $\tilde{Z} = H^T Z H \in \mathbb{R}^{k \times k}$
- Find eigen decomposition of \tilde{Z}

WPCMS complexity is $O((n^2 + nk + k^3) \cdot I)$

- **[Gradient]** $S := S - (G - G_+)$ where $G = W * (Z - D^2(X))$
- **[Projection]** $\Theta := \text{PCMS}(S, H)$

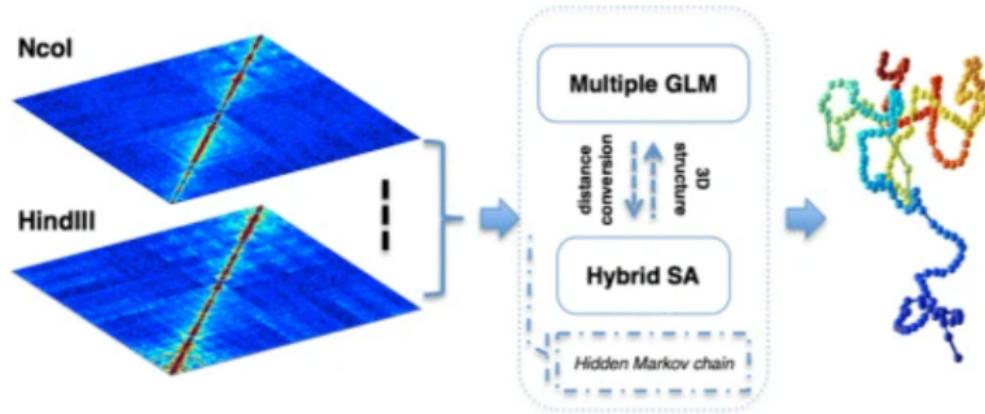
PoisMS complexity is $O((n^2 + nk + k^3) \cdot I \cdot E)$

- **[SOA]** $W = e^{-D^2(X)+\beta}$ and $Z = D^2(X) - \frac{C-W}{W}$.
- **[WPCMS]** $\Theta := \text{PCMS}_W(Z, H)$

HSA

"HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure", Chenchen Zou, Yuping Zhang, Zhengqing Ouyang, *Genome Biology*, 2016

- uses GLM for contact counts;
- characterizes the adjacency relationship of neighboring loci by a Gaussian Markov chain

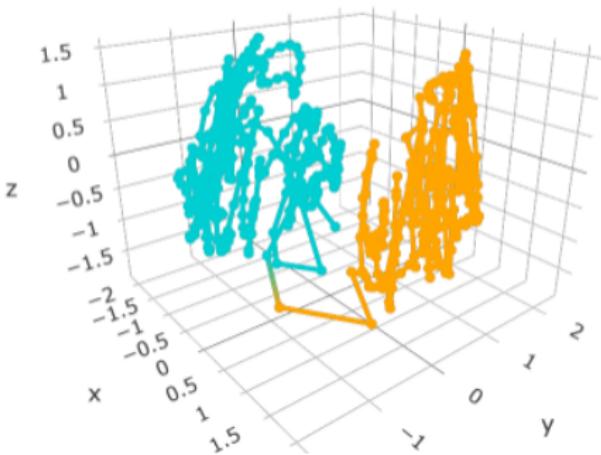
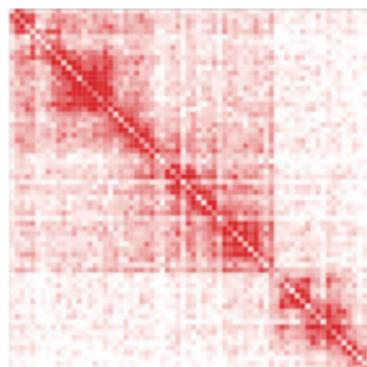


Conformation reconstruction

Motivation: obtain the reconstruction of the 3D chromatin architecture.

Why is it important? Chromatin conformation is a crucial component of numerous cellular processes including transcription.

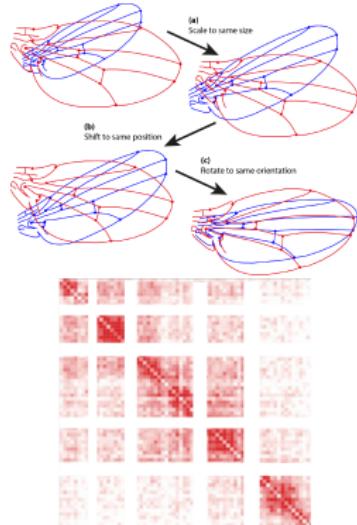
Why is it challenging? It is not possible to directly observe the 3D structure, however, you can obtain co-called *contact matrix*.



Chromatin reconstruction

Future directions:

- ① Develop suitable methodology for comparing two reconstructions, e.g. Procrustes distance robust to the local deformations.
- ② Develop methodology for validation through the contact matrix, e.g. cross-validation robust to high correlation in the data.



Weighted low-rank matrix approximation

Motivation: develop a generalization of the low-rank matrix approximation approach.

Why is it important? LRMA is the core of many machine learning techniques: data compression, dimension reduction and de-noising.

Goal: given some matrix $M \in \mathbb{R}^{n \times p}$ and weights $W \in [0, 1]^{n \times p}$ identify the “best” way to approximate M with a rank- k matrix X

$$\text{minimize } \|\sqrt{W} * (M - X)\|_F^2 \text{ w.r.t. } X \in \mathbb{R}^{n \times p} \text{ subject to } \text{rk}(X) \leq k$$

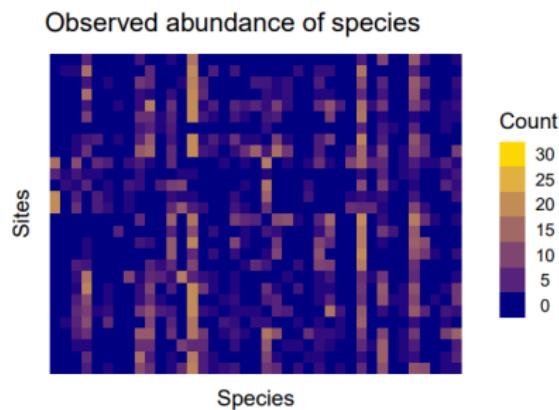
Well studied cases:

- Explicit solution when $W_{ij} = 1$. Used in principal components, linear discriminant analysis and canonical correlation analysis.
- Soft-impute algorithm when $W_{ij} \in \{0, 1\}$. Used in recommendation systems.

Weighted low-rank matrix approximation

Future directions:

- ① Use WLRMA to develop the framework for matrix-type generalized linear models.
- ② Study applications of these GLMs. For example, in *ecology* populations of species can be modeled via Poisson GLMs; in *item response theory* the matrix-type logistic regression can be used.

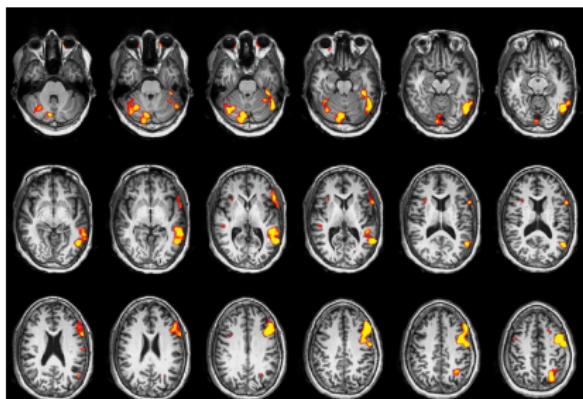


Structured canonical correlation analysis

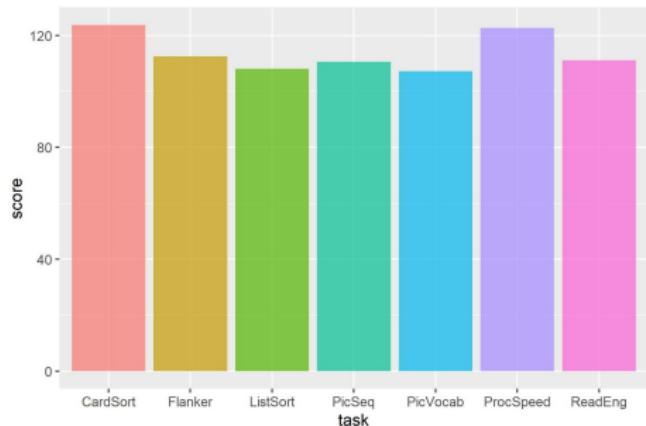
Motivation: how emotional disorders influence the brain activity?

Why is it important? It will help us to understand better the depression and anxiety phenomenon and design the personalized treatment in future.

Why is it challenging? The brain data has very specific structure: extremely high-dimensional with the structure imposed by the brain geometry.



$X \in \mathbb{R}^{n \times p}$ – brain activations



$Y \in \mathbb{R}^{n \times q}$ – behavior test scores

Structured canonical correlation analysis

Goal: understand if there is some correlation between measurement matrices X and Y .

Canonical correlation analysis + regularization + group penalty

maximize $\text{cor}(X\alpha, Y\beta)$ w.r.t. $\alpha \in \mathbb{R}^p, \beta \in \mathbb{R}^q$

subject to α is sparse

α follows the brain structure

Future directions:

- ① Extend the methodology to the case when some data is missing.
- ② Develop multiway CCA. For instance, if X , Y , Z correspond to the fMRI data, questionnaire and gene expression data.
- ③ Develop pooled CCA. For instance, if $(X_1, Y_1), \dots, (X_N, Y_N)$ represent the data from different institutions, each using different types of the questionnaire.