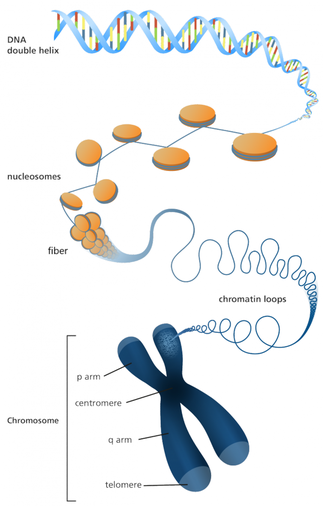
Chromatin reconstruction via Weighted Principal Curves

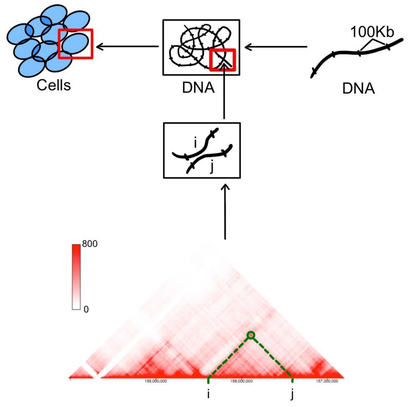
[Elena Tuzhilina] Stanford University, Department of Statistics
Joint work with T.Hastie and M.Segal

Motivation

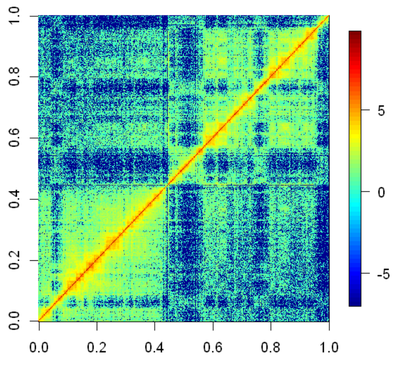


Chromatin is a highly organized DNA + protein structure which enables the approximately 2m of DNA contained in each human cell to be packaged into the nucleus. Three dimensional (3D) chromatin spatial organization is critical for numerous cellular processes, including transcription.
Genome architecture had been notoriously difficult to elucidate, but the recent advent of the suite of chromatin conformation capture assays, notably Hi-C, has transformed understanding of chromatin structure and provided downstream biological insights. The contact matrix resulting from Hi-C assays records the frequency with which pairs of binned genomic loci are cross-linked is commonly used to reconstruct chromatin conformation [1]. Most of existing approaches model chromatin as a polygonal chain and apply Multidimensional Scaling (MDS) techniques directly to the contact matrix [2, 3].
In this work we introduce a novel approach modelling chromatin by a smooth curve, develop Weighted Principal Curve technique and demonstrate its application to the real contact matrix data.

Contact matrix



A *contact matrix* is a symmetric matrix C, where C_{ij} equal to the contact counts between genomic loci $i,\ j$.



The logarithm of contact matrix for chromosome 20 and probe resolution 100 kilobases.

References

[1] T. Stevens et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544:59–64, 2017.

[2] N. Varoquaux et al. A statistical approach for inferring the three-dimensional structure of the genome. *Bioinformatics*, 30(12):123–456, 2014.

[3] Z. Zhang et al. 3d Chromosome Modeling with Semi-Definite Programming and Hi-C Data. *Journal of computational biology*, 20(11):831–846, 2013.

Notations

Frobenius norm:

$$||A||_F = \sqrt{\sum_{(i,j)} A_{ij}^2} = \sqrt{\mathrm{tr}(AA^T)}$$

Hadamard product: $(A * B)_{ij} = A_{ij}B_{ij}$

SVD decomposition: $A = UDV^T$

- $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ are orthogonal,
- $D \in \mathbb{R}^{m\times n}$ is a diagonal matrix with $D_{11} \geq D_{22} \geq \ldots \geq D_{mm} \geq 0$.

Principal curves

Goal: Given matrix $C \in \mathbb{R}^{n\times n}$ find 3D-embedding $x_1, \ldots, x_n \in \mathbb{R}^3$.

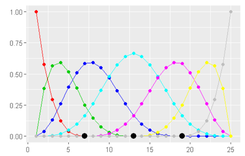
Multidimentional Scaling (MDS)
Interpret C as a similarity matrix and approximate it by Gram matrix.

$$\text{minimize } ||(C - XX^T)||_F \quad \text{w.r.t. } X \in \mathbb{R}^{n\times 3}$$

Principal Curve (PC)

$$x_1,\ x_2,\ \ldots,\ x_n \in \text{smooth curve.}$$

Spline basis matrix



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

Smooth curve constraint

$$X = H\Theta, \text{ where } \Theta \in \mathbb{R}^{k\times 3}.$$

$$\text{minimize } ||(C - H\Theta\Theta^T H^T)||_F \text{ w.r.t. } \Theta$$

Solution

- $H^T C H = UDV^T$
- $\Theta = U_{\cdot,1:3}\ \sqrt{D_{1:3,1:3}}\ V^T_{\cdot,1:3}$

$$\mathrm{PC}(C) = \mathrm{argmin}_\Theta\ ||(C - H\Theta\Theta^T H^T)||_F.$$

Weighted Principal Curve

Problem: Suppose $W \in [0,1]^{n\times n}$

$$\text{minimize } ||\sqrt{W} * (C - H\Theta\Theta^T H^T)||_F \quad \text{w.r.t. } \Theta$$

WPC algorithm

1. *[Initialize]* Generate random $\Theta \in \mathbb{R}^{k\times 3}$
2. Repeat until convergence:
   - 2.1 $\tilde{C} = H\Theta\Theta^T H^T$
   - 2.2 *[Mix]* $\tilde{C} := W * C + (1 - W) * \tilde{C}$
   - 2.3 *[Project]* $\Theta := \mathrm{PC}(\tilde{C})$

$$\mathrm{PC_W}(C) = \mathrm{argmin}_\Theta\ ||\sqrt{W} * (C - H\Theta\Theta^T H^T)||_F.$$

Projected gradient descent

$$f(M) = ||\sqrt{W} * (C - HMH^T)||_F$$

$$S^k_+(3) = \{M \in \mathbb{R}^{k\times k} : M \succeq 0, \mathrm{rk}(M) = 3\}$$

$$\text{minimize } f(M) \text{ w.r.t. } M \in S^k_+(3)$$

1. *[Initialize]* Generate random $M \in S^k_+(3)$
2. Repeat until convergence:
   - 2.1 *[Gradient]* $M := M - \nabla f(M)$
   - 2.2 *[Projection]* $M := \mathrm{proj}_{S^k_+(3)}(M)$

Extensions: For *[Mix]* step

$$\tilde{C}_\alpha := \alpha W * C + (1 - \alpha W) * \tilde{C}$$

and add line search.

Probability model

Poisson model

$$C_{ij} \sim Pois(\lambda_{ij}),\ \ \log(\lambda_{ij}) = \alpha\langle x_i, x_j\rangle + \beta$$

Negative log-likelihood

$$\ell(X) = \sum_{1\leq i,j\leq n} e^{\alpha\langle x_i, x_j\rangle + \beta} - c_{ij}(\alpha\langle x_i, x_j\rangle + \beta).$$

$$\text{minimize } \ell(X) \text{ w.r.t. } X \quad \text{subject to } X = H\Theta$$

Second order approximation (SOA)

$$\ell(X) \approx \ell_{SOA}(X) = \frac{\alpha^2}{2} \cdot ||\sqrt{W} * (Z - X^T X)||^2_F$$

$$W = \alpha XX^T + \beta \text{ and } Z = XX^T + \tfrac{1}{\alpha}(\tfrac{C}{W} - 1)$$

Under constraint $X = H\Theta$ ***the solution is*** $\Theta = \mathrm{PC_W}(Z)$ ***!***

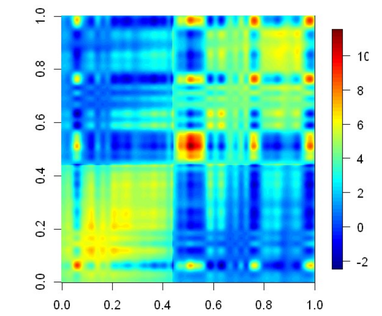
WPois algorithm (basic)

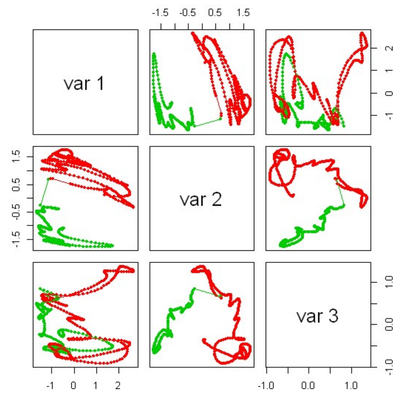
1. *[Initialize]* Generate random Θ
2. Repeat until convergence
   - 2.1 *[SOA]* For current guess Θ calculate
     - $W = e^{\alpha \cdot H\Theta\Theta^T H^T + \beta}$
     - $Z = H\Theta\Theta^T H^T + \frac{1}{\alpha}(\frac{C}{W} - 1)$
   - 2.2 *[WPC]* Solve WPC problem Update $\Theta := WPC(Z, W)$

Extensions:

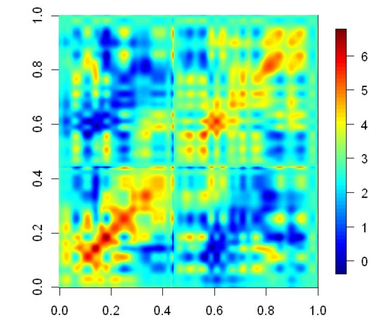
- Do a few steps of WPC
- Add learning rate and line search to SOA step (equivalence to Newton's Method)

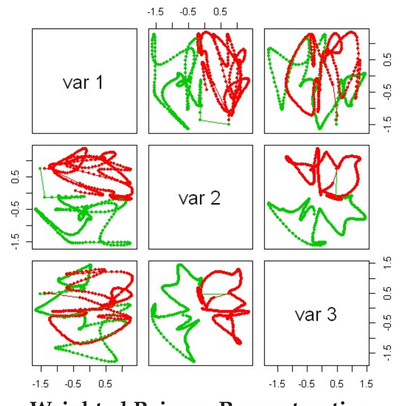
Reconstruction





Principal Curves Reconstruction





Weighted Poisson Reconstruction