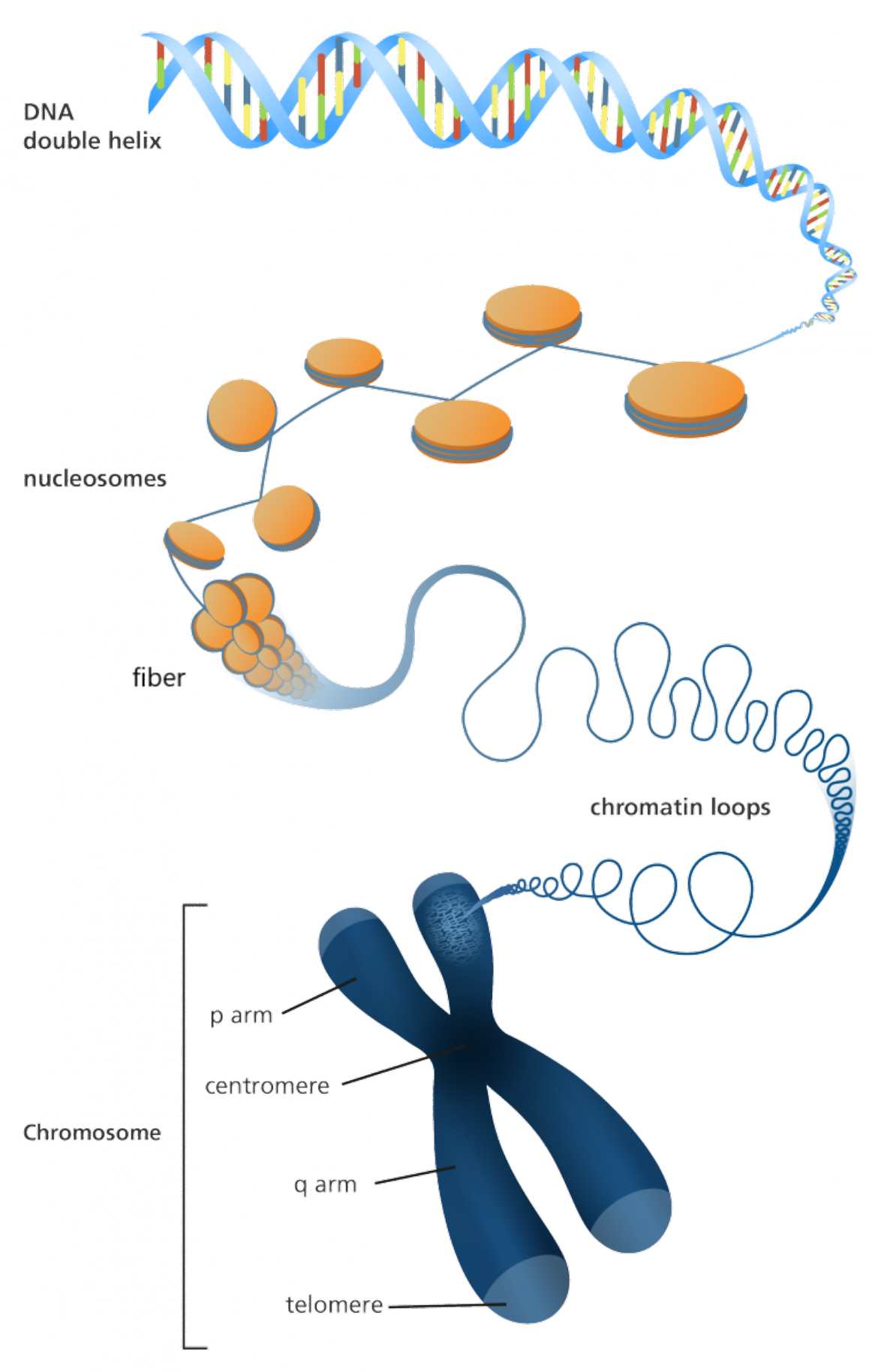
Statistical curve models for inferring 3D chromatin architecture

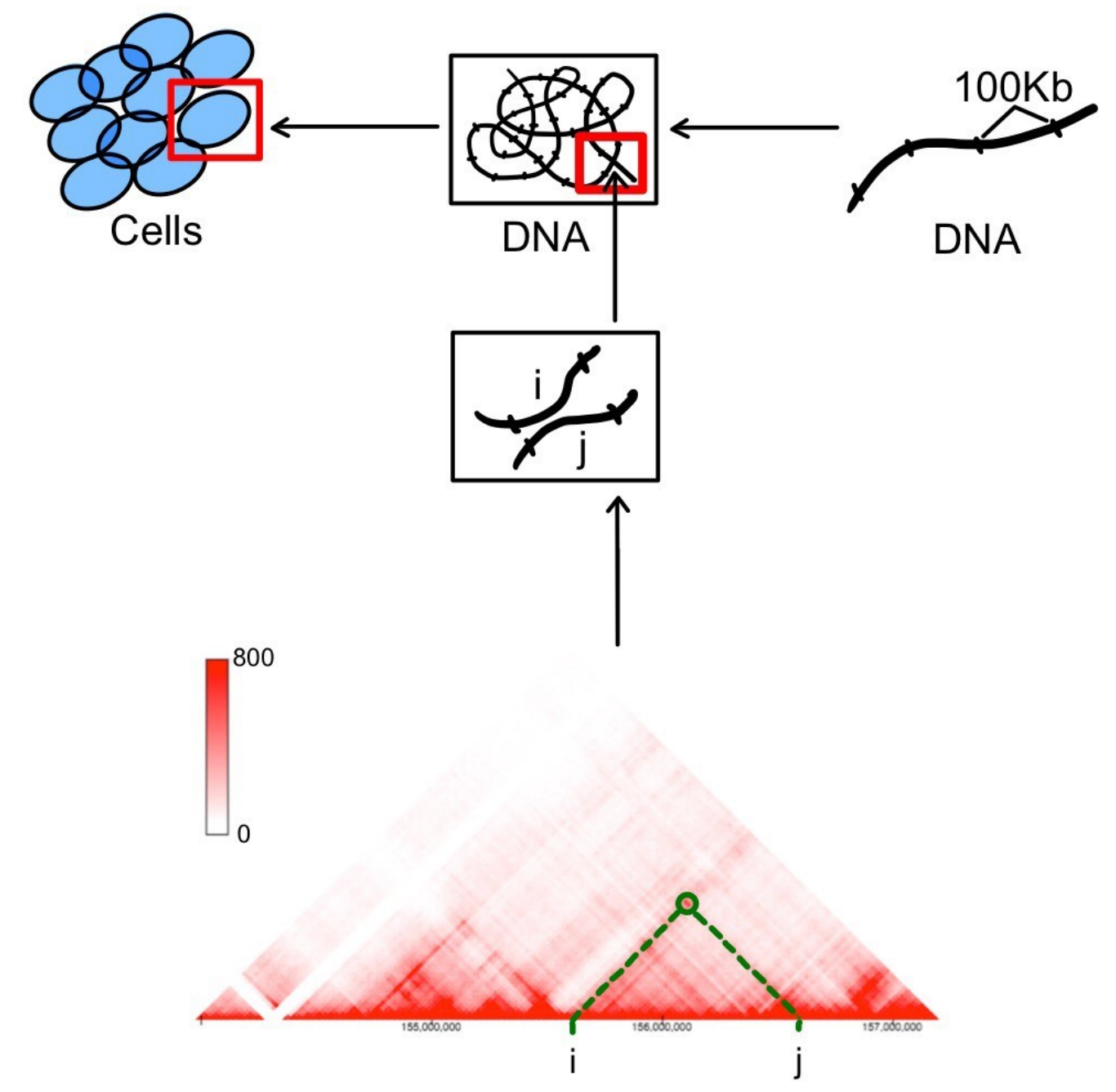
[Elena Tuzhilina] Stanford University, Department of Statistics
Joint work with Trevor Hastie and Mark Segal

Motivation

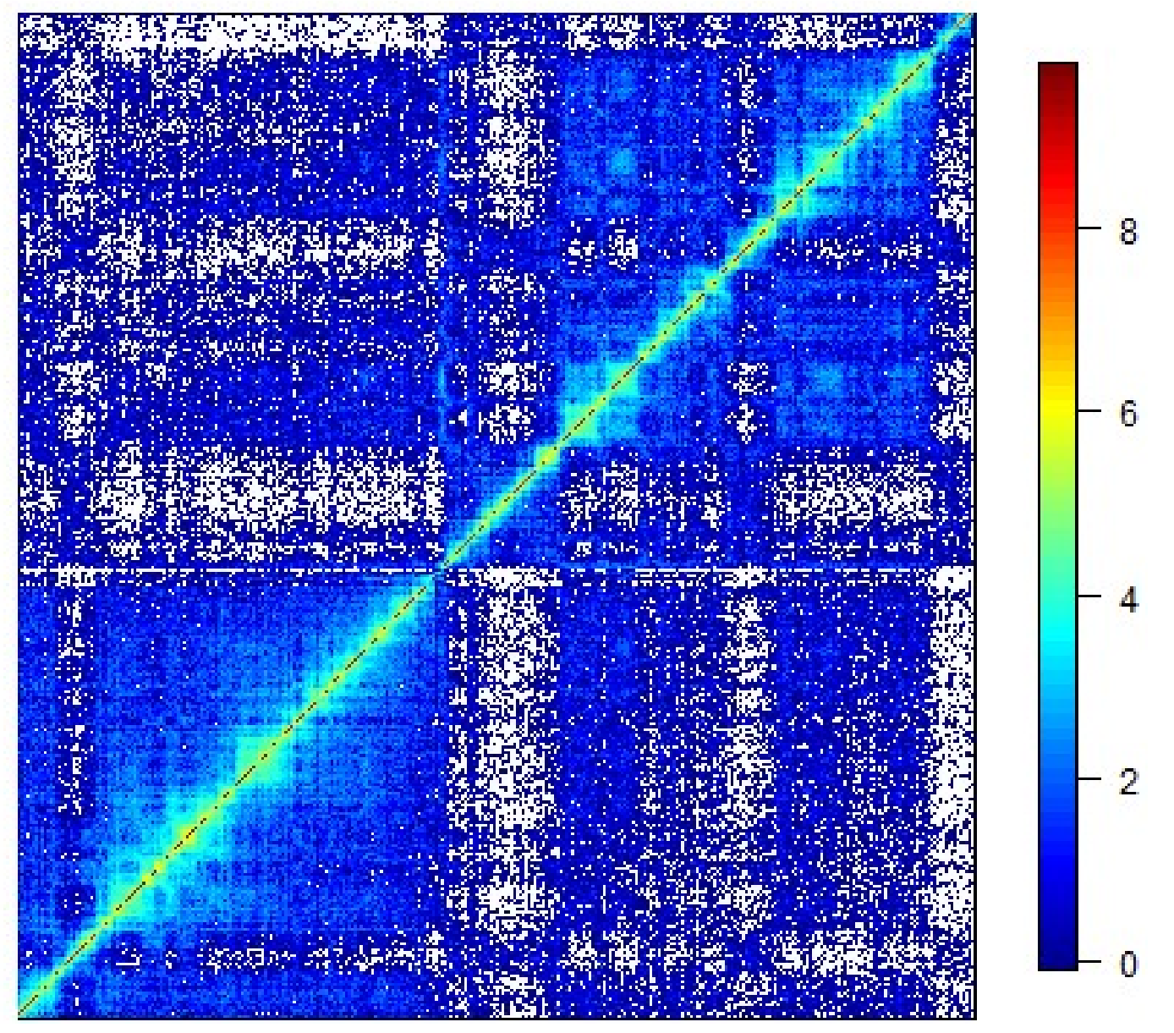
Chromatin is a highly organized DNA and protein structure which enables the approximately two meters of DNA contained in each human cell to be packaged into the nucleus. Three dimensional (3D) chromatin spatial organization is critical for numerous cellular processes, including transcription. Genome 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. Most of existing approaches model chromatin as a *polygonal chain* and apply Multidimensional Scaling (MDS) techniques directly to the contact matrix. In this work we introduce a novel approach modelling chromatin by a *smooth curve*.



Data



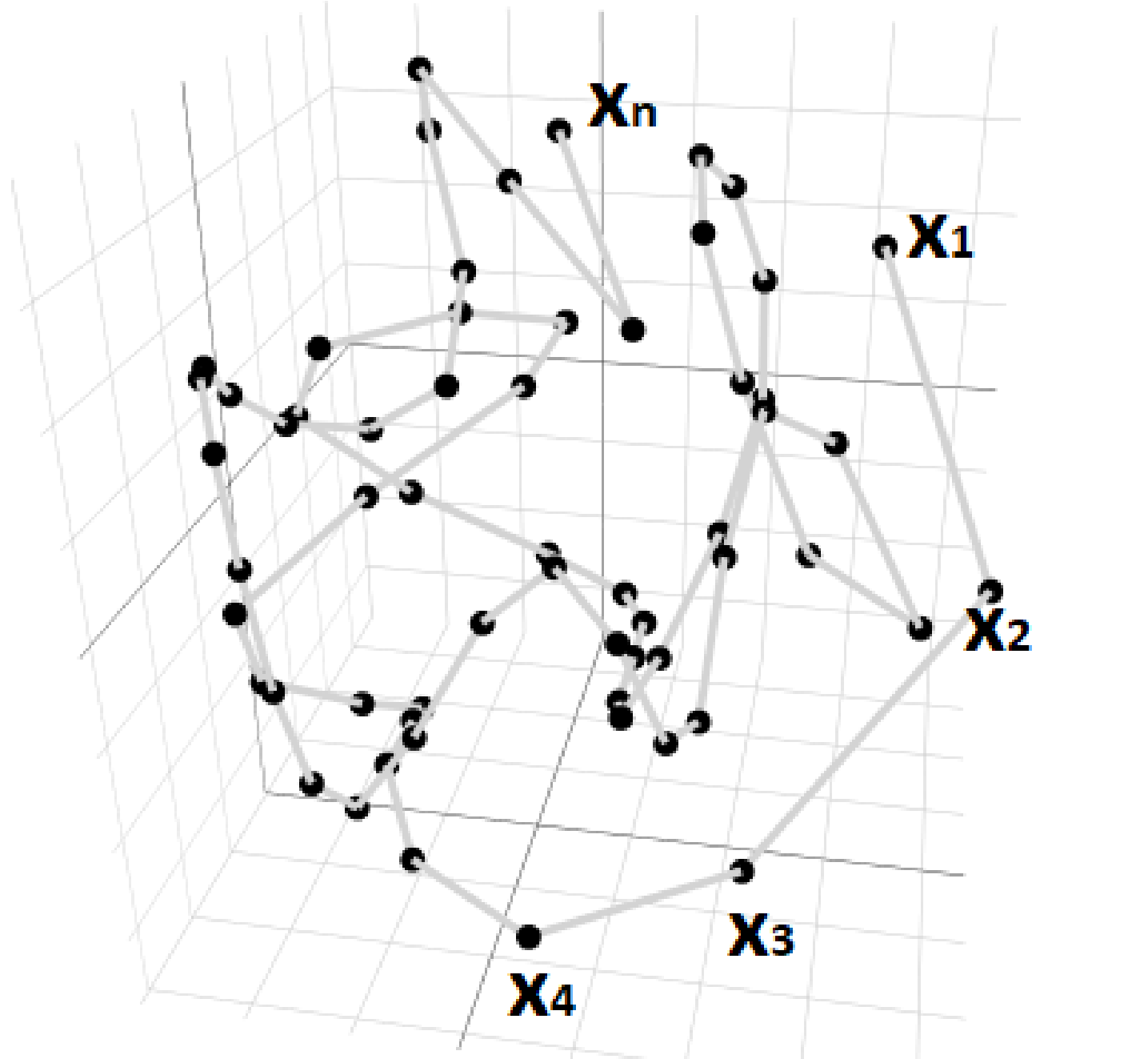
Contact matrix: $C = [C_{ij}] \in \mathbb{Z}_+^{n\times n}$ with elements representing contact frequencies between genomic loci i and j.



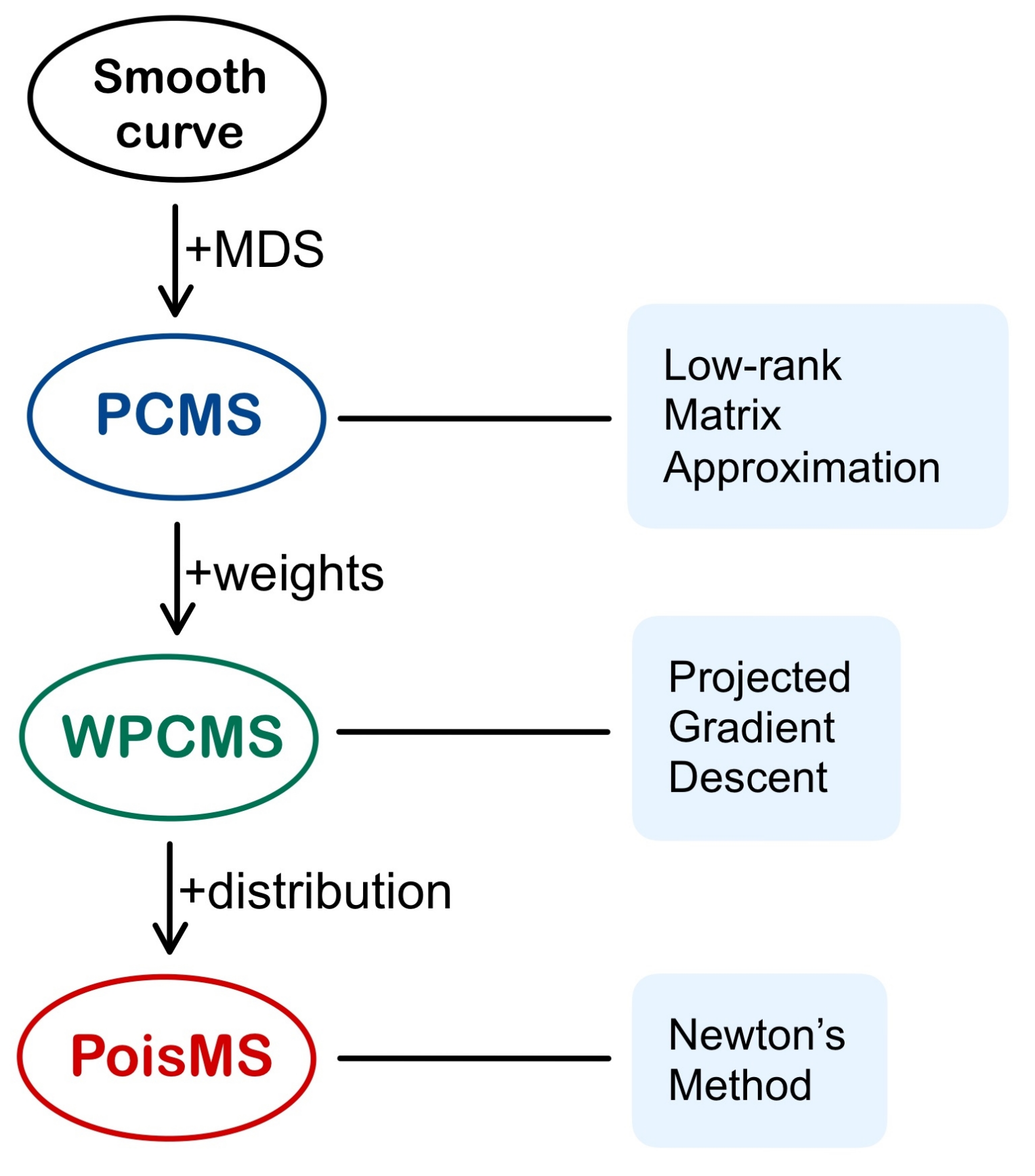
The heatmap of $\log(C)$ for chromosome 20 and probe resolution 100kb. Resulting number of genomic loci is $n = 625$.

Reconstruction challenge

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



Outline



Smooth curve

+MDS

PCMS — Low-rank Matrix Approximation

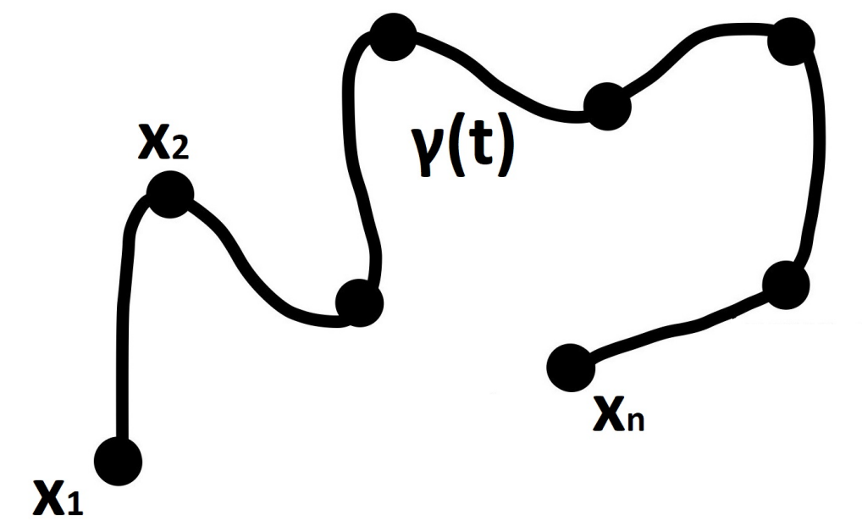
+weights

WPCMS — Projected Gradient Descent

+distribution

PoisMS — Newton's Method

Smooth curves



$$x_1,\ x_2,\ \dots,\ x_n \in \text{smooth curve } \gamma(t)$$

Assumptions:

- $\gamma(t) = \begin{pmatrix}\gamma_1(t)\\ \gamma_2(t)\\ \gamma_3(t)\end{pmatrix}$ and $\gamma_j(t)$ are splines
- $\gamma_j(t) = \sum_{\ell=1}^k \Theta_{\ell j}\ h_\ell(t)$ where

  $h_1(t), \dots, h_k(t)$ – spline basis
- parametrization $x_i = \gamma(i)$
- k = reconstruction degrees-of-freedom

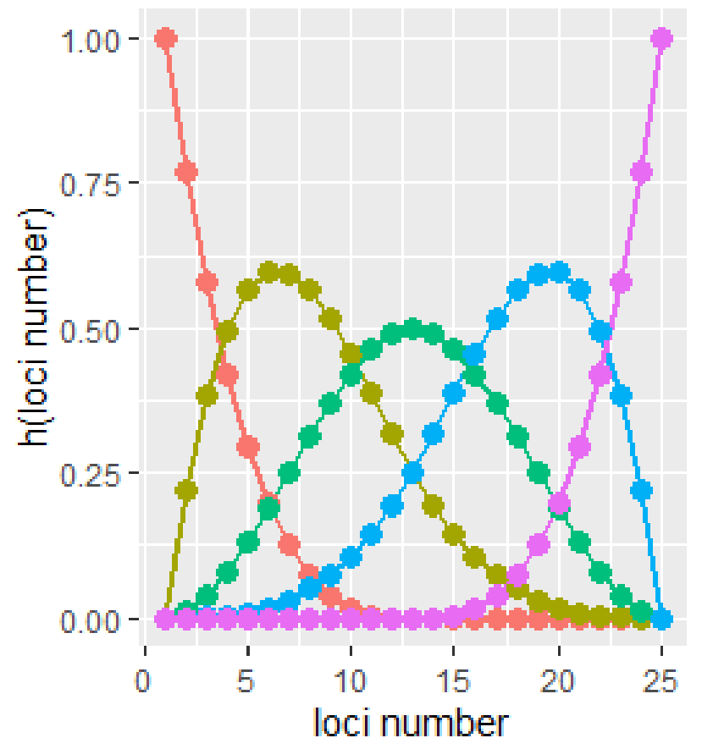
Smooth curve constraint:

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

Conformation matrix:

$$X = \begin{pmatrix} -x_1^\top - \\ \vdots \\ -x_n^\top - \end{pmatrix} \in \mathbb{R}^{n\times 3}$$

Spline basis matrix:



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

Principal curve metric scaling

PCMS = classical MDS + smooth curve

- convert C to a *similarity* matrix Z
- approximate Z by inner products

  $$S(X) = XX^\top$$
- add constraint $X = H\Theta$

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

Solution: via eigen decomposition of

$$H^\top Z H \in \mathbb{R}^{k\times k}$$

Weighted modification

Motivation: diagonal dominance and sparsity of C

WPCMS = PCMS + weight + distance

- convert C to a *dissimilarity* matrix Z
- approximate Z by distances

  $D^2(X)$ with $d_{ij}^2 = \|x_i - x_j\|^2$
- add weights

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

Solution: run projected gradient descent in the space of $S(X)$, where projection is performed via PCMS

```
1. [Initialize] Generate $X$
2. Repeat until convergence:
    • [Gradient] $G = W*(Z-D^2(X))$
      $S := S - (G - \text{diag}(G \cdot 1))$
    • [Projection] $X := \text{PCMS}(S)$
```

Add distribution

PoisMS = WPCMS + Poisson GLM

Model: $C_{ij} \sim \text{Pois}(\lambda_{ij})$, where

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

Negative log-likelihood:

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

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

Solution: perform Newton's method step via WPCMS

Idea: at current guess X_0

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

```
1. [Initialize] Generate $X$
2. Repeat until convergence:
    • [SOA] $\begin{cases} W = e^{-D^2(X)+\beta} \\ Z = D^2(X) - \frac{C-W}{W} \end{cases}$
    • [Newton] $X := \text{WPCMS}(Z, W)$
    • [Nuisance]
      $\beta := \log\left(\frac{\sum_{i\ j} C_{ij}}{\sum_{i\ j} e^{-\|x_i - x_j\|^2}}\right)$
```

Multiplex FISH

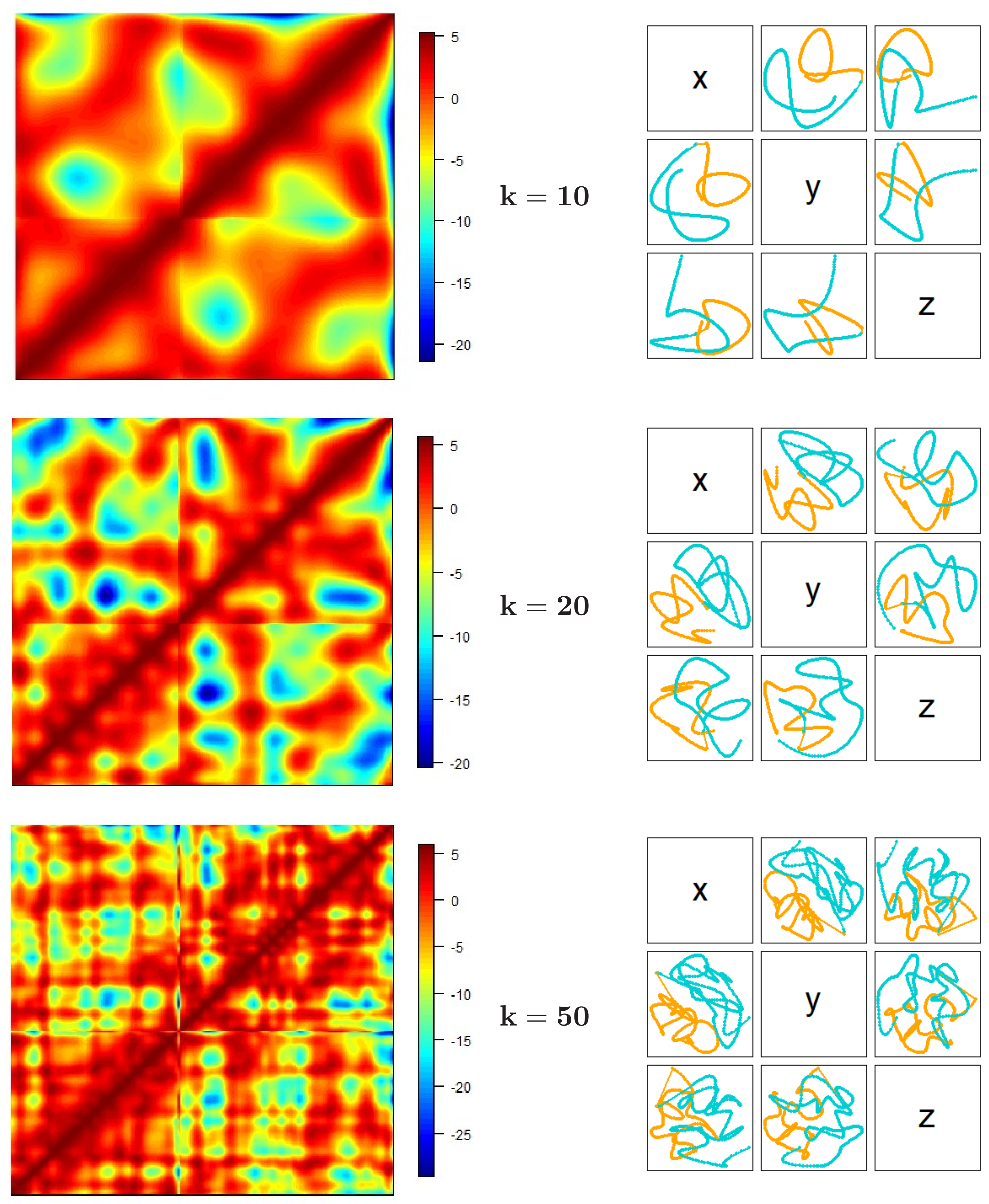
- low resolution ($\approx$ 30 genomic loci)
- many replicates ($>$ 100)

Idea: use Procrustes distance to measure dissimilarity between various reconstructions

Results

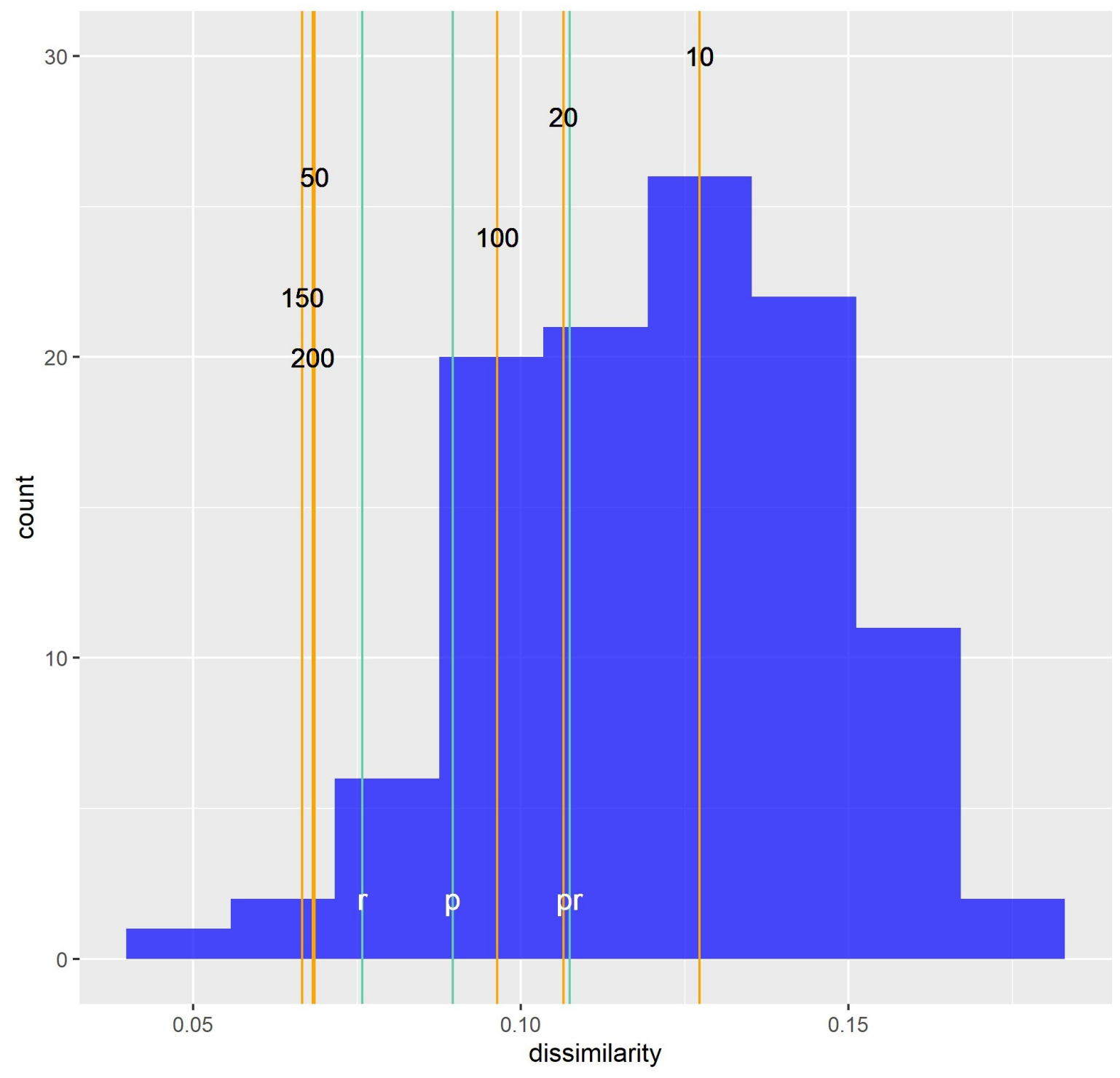
Approximation
$\log(\Lambda) = -D^2(X) + \beta$

3D conformation X



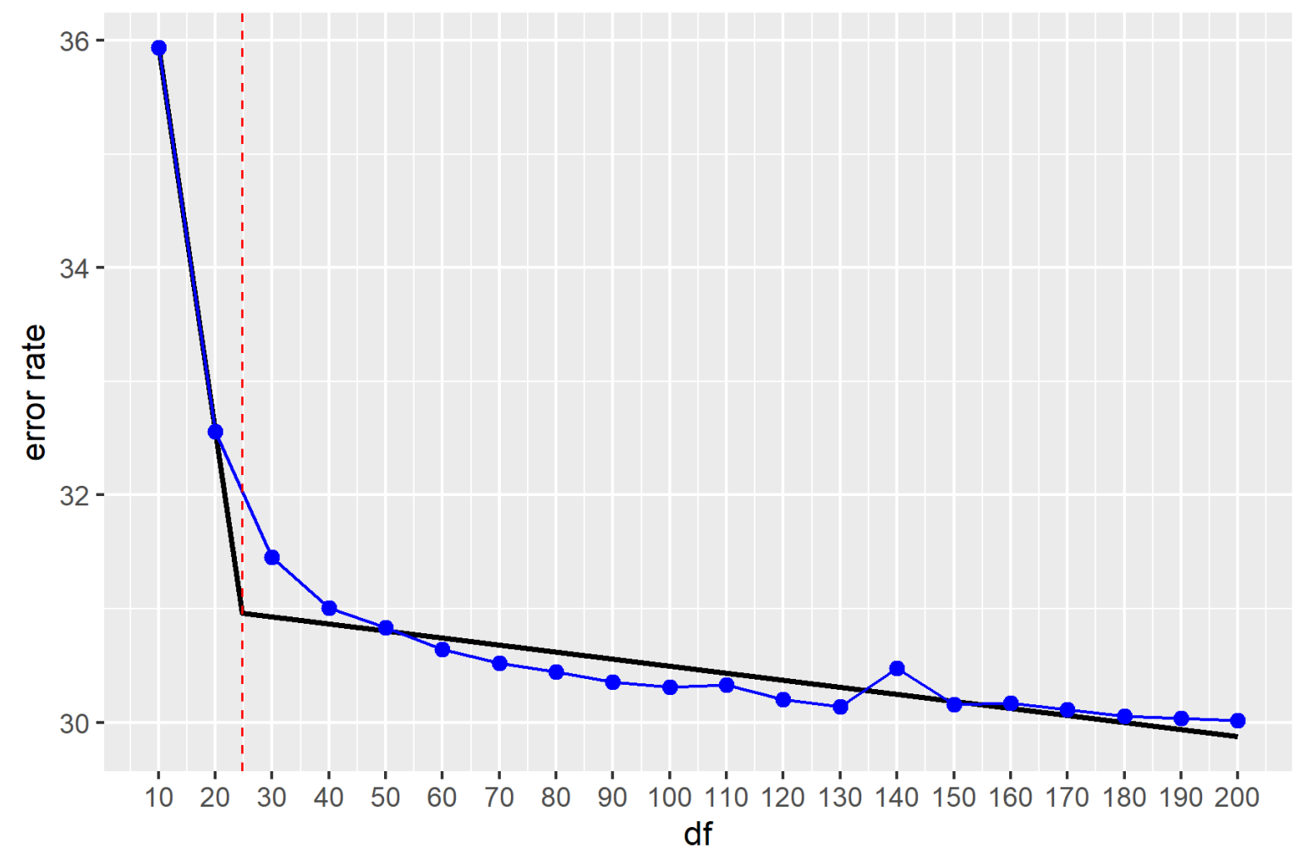
Validate reconstruction

1. use FISH to construct a gold standard
2. use the gold standard to compute the reference distribution
3. position the reconstructions PoisMS and HSA



Pick degrees-of-freedom

1. calculate X, β for a grid of k
2. for each k measure Poisson deviance
3. use elbow method to select the best value of k



Hyperparameter: k controls the spline basis size and how wiggly is the resulting reconstruction