Modeleing Complex Phenotypes

DCM&B seminar

Sayan Mukherjee

Departments of Statistical Science, Computer Science, Mathematics
Duke University
www.stat.duke.edu/∼sayan

Part I – DE. Runcie (UC Davis)
Part II – K. Turner (U Chicago), D. Boyer (Duke)

Nov 12, 2014

Two parts

(1) Bayesian sparse factor model to estimate genetic covariance.

Two parts

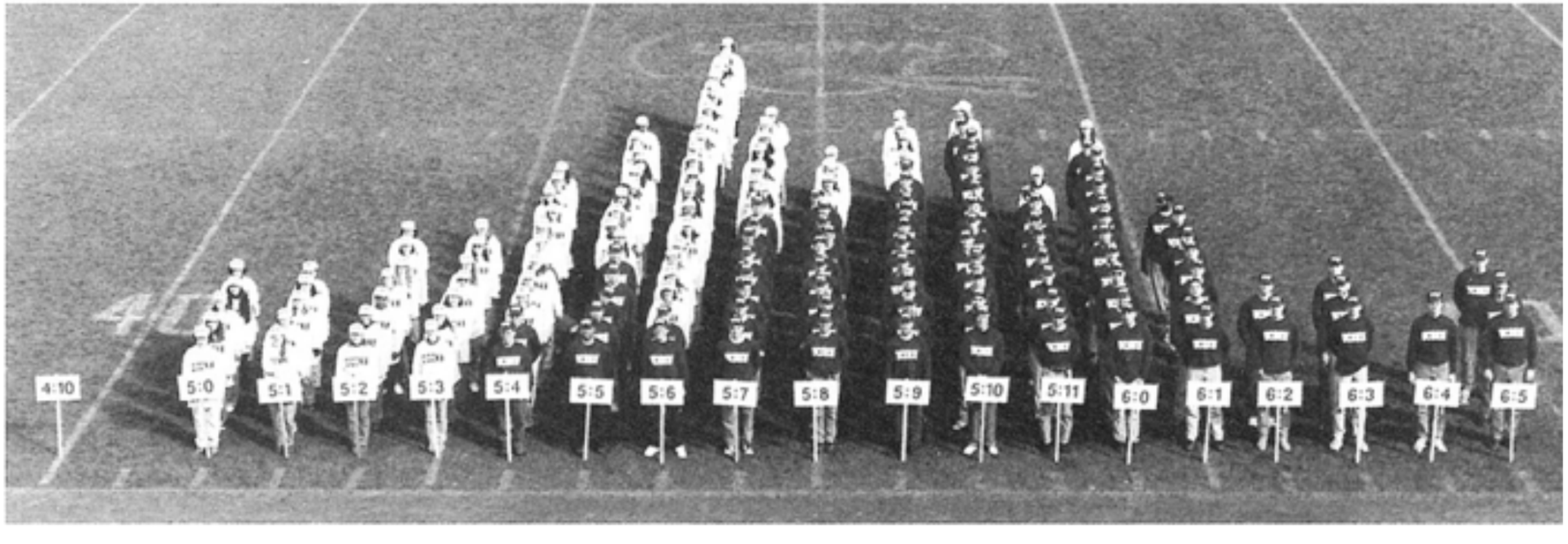
(1) Bayesian sparse factor model to estimate genetic covariance.

(2) Quantitative genetics of shapes.

Quantitative genetics

Genetics of multiple traits

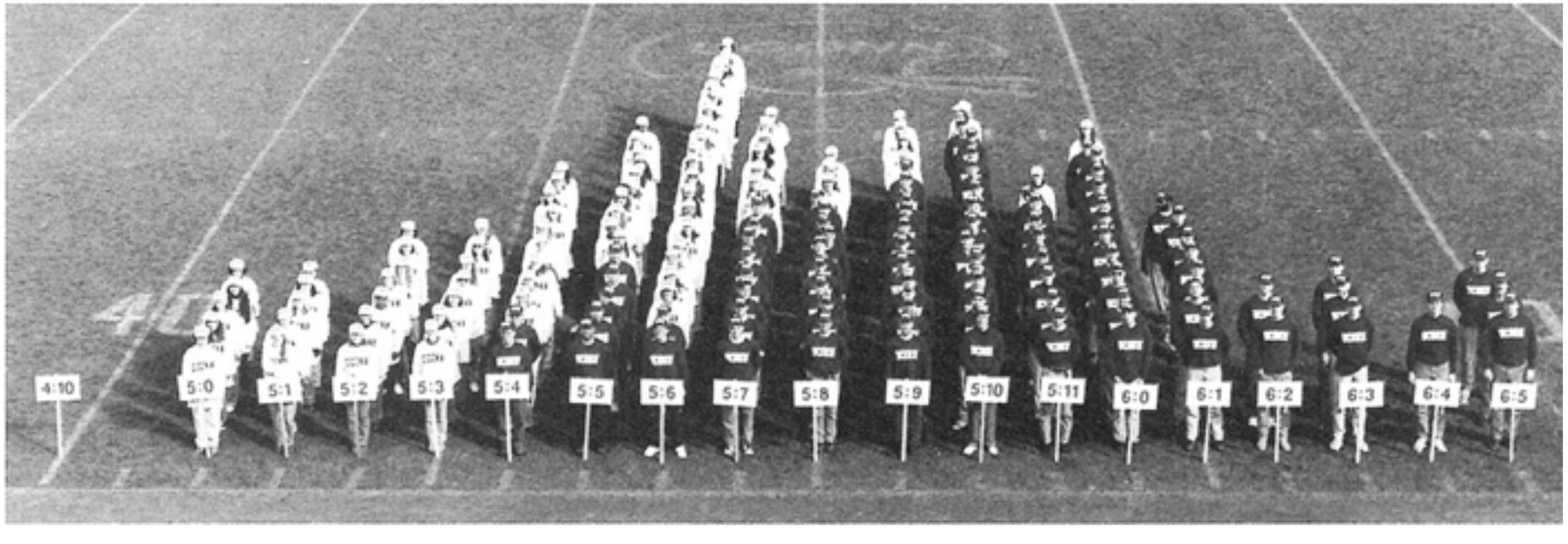
Phenotypic traits are often considered individually



Linda Strausbaugh (Genetics 147:5, 1997)

Genetics of multiple traits

Phenotypic traits are often considered individually



Linda Strausbaugh (Genetics 147:5, 1997)

Important phenotypes often involve many traits

BBC

Some objectives in quantitative genetics

Partition total phenotypic (trait) variation into genetic and environmental components.

$$\mathbf{P} = \mathbf{G} + \mathbf{E}.$$

G-matrix: matrix of genetic covariance among traits, $\mathbf{G}$.
E-matrix: matrix covariance among traits due to environment $\mathbf{E}$.

Some objectives in quantitative genetics

Partition total phenotypic (trait) variation into genetic and environmental components.

$$\mathbf{P} = \mathbf{G} + \mathbf{E}.$$

G-matrix: matrix of genetic covariance among traits, $\mathbf{G}$.
E-matrix: matrix covariance among traits due to environment $\mathbf{E}$.

Broad-sense heritability = genetic effects on phenotype, can be further partitioned into additive, dominant, and interaction effects.

Lande's equation

Only the additive effects can be passed from parent to offsping: narrow-sense heritability, h^2

Fisher's fundamental theorem (1930):
"The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time."

Lande's equation

Only the additive effects can be passed from parent to offsping: narrow-sense heritability, h^2

Fisher's fundamental theorem (1930):
"The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time."

Lande or breeder's equation:

$$R = h^2 s,$$

R - response to selection, S - selection differential.

Multivariate Lande's equation

G: matrix of additive genetic covariance among traits, **G**

Multivariate Lande's equation

G: matrix of additive genetic covariance among traits, **G**

Lande or breeder's equation:

$$\Delta \mathbf{y} = \mathbf{G}\mathbf{s}$$

$\mathbf{Y} \sim \mathrm{N}_p$: traits are multivariate normal
$\mathbf{s} = \frac{\partial F(\bar{\mathbf{Y}})}{\partial \mathbf{y}}$: selection gradient.

Estimating **G**

Previous aproaches

(1) Pairwise covariances followed by clustering – Ayroles and Stone.

Estimating **G**

Previous aproaches

(1) Pairwise covariances followed by clustering – Ayroles and Stone.

(2) Methods based on moments estimators – Hine and Blows, McGraw.

Estimating **G**

Previous aproaches

(1) Pairwise covariances followed by clustering – Ayroles and Stone.

(2) Methods based on moments estimators – Hine and Blows, McGraw.

(3) Linear mixed effects models – Henderson, Kruuk, Kirkpatrick and Meyer, De Los Campos and Gianola.

Estimating **G**

Previous aproaches

(1) Pairwise covariances followed by clustering – Ayroles and Stone.

(2) Methods based on moments estimators – Hine and Blows, McGraw.

(3) Linear mixed effects models – Henderson, Kruuk, Kirkpatrick and Meyer, De Los Campos and Gianola.

LMM model that scales to thousands of traits.

Genetics of many traits

Today we can measure thousands of traits simultaneously

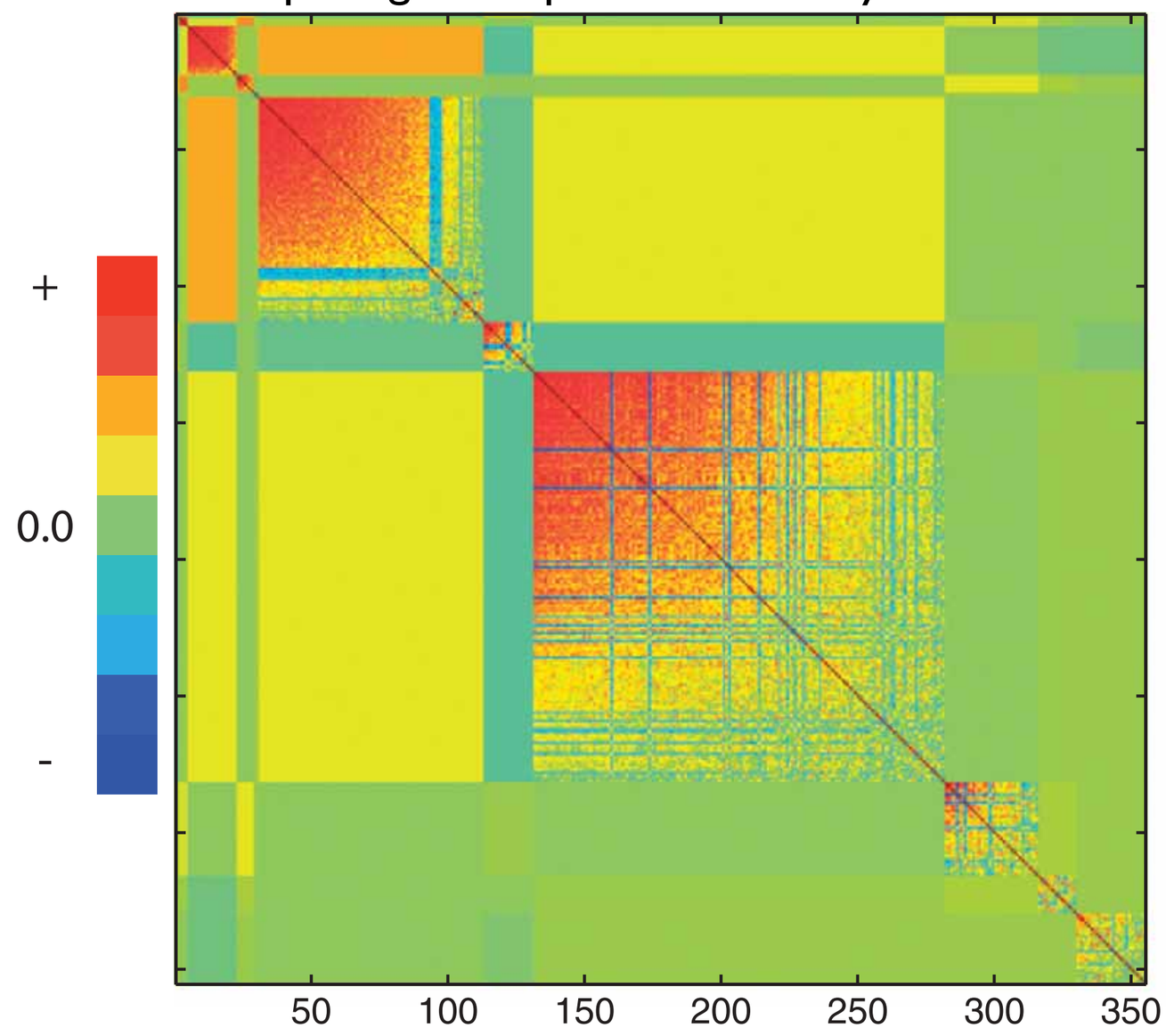
Genome-wide gene expression

Proteomics / metabolomics

morphometrics

genotype-environment interactions

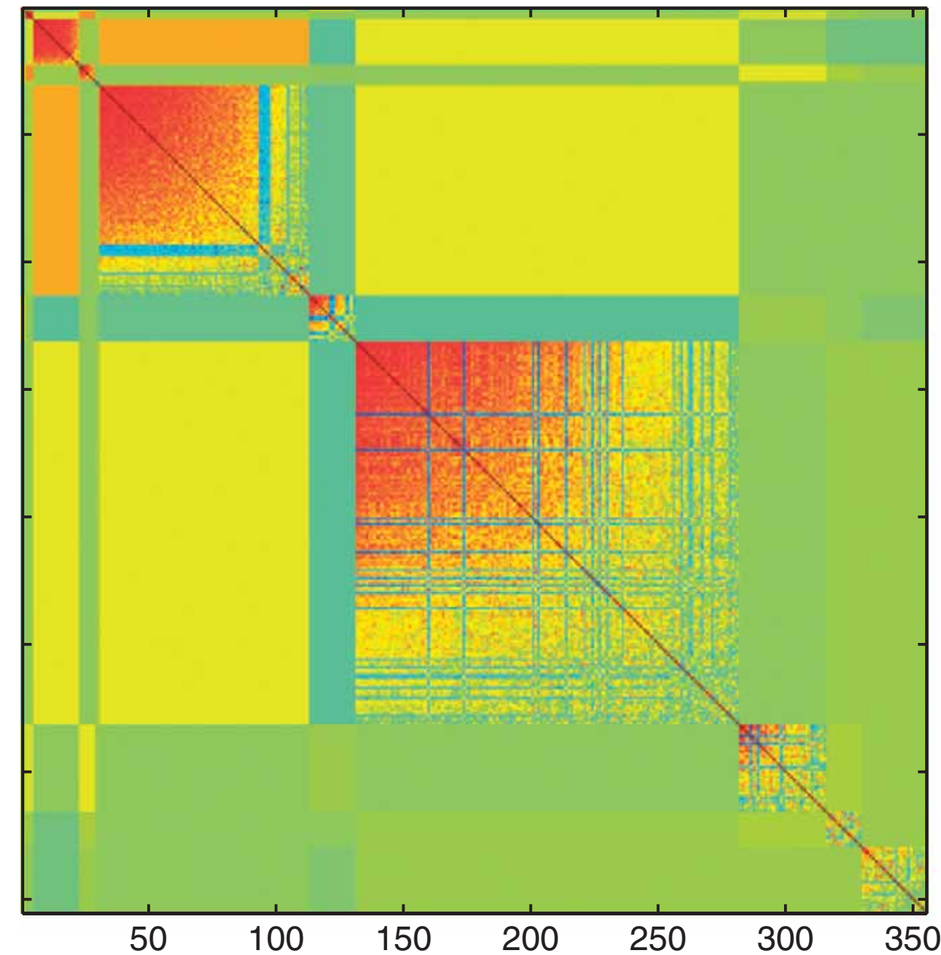
Drosophila gene expression from Ayroles et al 2009



New methods are necessary to take advantage of these data

Bayesian genetic sparse factor model

Ayroles et al 2009

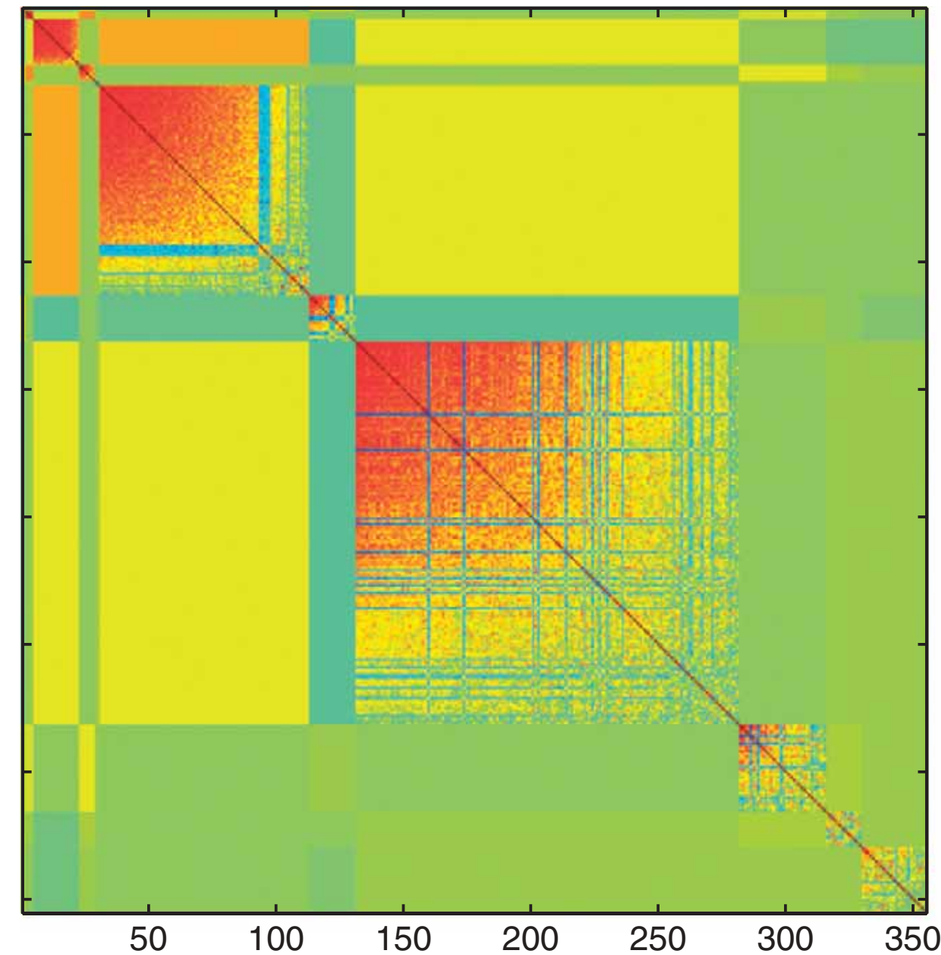


Goal:

- Reduce high-dimensional data to its underlying structure
- Estimate evolutionary parameters
- Handle complicated experimental designs or complex pedigrees
- Be scalable to large numbers of traits

Bayesian genetic sparse factor model

Ayroles et al 2009



Goal:

Reduce high-dimensional data to its underlying structure

Estimate evolutionary parameters

Handle complicated experimental designs or complex pedigrees

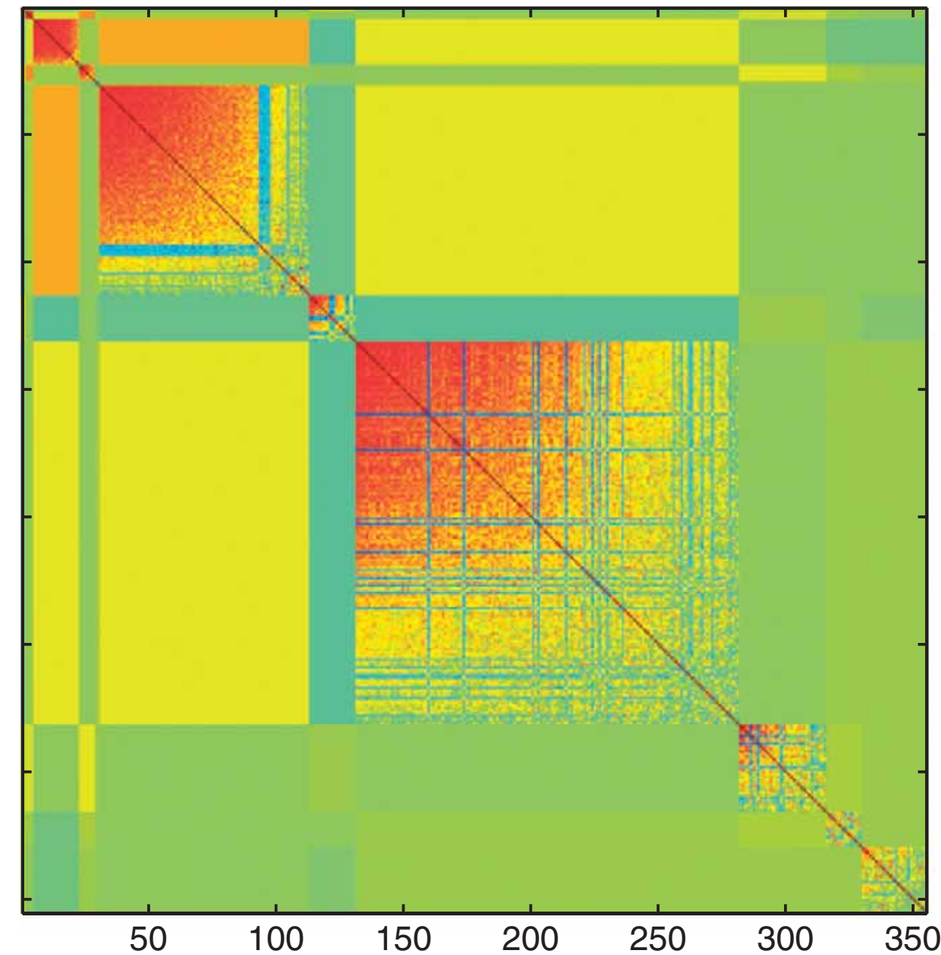
Be scalable to large numbers of traits

Methods: Bayesian dimension reduction

Sparse estimation of the **G** matrix based on an animal model

Bayesian genetic sparse factor model

Ayroles et al 2009



Goal:

Reduce high-dimensional data to its underlying structure

Estimate evolutionary parameters

Handle complicated experimental designs or complex pedigrees

Be scalable to large numbers of traits

Methods: Bayesian dimension reduction

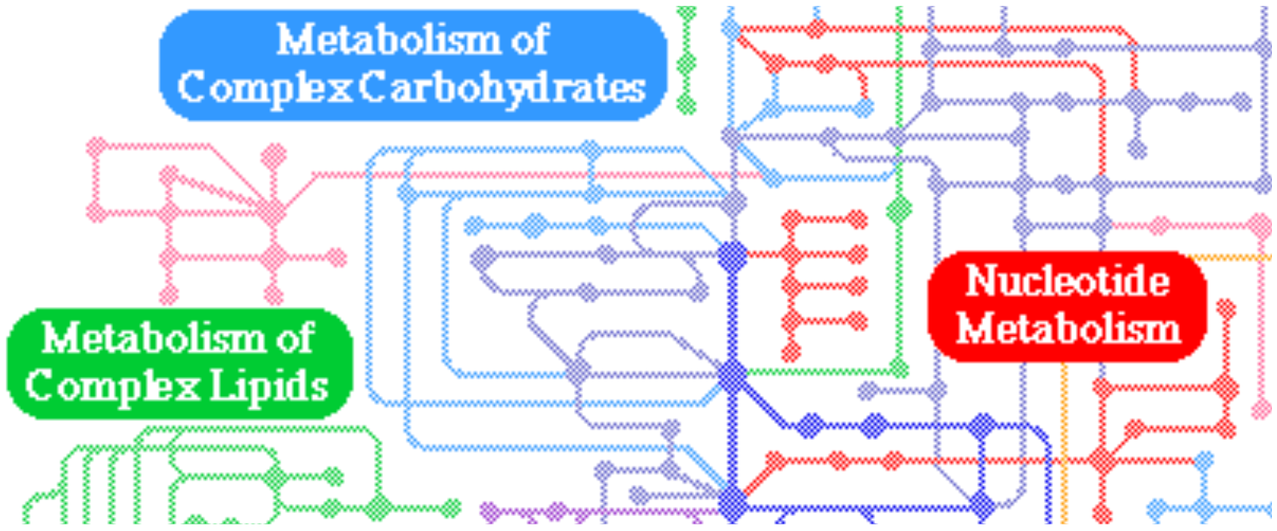
Sparse estimation of the **G** matrix based on an animal model

Case study:

An application to *Drosophila* gene expression data

Quantitative Genetics of Gene Expression

Gene expression is a readout of cellular activities



Metabolism, and cell-signaling activity is difficult to measure

but may be key determinants of fitness

A factor model for G

Animal model for multiple traits

phenotypes

fixed effects

Random genetic effects

residual error

$$\mathbf{y}_i = \mathbf{x}_i^T \mathbf{b}_i + \mathbf{u}_i + \epsilon_i$$

A factor model for G

Animal model for multiple traits

phenotypes — fixed effects — Random genetic effects — residual error

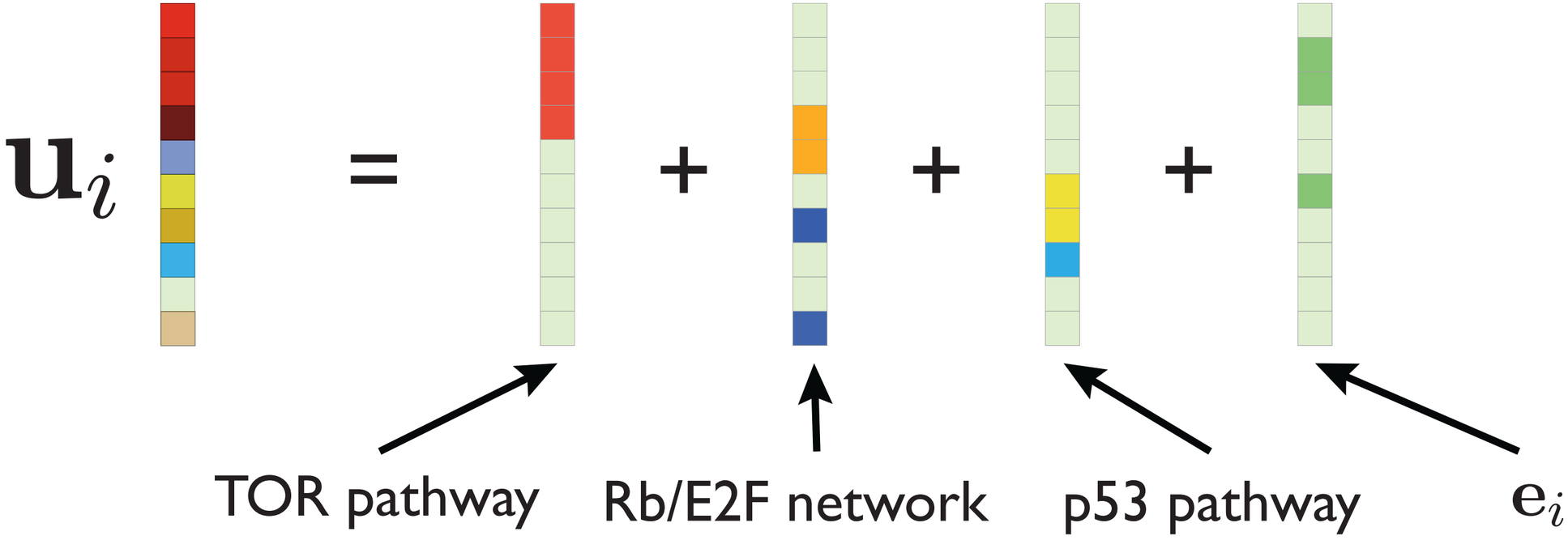
$$\mathbf{y}_i = \mathbf{x}_i^T \mathbf{b}_i + \mathbf{u}_i + \epsilon_i$$

genes

$$\mathbf{u}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{G})$$

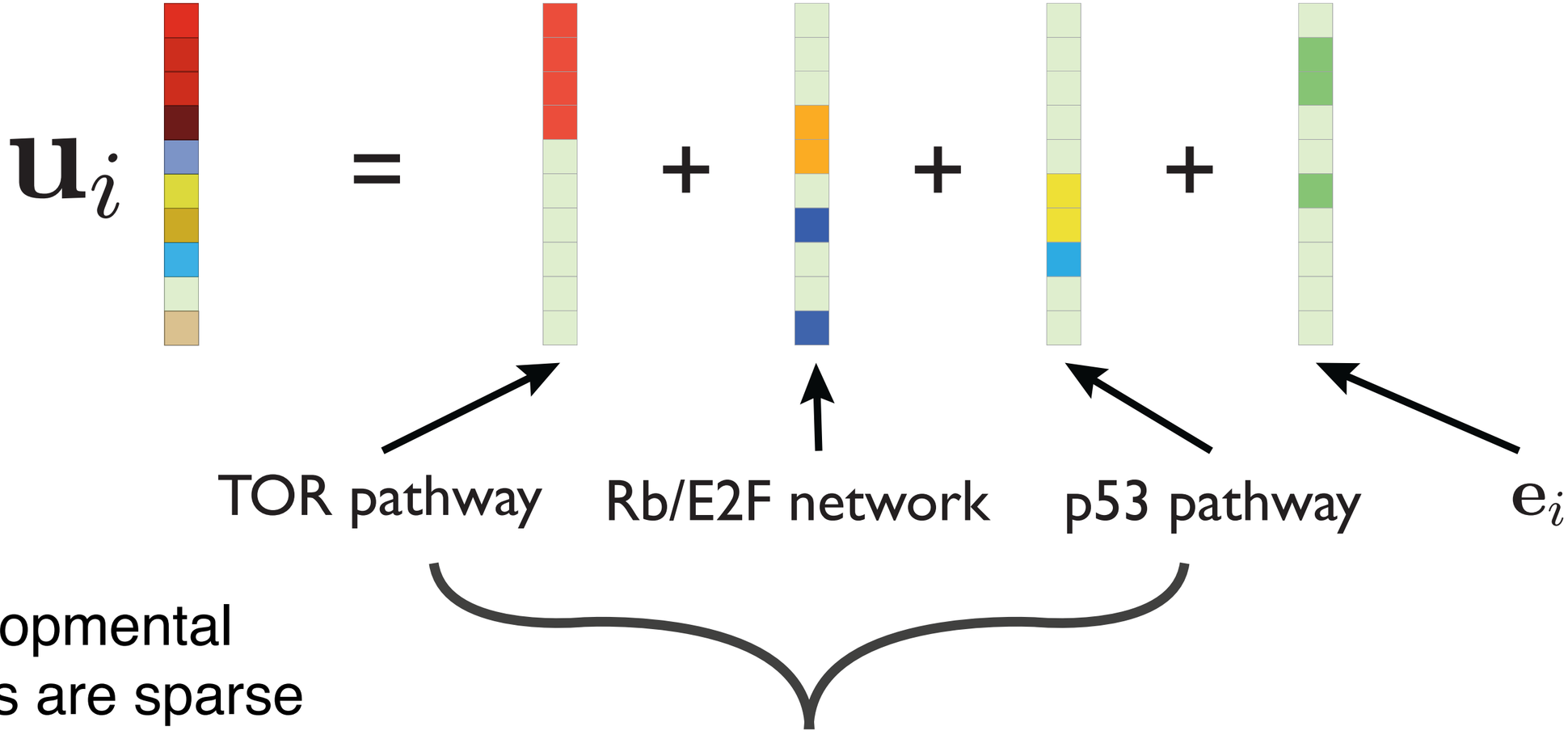
A factor model for G

Model **u** as output of development



A factor model for G

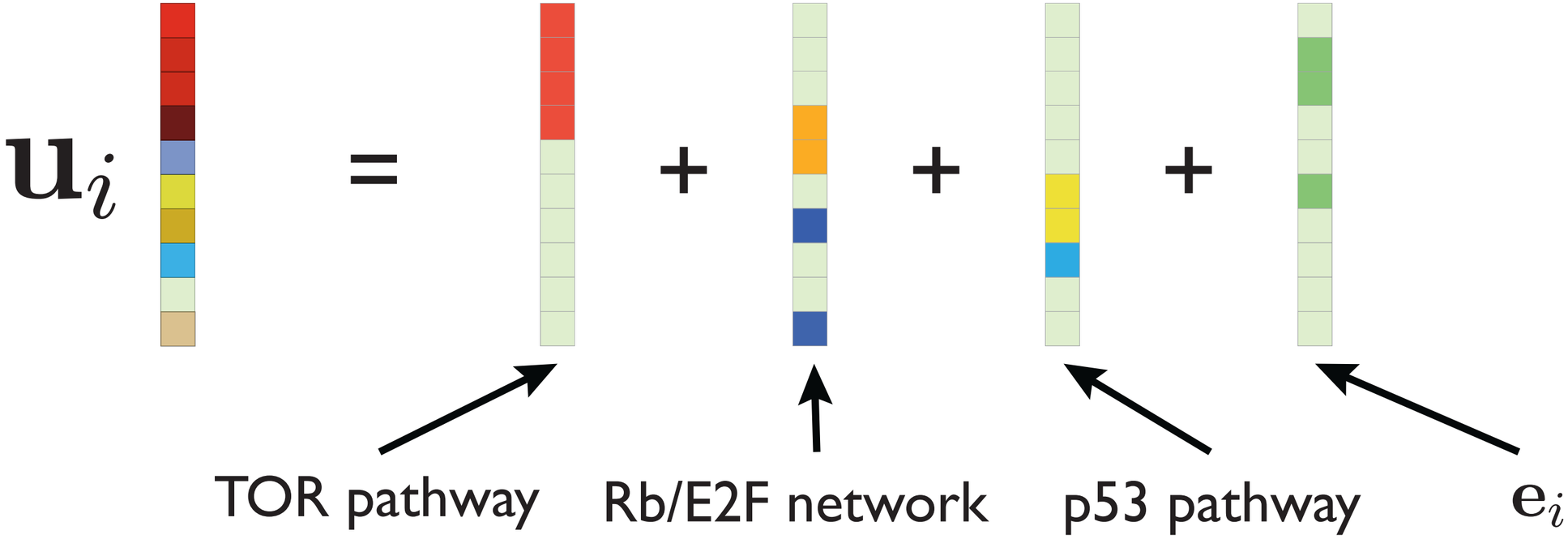
Model **u** as output of development



Developmental effects are sparse

A factor model for G

Model **u** as output of development

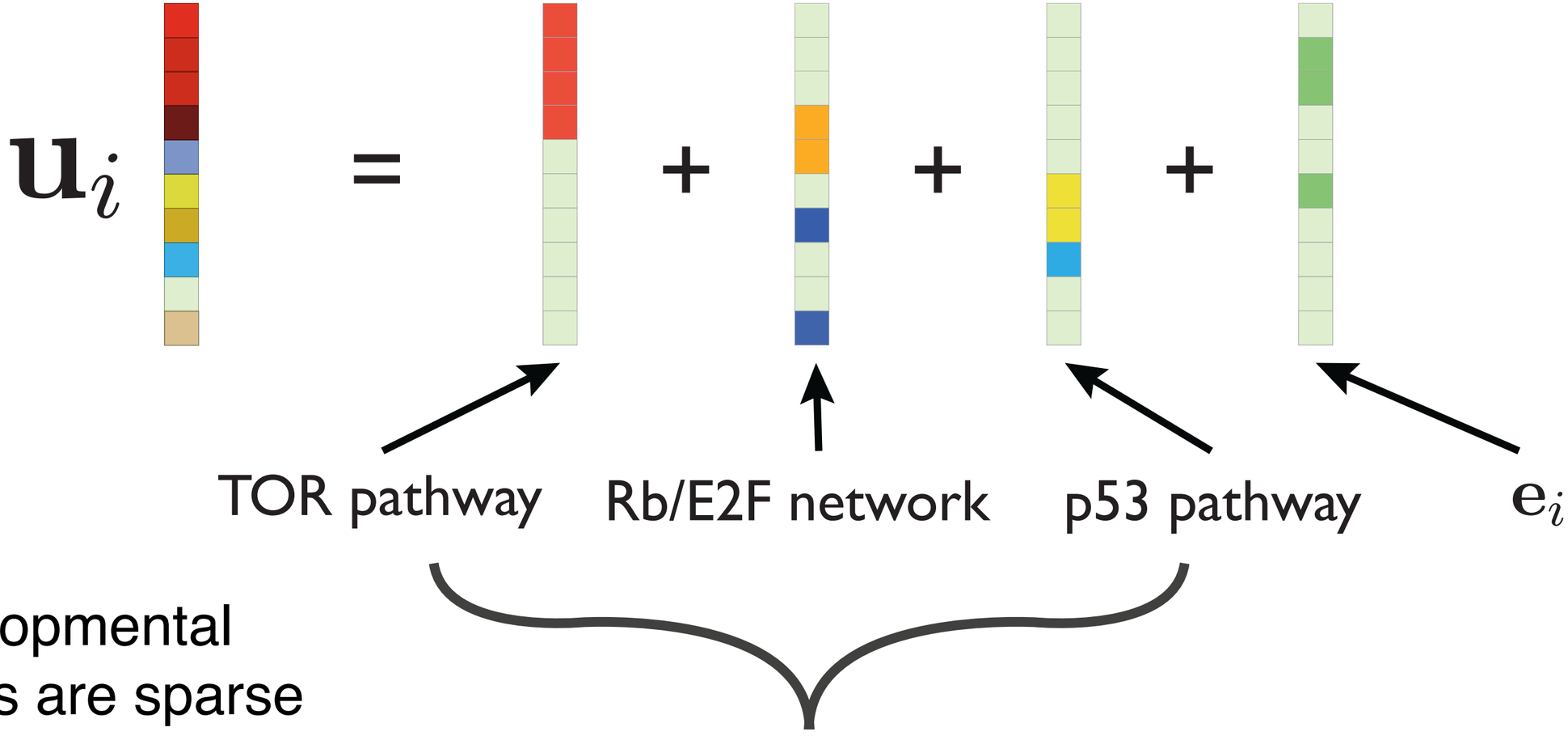


Developmental effects are sparse

1) Few underlying developmental pathways are genetically variable

A factor model for G

Model **u** as output of development



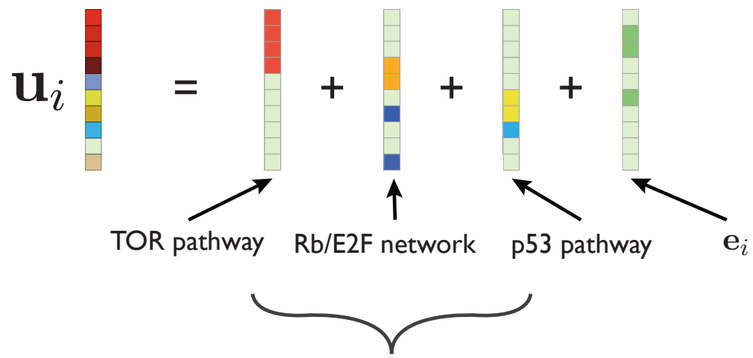
Developmental effects are sparse

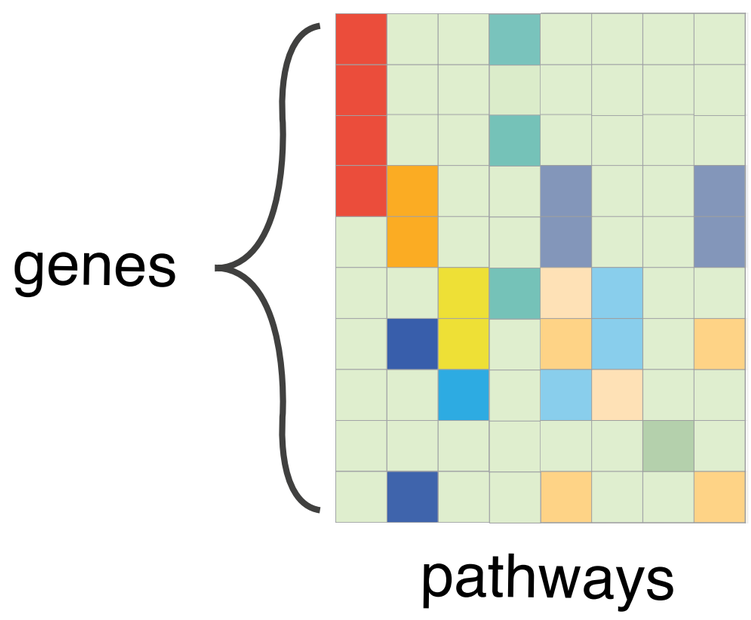
1) Few underlying developmental pathways are genetically variable

2) Each pathway affects a low number of genes

A factor model for G

Sparsity assumptions are key for high-dimensional data

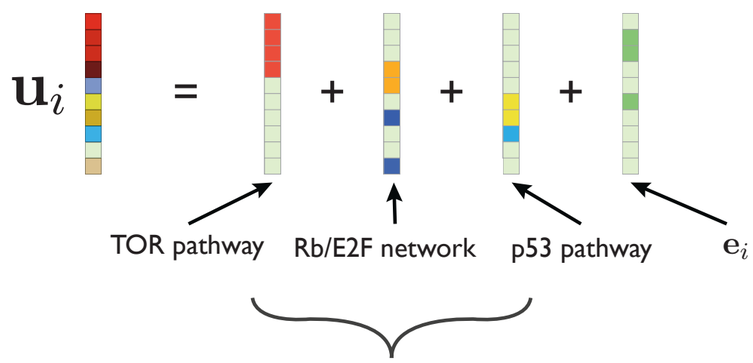


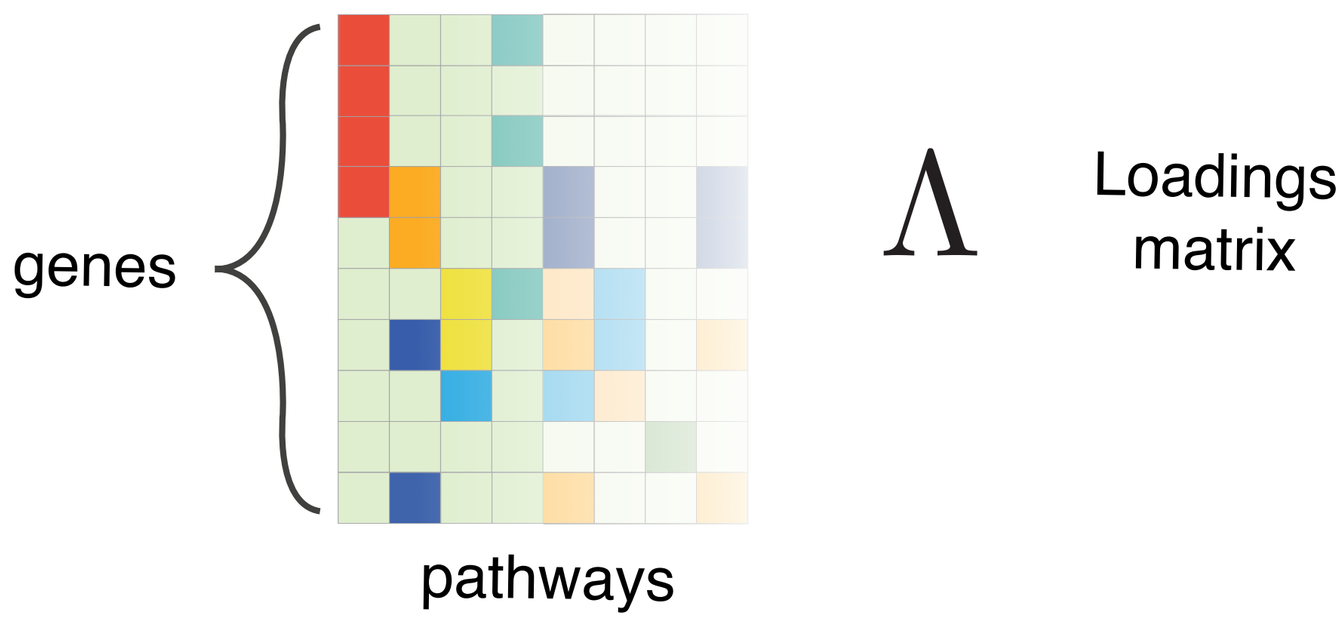


Λ Loadings matrix

A factor model for G

Sparsity assumptions are key for high-dimensional data

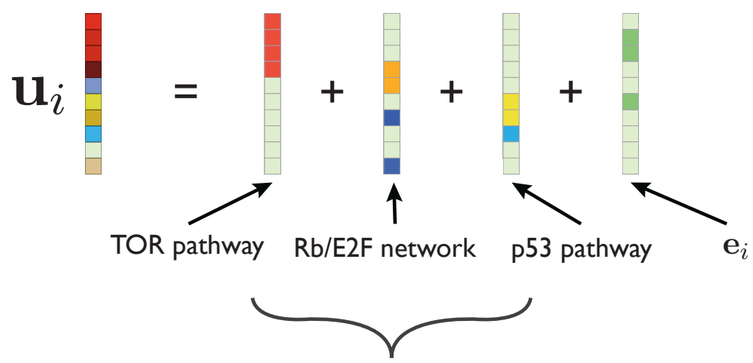


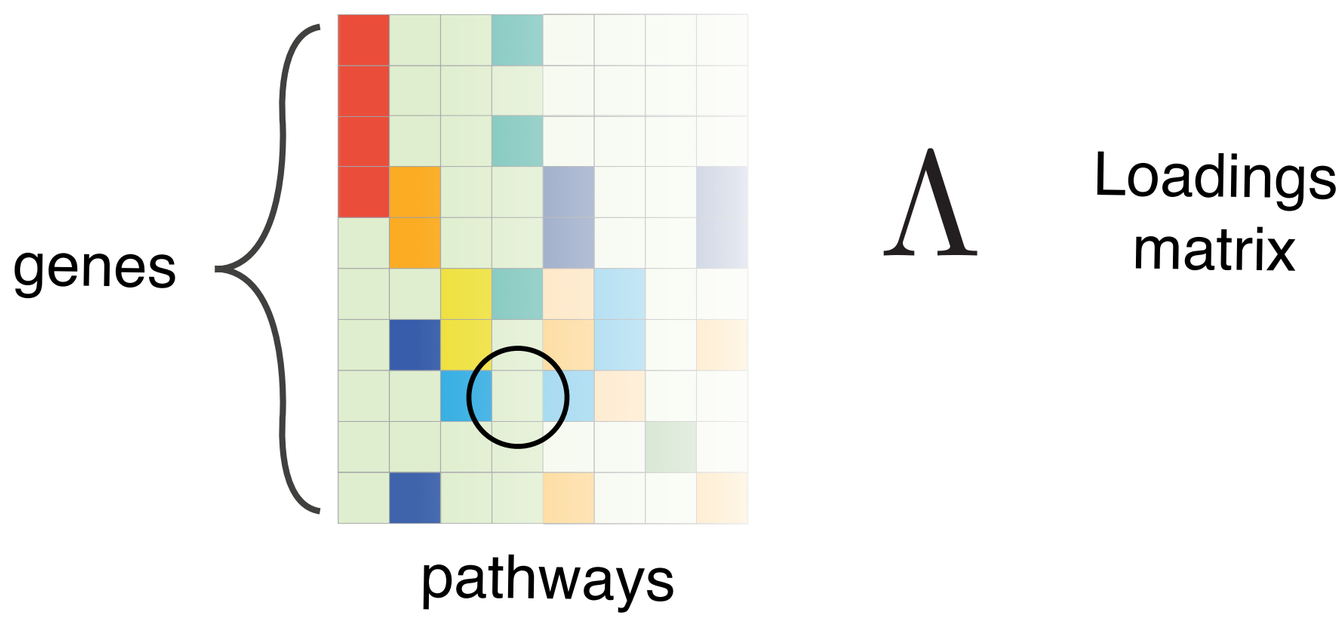


Few underlying pathways = few parameters to estimate

A factor model for G

Sparsity assumptions are key for high-dimensional data





Few underlying pathways = few parameters to estimate

Few effects per pathway = pathways are robust and interpretable

A factor model for G

genetic effects

measured traits

underlying traits

$$\mathbf{u}_i = \Lambda \mathbf{f}_i + \mathbf{e}_i$$

Loadings matrix

A factor model for G

genetic effects

measured traits

underlying traits

$$\mathbf{u}_i = \Lambda \mathbf{f}_i + \mathbf{e}_i$$

Loadings matrix

Residual covariance

$$\mathbf{G} = \Lambda\Lambda^T + \Sigma_{\mathbf{e}}$$

Genetic covariances

Bayesian genetic sparse factor model

Bayes' Theorem

$$\overbrace{p(\mathbf{G} \mid \mathbf{Y})}^{\text{Posterior}} = \frac{\overbrace{p(\mathbf{Y} \mid \mathbf{G})}^{\text{Likelihood}}\overbrace{\pi(\mathbf{G})}^{\text{Prior}}}{p(\mathbf{Y})}$$

Bayesian genetic sparse factor model

Bayes' Theorem

Posterior Likelihood Prior

$$p(\mathbf{G} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathbf{G})\pi(\mathbf{G})}{p(\mathbf{Y})}$$

Animal model likelihood

$$p(\mathbf{Y} \mid \mathbf{G}) \qquad \mathbf{y}_i \sim \mathrm{N}\left(\mathbf{x}_i\mathbf{b} + \mathbf{u}_i, \mathbf{R}\right)$$

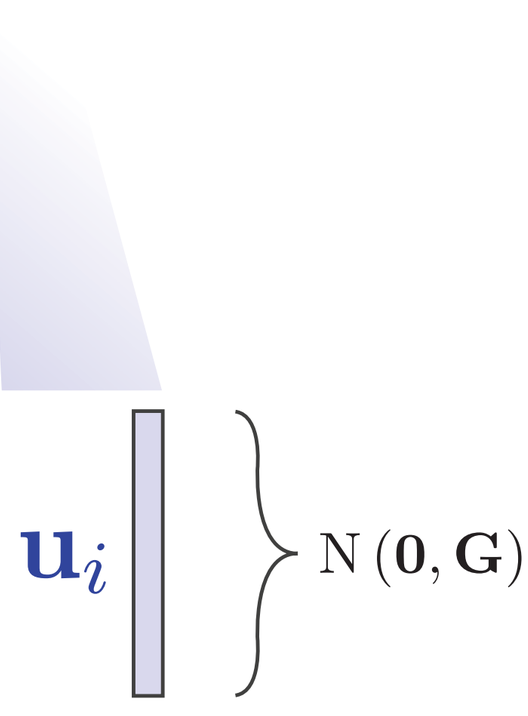
Bayesian genetic sparse factor model

Bayes' Theorem

$$\underbrace{p(\mathbf{G} \mid \mathbf{Y})}_{\text{Posterior}} = \frac{\overbrace{p(\mathbf{Y} \mid \mathbf{G})}^{\text{Likelihood}}\overbrace{\pi(\mathbf{G})}^{\text{Prior}}}{p(\mathbf{Y})}$$

Animal model likelihood

$$p(\mathbf{Y} \mid \mathbf{G}) \qquad \mathbf{y}_i \sim \mathrm{N}\left(\mathbf{x}_i\mathbf{b} + \mathbf{u}_i, \mathbf{R}\right)$$



Bayesian genetic sparse factor model

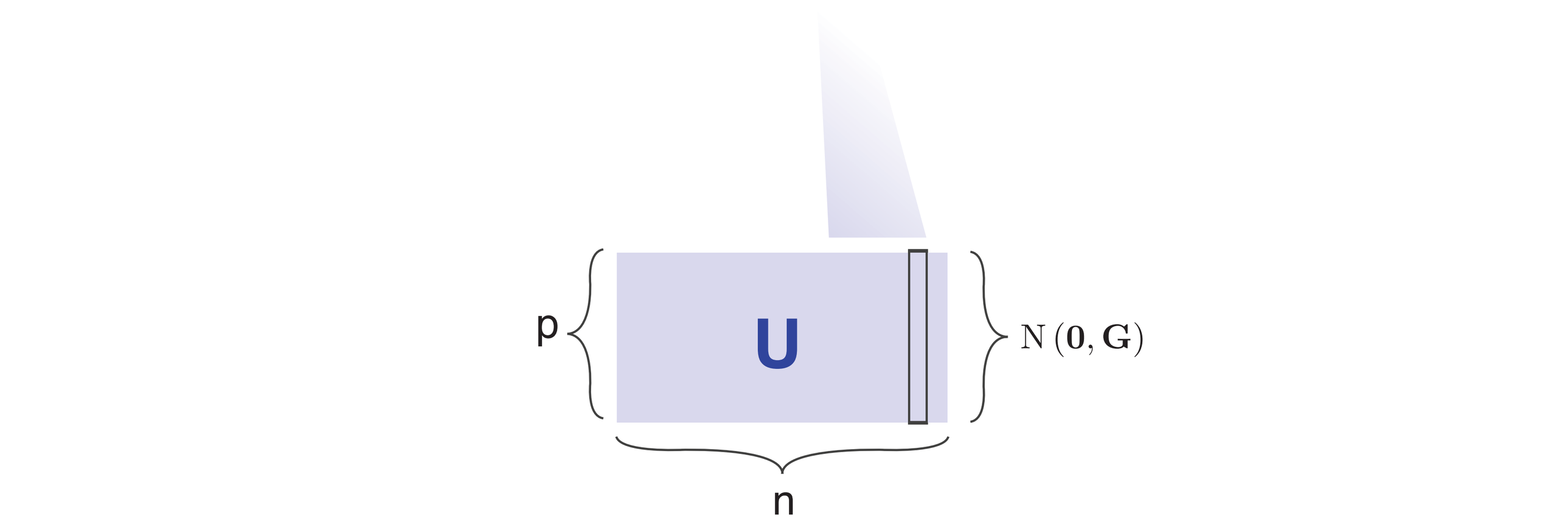
Bayes' Theorem

Posterior Likelihood Prior

$$p(\mathbf{G} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathbf{G})\pi(\mathbf{G})}{p(\mathbf{Y})}$$

Animal model likelihood

$$p(\mathbf{Y} \mid \mathbf{G}) \qquad \mathbf{y}_i \sim \mathrm{N}\left(\mathbf{x}_i\mathbf{b} + \mathbf{u}_i, \mathbf{R}\right)$$



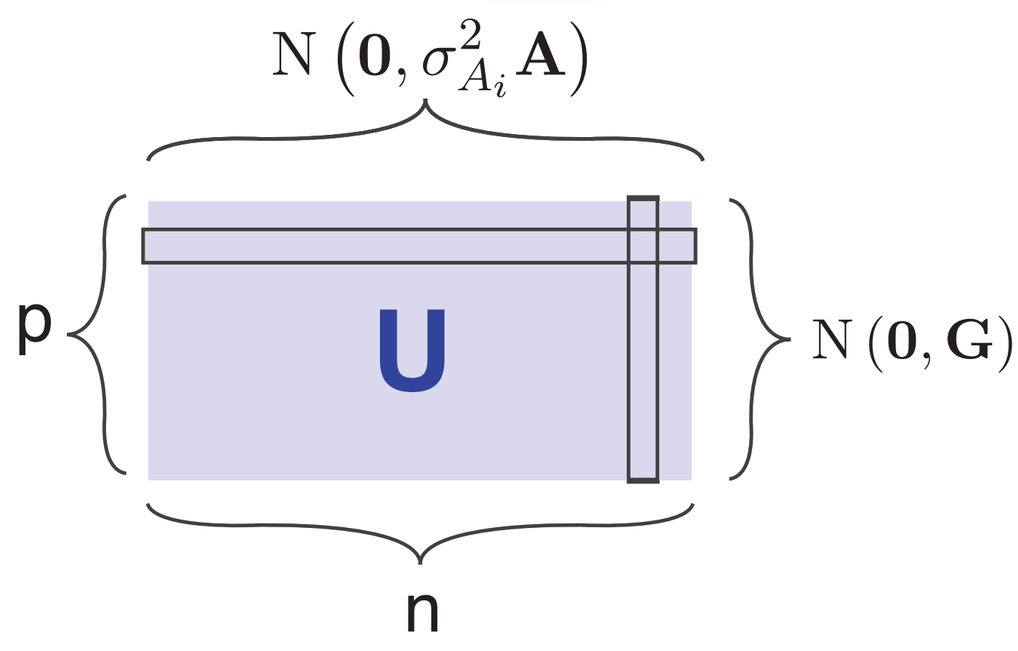
Bayesian genetic sparse factor model

Bayes' Theorem

$$\overbrace{p(\mathbf{G} \mid \mathbf{Y})}^{\text{Posterior}} = \frac{\overbrace{p(\mathbf{Y} \mid \mathbf{G})}^{\text{Likelihood}}\overbrace{\pi(\mathbf{G})}^{\text{Prior}}}{p(\mathbf{Y})}$$

Animal model likelihood

$$p(\mathbf{Y} \mid \mathbf{G}) \qquad \mathbf{y}_i \sim \mathrm{N}\left(\mathbf{x}_i\mathbf{b} + \mathbf{u}_i, \mathbf{R}\right)$$



The animal model – singe trait

For a single trait

$$\mathbf{y}_i = \mathbf{X}\mathbf{b}_i + \mathbf{Z}\mathbf{u}_i + \mathbf{e}_i,$$

The animal model – singe trait

For a single trait

$$\mathbf{y}_i = \mathbf{X}\mathbf{b}_i + \mathbf{Z}\mathbf{u}_i + \mathbf{e}_i,$$

$\mathbf{y}_i$: vector of measurements of trait i

The animal model – singe trait

For a single trait

$$\mathbf{y}_i = \mathbf{X}\mathbf{b}_i + \mathbf{Z}\mathbf{u}_i + \mathbf{e}_i,$$

$\mathbf{y}_i$: vector of measurements of trait i
$\mathbf{b}_i$: vector of fixed effects and environmental covariates of trait i

The animal model – singe trait

For a single trait

$$\mathbf{y}_i = \mathbf{X}\mathbf{b}_i + \mathbf{Z}\mathbf{u}_i + \mathbf{e}_i,$$

$\mathbf{y}_i$: vector of measurements of trait i
$\mathbf{b}_i$: vector of fixed effects and environmental covariates of trait i
$\mathbf{X}$: design matrix

The animal model – singe trait

For a single trait

$$\mathbf{y}_i = \mathbf{X}\mathbf{b}_i + \mathbf{Z}\mathbf{u}_i + \mathbf{e}_i,$$

$\mathbf{y}_i$: vector of measurements of trait i
$\mathbf{b}_i$: vector of fixed effects and environmental covariates of trait i
$\mathbf{X}$: design matrix
$\mathbf{u}_i$: random additive genetic effects with known covariance $\sigma_i^2\mathbf{A}$

The animal model – singe trait

For a single trait

$$\mathbf{y}_i = \mathbf{X}\mathbf{b}_i + \mathbf{Z}\mathbf{u}_i + \mathbf{e}_i,$$

$\mathbf{y}_i$: vector of measurements of trait i
$\mathbf{b}_i$: vector of fixed effects and environmental covariates of trait i
$\mathbf{X}$: design matrix
$\mathbf{u}_i$: random additive genetic effects with known covariance $\sigma_i^2\mathbf{A}$
$\mathbf{Z}$: relates random effects to observations

The animal model – singe trait

For a single trait

$$\mathbf{y}_i = \mathbf{X}\mathbf{b}_i + \mathbf{Z}\mathbf{u}_i + \mathbf{e}_i,$$

$\mathbf{y}_i$: vector of measurements of trait i
$\mathbf{b}_i$: vector of fixed effects and environmental covariates of trait i
$\mathbf{X}$: design matrix
$\mathbf{u}_i$: random additive genetic effects with known covariance $\sigma_i^2\mathbf{A}$
$\mathbf{Z}$: relates random effects to observations
$\mathbf{e}_i$: error independent to random effects.

The animal model – multiple traits

For p traits

$$\mathbf{Y} = \mathbf{XB} + \mathbf{ZU} + \mathbf{E},$$

The animal model – multiple traits

For p traits

$$\mathbf{Y} = \mathbf{XB} + \mathbf{ZU} + \mathbf{E},$$

$\mathbf{Y}$: $n \times p$ vector of measured traits

The animal model – multiple traits

For p traits

$$\mathbf{Y} = \mathbf{XB} + \mathbf{ZU} + \mathbf{E},$$

$\mathbf{Y}$: $n \times p$ vector of measured traits

$\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_p]$: $\quad \mathbf{U} \sim \mathrm{MN}_{r,p}(\mathbf{0}, \mathbf{A}, \mathbf{G})$

The animal model – multiple traits

For p traits

$$\mathbf{Y} = \mathbf{XB} + \mathbf{ZU} + \mathbf{E},$$

$\mathbf{Y}$: $n \times p$ vector of measured traits

$\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_p]$: $\quad \mathbf{U} \sim \mathrm{MN}_{r,p}(\mathbf{0}, \mathbf{A}, \mathbf{G})$

$\mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_p]$: $\quad \mathbf{E} \sim \mathrm{MN}_{n,p}(\mathbf{0}, \mathbf{I}_n, \mathbf{R})$

The animal model – multiple traits

For p traits

$$\mathbf{Y} = \mathbf{XB} + \mathbf{ZU} + \mathbf{E},$$

$\mathbf{Y}$: $n \times p$ vector of measured traits
$\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_p]$: $\quad \mathbf{U} \sim \mathrm{MN}_{r,p}(\mathbf{0}, \mathbf{A}, \mathbf{G})$
$\mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_p]$: $\quad \mathbf{E} \sim \mathrm{MN}_{n,p}(\mathbf{0}, \mathbf{I}_n, \mathbf{R})$

$\mathbf{G}$: $p \times p$ genetic covariance matrix
$\mathbf{A}$: rank $r \leq n$ matrix of relatedness

The animal model – multiple traits

For p traits

$$\mathbf{Y} = \mathbf{XB} + \mathbf{ZU} + \mathbf{E},$$

$\mathbf{Y}$: $n \times p$ vector of measured traits
$\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_p]$: $\quad \mathbf{U} \sim \mathrm{MN}_{r,p}(\mathbf{0}, \mathbf{A}, \mathbf{G})$
$\mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_p]$: $\quad \mathbf{E} \sim \mathrm{MN}_{n,p}(\mathbf{0}, \mathbf{I}_n, \mathbf{R})$

$\mathbf{G}$: $p \times p$ genetic covariance matrix
$\mathbf{A}$: rank $r \leq n$ matrix of relatedness
$\mathbf{E}$: $p \times p$ residual covariance matrix

The animal model – multiple traits

For p traits

$$\mathbf{Y} = \mathbf{XB} + \mathbf{ZU} + \mathbf{E},$$

$\mathbf{Y}$: $n \times p$ vector of measured traits
$\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_p]$: $\quad \mathbf{U} \sim \mathrm{MN}_{r,p}(\mathbf{0}, \mathbf{A}, \mathbf{G})$
$\mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_p]$: $\quad \mathbf{E} \sim \mathrm{MN}_{n,p}(\mathbf{0}, \mathbf{I}_n, \mathbf{R})$

$\mathbf{G}$: $p \times p$ genetic covariance matrix
$\mathbf{A}$: rank $r \leq n$ matrix of relatedness
$\mathbf{E}$: $p \times p$ residual covariance matrix

$$p(\mathbf{V} \mid \mathbf{M}, \mathbf{\Omega}, \mathbf{\Sigma}) = \frac{\exp\left(-\frac{1}{2}\mathrm{tr}[\mathbf{\Omega}^{-1/2}(\mathbf{V} - \mathbf{M})^T \mathbf{\Sigma}^{-1}(\mathbf{V} - \mathbf{M})]\right)}{(2\pi)^{np/2}|\mathbf{\Omega}|^{n/2}|\mathbf{\Sigma}|^{p/2}}.$$

Hierarchical factor model

Model k latent traits that linearly relate to observed traits.

Specification of $\mathbf{U}$ and $\mathbf{E}$.

$$
\begin{gathered}
\mathbf{U} = \mathbf{F}_a \boldsymbol{\Lambda}^T + \boldsymbol{\Delta}, \quad \mathbf{E} = \mathbf{F}_e \boldsymbol{\Lambda}^T + \boldsymbol{\Xi} \\
\mathbf{F}_a \sim \mathrm{MN}_{r,k}(\mathbf{0}, \mathbf{A}, \boldsymbol{\Sigma}_a), \quad \mathbf{F}_e \sim \mathrm{MN}_{n,k}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Sigma}_e) \\
\boldsymbol{\Delta} \sim \mathrm{MN}_{r,p}(\mathbf{0}, \mathbf{A}, \boldsymbol{\Psi}_a), \quad \boldsymbol{\Xi} \sim \mathrm{MN}_{n,p}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi}_e) \\
\boldsymbol{\Lambda} \sim \pi(\theta),
\end{gathered}
\tag{1}
$$

Partition of variation and heritability

$\mathbf{F}_a$ and $\mathbf{F}_e$ among-individual variation in the latent traits.
$\mathbf{\Sigma}_a$ and $\mathbf{\Sigma}_e$ model within individual covariance of the factors:

$$\mathbf{\Sigma}_a = \text{Diag}(\sigma^2_{a_j}), \mathbf{\Sigma}_e = \text{Diag}(\sigma^2_{e_j}).$$

Partition of variation and heritability

$\mathbf{F}_a$ and $\mathbf{F}_e$ among-individual variation in the latent traits.
$\mathbf{\Sigma}_a$ and $\mathbf{\Sigma}_e$ model within individual covariance of the factors:

$$\mathbf{\Sigma}_a = \text{Diag}(\sigma^2_{a_j}), \mathbf{\Sigma}_e = \text{Diag}(\sigma^2_{e_j}).$$

$\mathbf{\Lambda}$ is not identifiable without constraints (rotation and scaling).
Column variances sum to one

$$\mathbf{\Sigma}_a + \mathbf{\Sigma}_e = \mathbf{I}_k, \quad \mathbf{\Sigma}_{h^2} = \mathbf{\Sigma}_a = \mathbf{I}_k - \mathbf{\Sigma}_e$$

Narrow sense heritability

$$h^2_j = \frac{\sigma^2_{a_j}}{\sigma^2_{a_j} + \sigma^2_{e_j}} = \sigma^2_{a_j}.$$

Partition of variance by factors

Recovering **G** and **R**

$$\begin{aligned} \mathbf{G} &= \mathbf{\Lambda}\mathbf{\Sigma}_{h^2}\mathbf{\Lambda}^T + \mathbf{\Psi}_a, \\ \mathbf{R} &= \mathbf{\Lambda}(\mathbf{I}_k - \mathbf{\Sigma}_{h^2})\mathbf{\Lambda}^T + \mathbf{\Psi}_e. \end{aligned} \tag{2}$$

Partition of variance by factors

Recovering **G** and **R**

$$\begin{aligned} \mathbf{G} &= \mathbf{\Lambda}\mathbf{\Sigma}_{h^2}\mathbf{\Lambda}^T + \mathbf{\Psi}_a, \\ \mathbf{R} &= \mathbf{\Lambda}(\mathbf{I}_k - \mathbf{\Sigma}_{h^2})\mathbf{\Lambda}^T + \mathbf{\Psi}_e. \end{aligned} \tag{2}$$

Total phenotypic covariance $\mathbf{P} = \mathbf{G} + \mathbf{R}$:

$$\mathbf{P} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}_a + \mathbf{\Psi}_e. \tag{3}$$

Constraints of **G** and **R**

Informative priors on covariance matrices

(1) Limited number of pathways are relevant for trait variation or number of factors is low, $k \ll p$.

Bayesian genetic sparse factor model

Bhattachyra and Dunson (2011) *Sparse Bayesian infinite factor models*

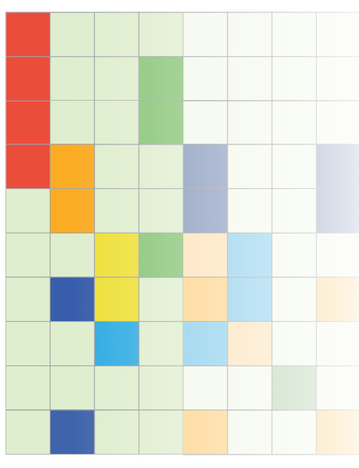
$$\pi(\mathbf{G}) = \pi\left(\Lambda\Lambda^T + \Sigma_{\mathbf{e}}\right)$$

Bayesian genetic sparse factor model

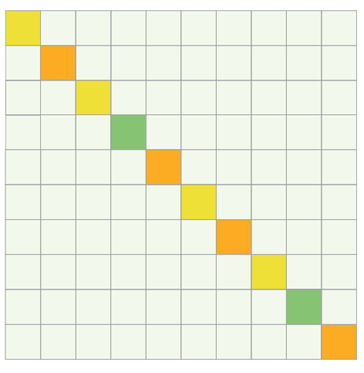
Bhattachyra and Dunson (2011) *Sparse Bayesian infinite factor models*

$$\pi(\mathbf{G}) = \pi\left(\Lambda\Lambda^T + \Sigma_{\mathbf{e}}\right)$$

Λ



$\Sigma_{\mathbf{e}}$

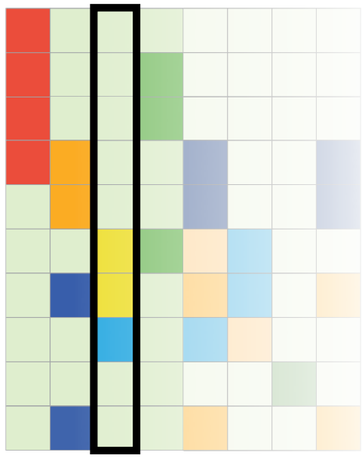


Bayesian genetic sparse factor model

Bhattachyra and Dunson (2011) *Sparse Bayesian infinite factor models*

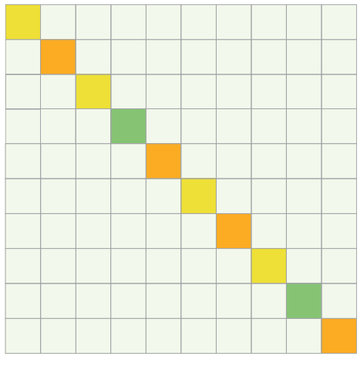
$$\pi(\mathbf{G}) = \pi\left(\Lambda\Lambda^T + \Sigma_{\mathbf{e}}\right)$$

Λ



factors

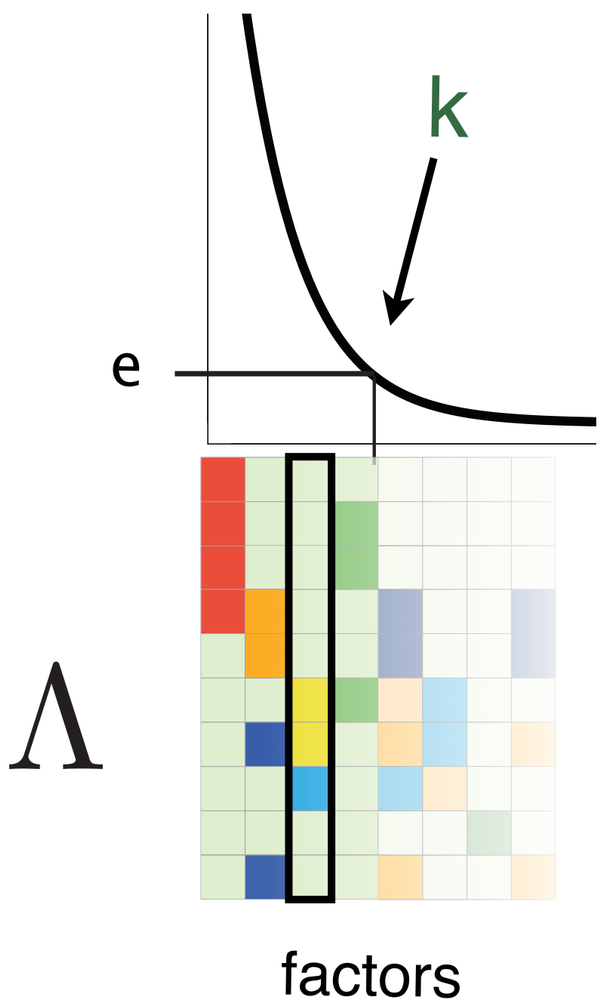
$\Sigma_{\mathbf{e}}$

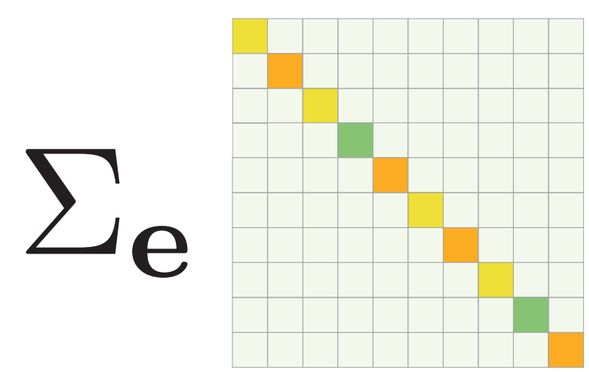


Bayesian genetic sparse factor model

Bhattachyra and Dunson (2011) *Sparse Bayesian infinite factor models*

$$\pi(\mathbf{G}) = \pi\left(\Lambda\Lambda^T + \Sigma_{\mathbf{e}}\right)$$

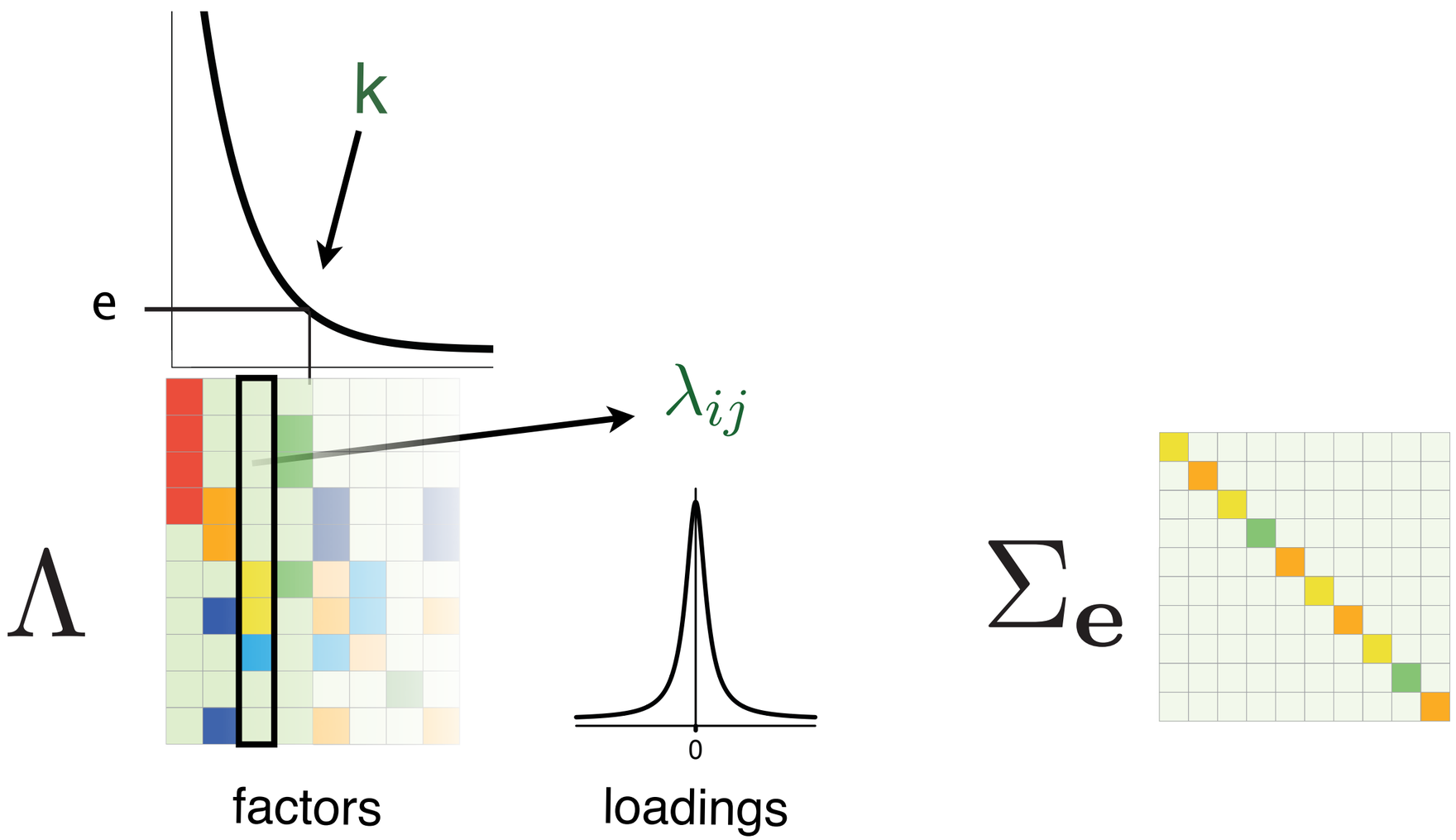




Bayesian genetic sparse factor model

Bhattachyra and Dunson (2011) *Sparse Bayesian infinite factor models*

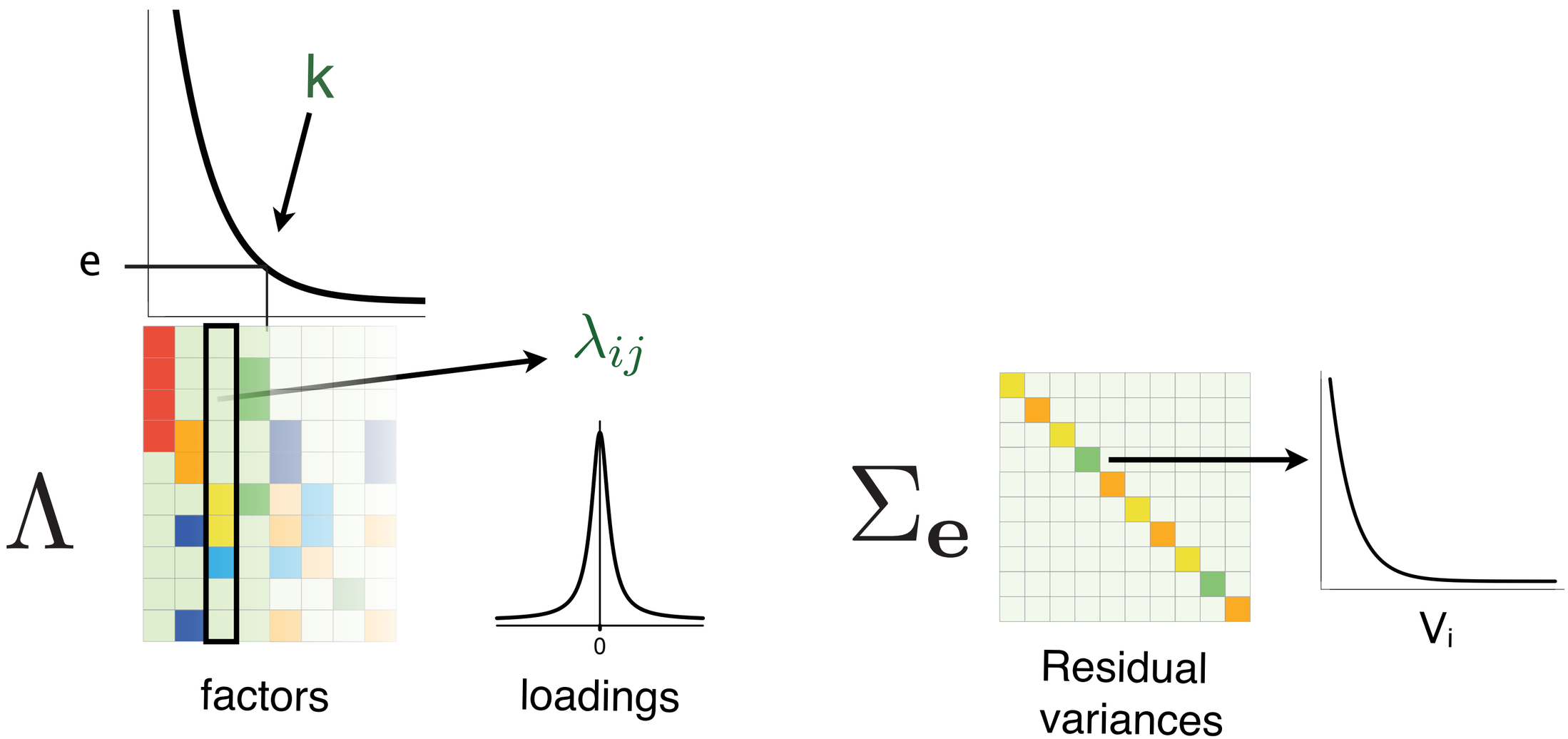
$$\pi(\mathbf{G}) = \pi\left(\Lambda\Lambda^T + \Sigma_{\mathbf{e}}\right)$$



Bayesian genetic sparse factor model

Bhattachyra and Dunson (2011) *Sparse Bayesian infinite factor models*

$$\pi(\mathbf{G}) = \pi\left(\Lambda\Lambda^T + \Sigma_{\mathbf{e}}\right)$$



Prior specification on $\boldsymbol{\Lambda}$

Based on (Bhattacharya and Dunson, 2011)

$$\lambda_{im} \mid \phi_{im}, \tau_m \sim \mathrm{N}\left(0, \phi_{im}^{-1}\tau_m^{-1}\right)$$
$$\phi_{im} \sim \mathrm{Ga}(\nu/2, \nu/2),$$
$$\tau_m = \prod_{\ell=1}^{m} \delta_\ell,$$
$$\delta_1 \sim \mathrm{Ga}(a_1, b_1),$$
$$\delta_\ell \sim \mathrm{Ga}(a_2, b_2) \text{ for } \ell = 2, ..., k.$$

Heritability prior (Zhou and Stephens, pers. comm.)

$$\pi(h_i^2 = \ell/n_h) = 1/n_h, \text{ where } \ell = 0 \ldots (n_h - 1).$$

Prior specification on $\mathbf{\Lambda}$

Based on (Bhattacharya and Dunson, 2011)

$$\lambda_{im} \mid \phi_{im}, \tau_m \sim \mathrm{N}\left(0, \phi_{im}^{-1}\tau_m^{-1}\right)$$
$$\phi_{im} \sim \mathrm{Ga}(\nu/2, \nu/2),$$
$$\tau_m = \prod_{\ell=1}^{m} \delta_\ell,$$
$$\delta_1 \sim \mathrm{Ga}(a_1, b_1),$$
$$\delta_\ell \sim \mathrm{Ga}(a_2, b_2) \text{ for } \ell = 2, ..., k.$$

Heritability prior (Zhou and Stephens, pers. comm.)

$$\pi(h_i^2 = \ell/n_h) = 1/n_h, \text{ where } \ell = 0 \ldots (n_h - 1).$$

Remaining variables: $\mathbf{B}_{ij} \sim \mathrm{N}(0, \sigma^2 > 10^6)$, $\quad (\Psi_u, \Psi_e) \sim \mathrm{IG}$.

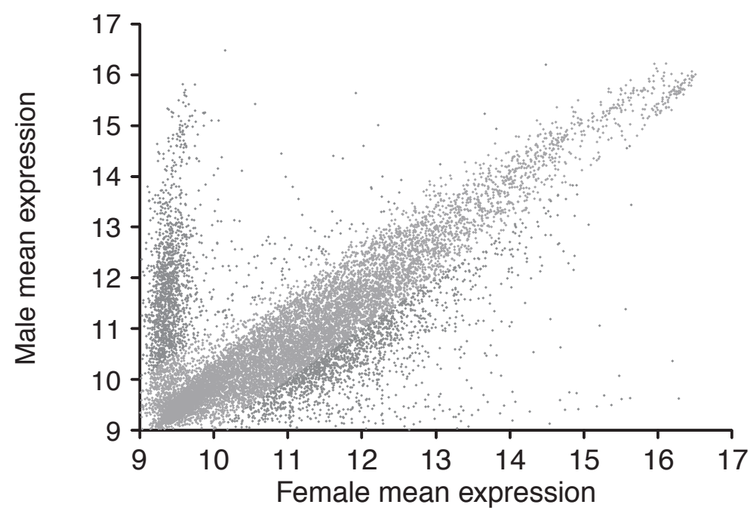
Case study: *Drosophila* gene expression

As a demonstration, we collected gene expression from:

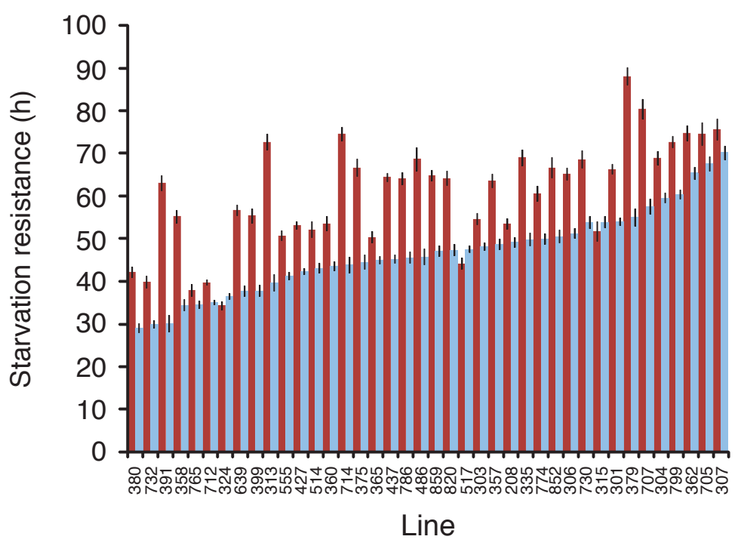
Ayroles et al (2009) Systems genetics of complex traits in *Drosophila melanogaster*. Nat Genet, 41, 299–307.

40 lines of *D. melanogaster*

DGRP



gene expression of >10,000 genes



Phenotype data on 7 fitness-related traits

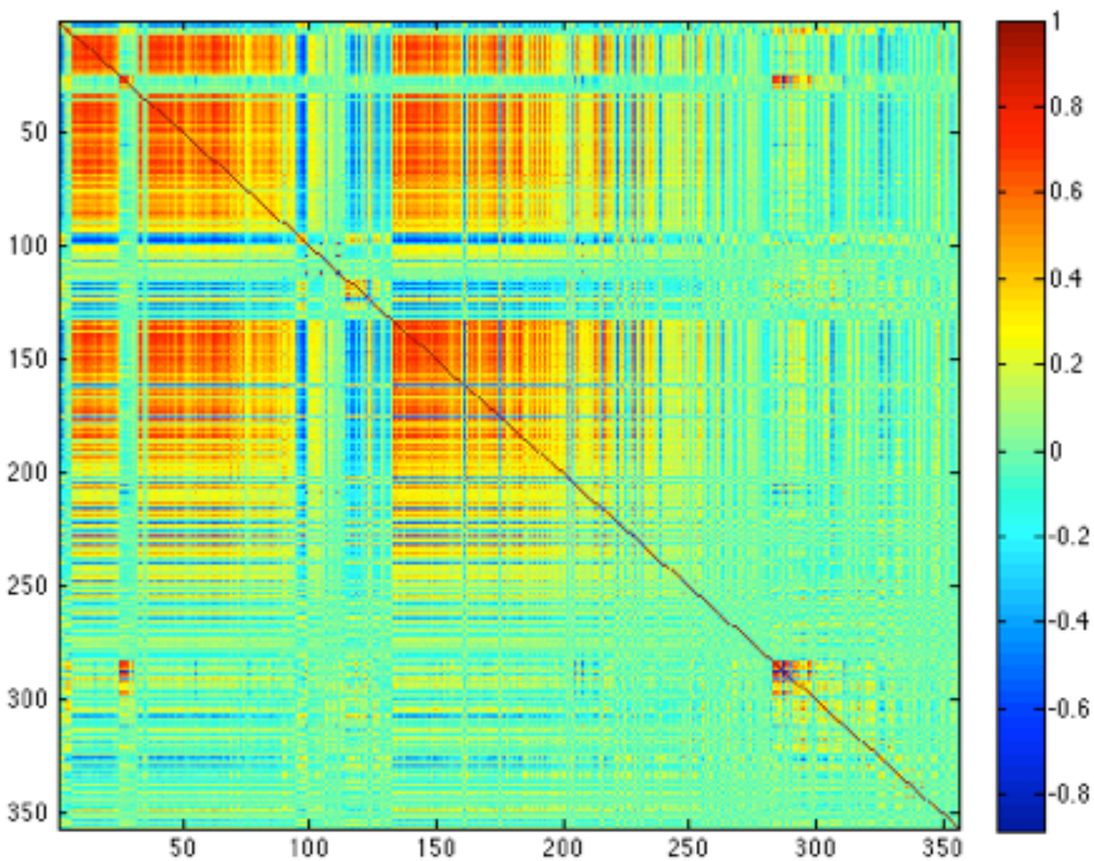
Case study: *Drosophila* gene expression

We ran our genetic factor model on the 355 genes correlated with Starvation Resistance

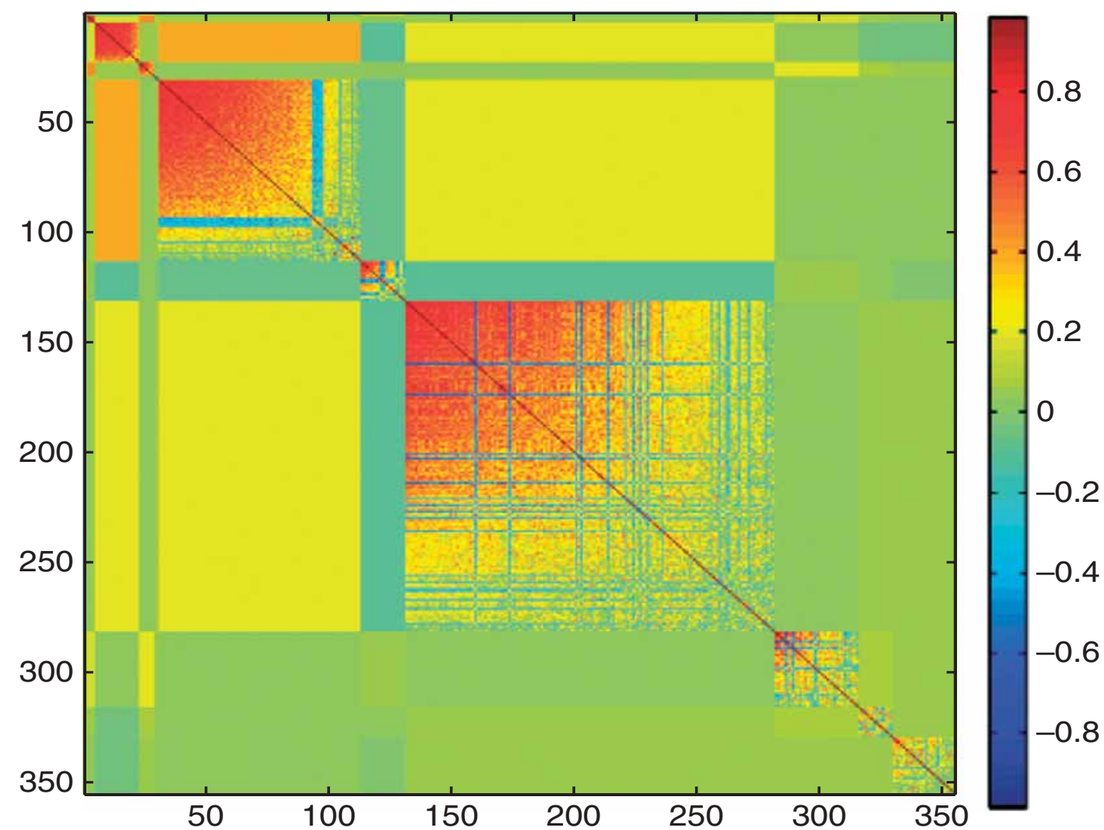
Case study: *Drosophila* gene expression

We ran our genetic factor model on the 355 genes correlated with Starvation Resistance

Our estimate



Ayroles *et al.* 2009

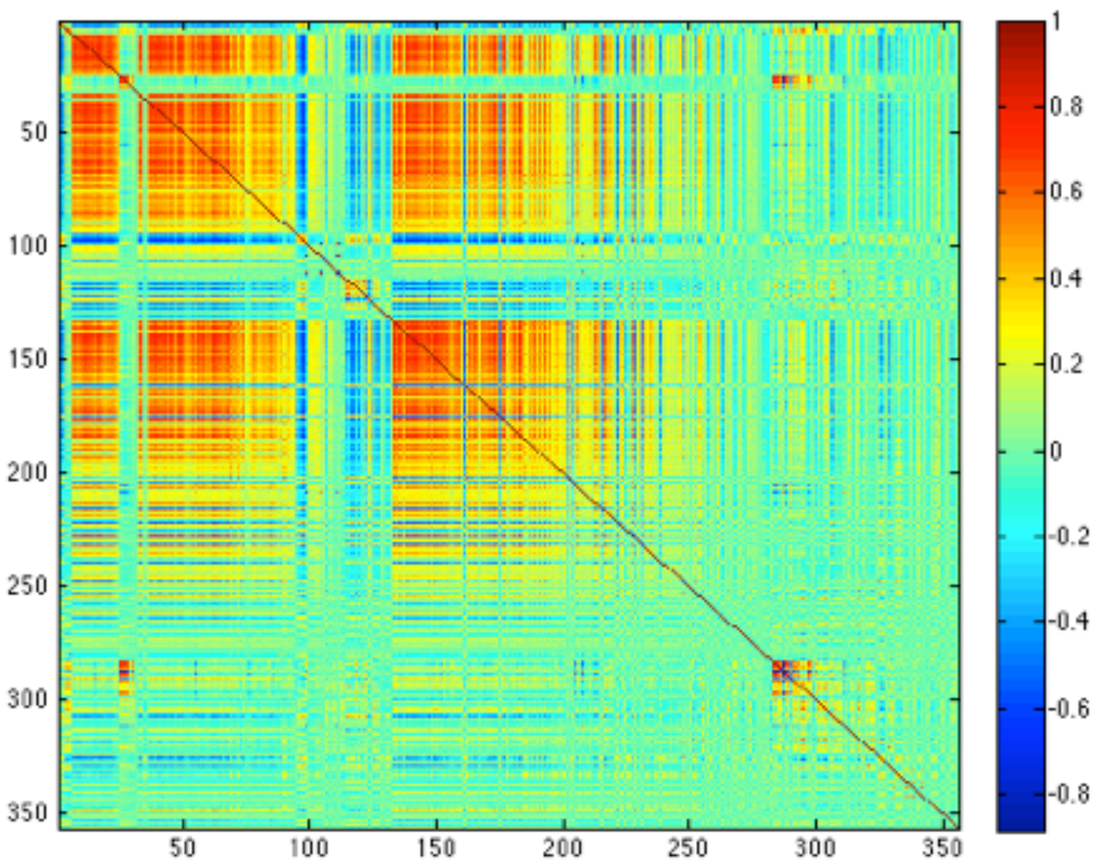


Our estimate of genetic (line) correlations is very similar to the estimate by Ayroles *et al.*

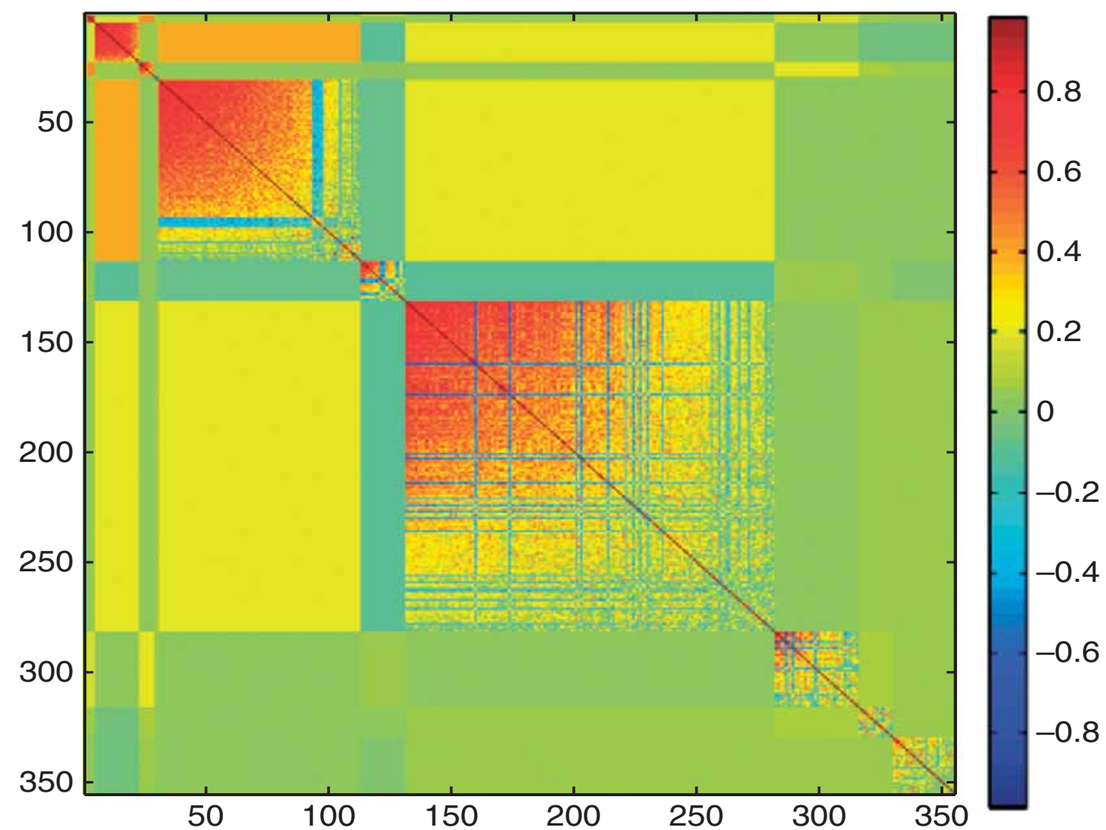
Case study: *Drosophila* gene expression

We ran our genetic factor model on the 355 genes correlated with Starvation Resistance

Our estimate



Ayroles *et al.* 2009



Our estimate of genetic (line) correlations is very similar to the estimate by Ayroles *et al.*

We fit >60,000 covariances with fewer than 4,000 parameters

We recover a true covariance matrix

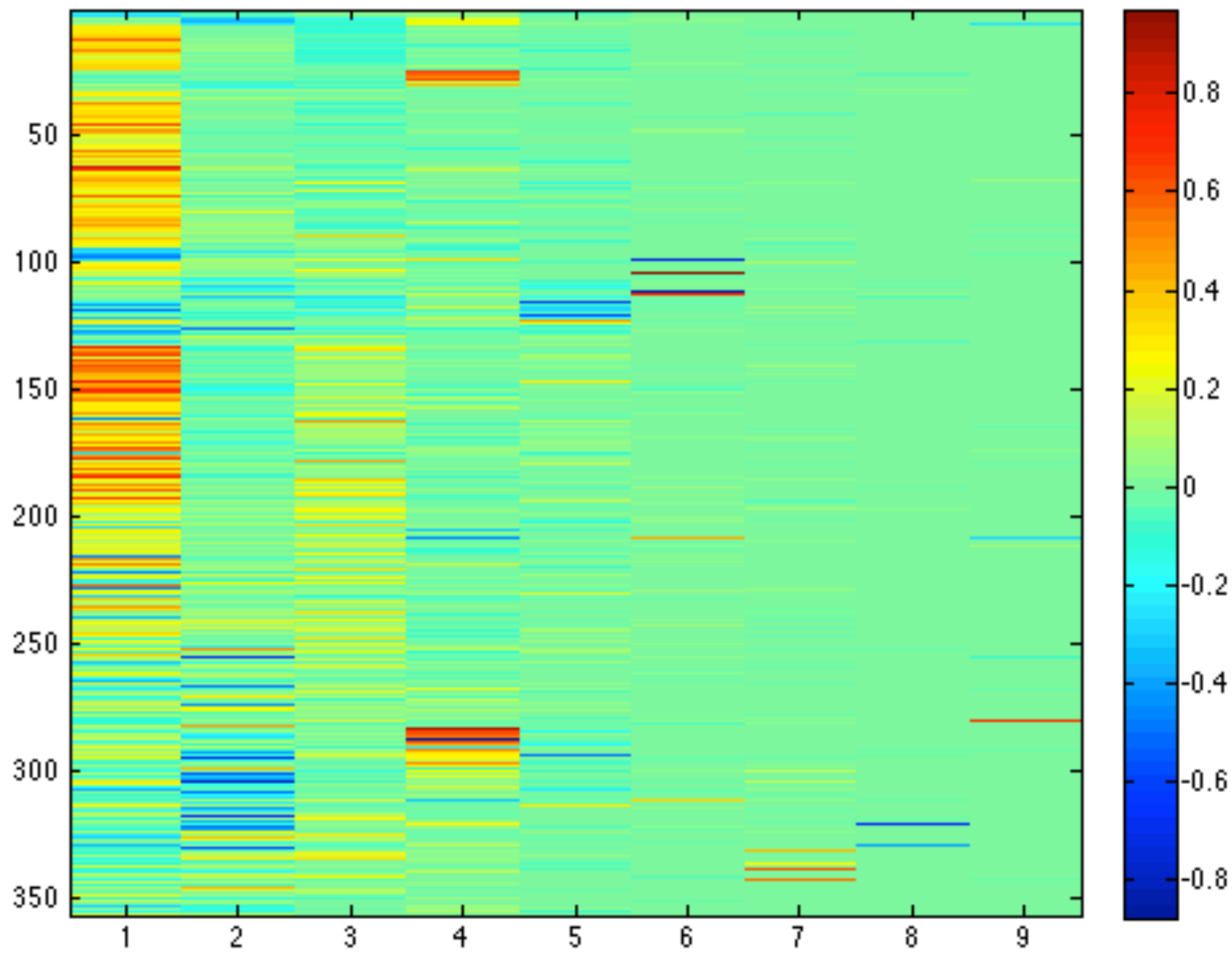
Case study: *Drosophila* gene expression

We estimate that the genetic covariation in expression could be explained by 9 factors

Factor 1 is dense but the remainder are very sparse.

Λ

Loadings matrix



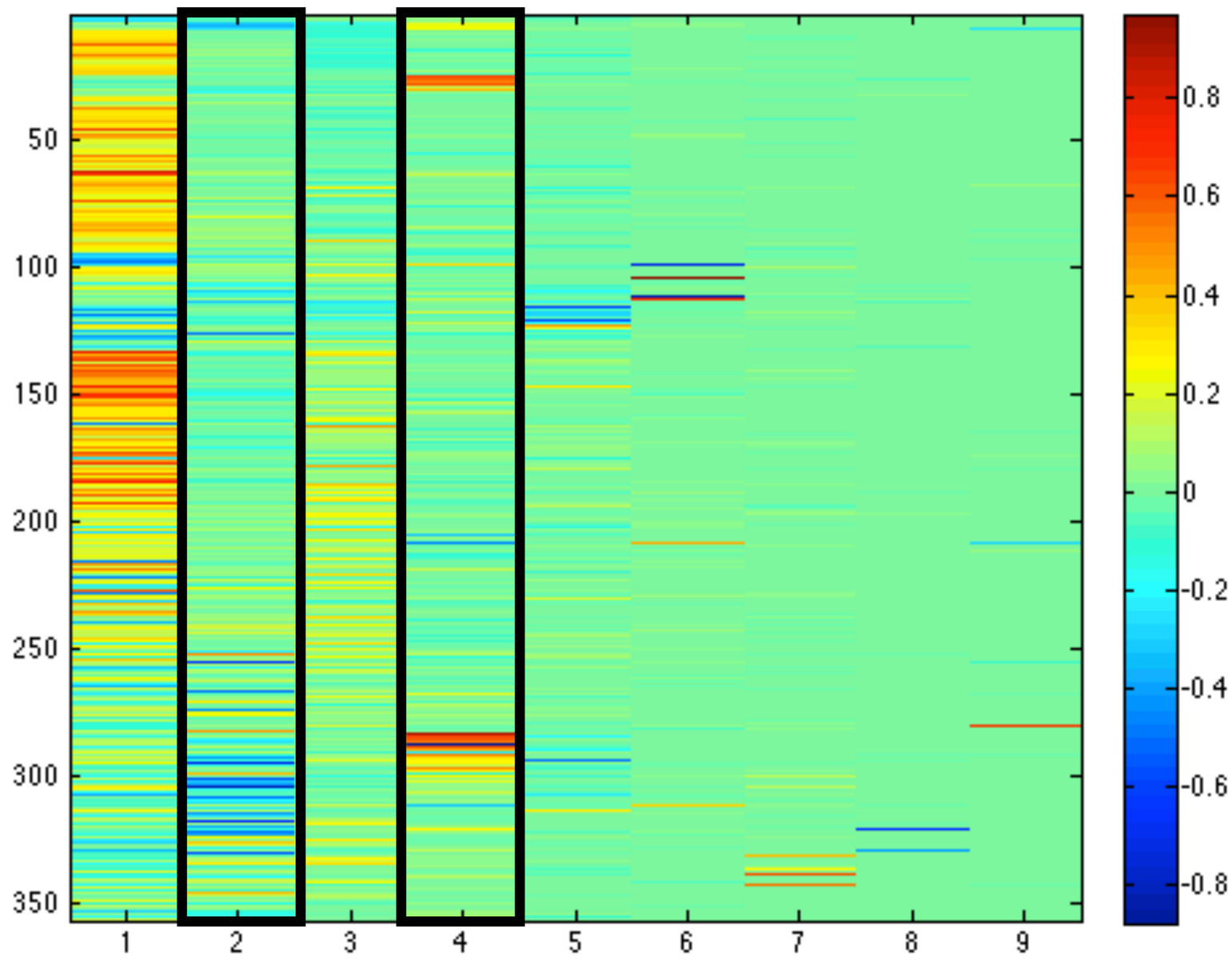
Case study: *Drosophila* gene expression

We estimate that the genetic covariation in expression could be explained by 9 factors

Factor 1 is dense but the remainder are very sparse.

Λ

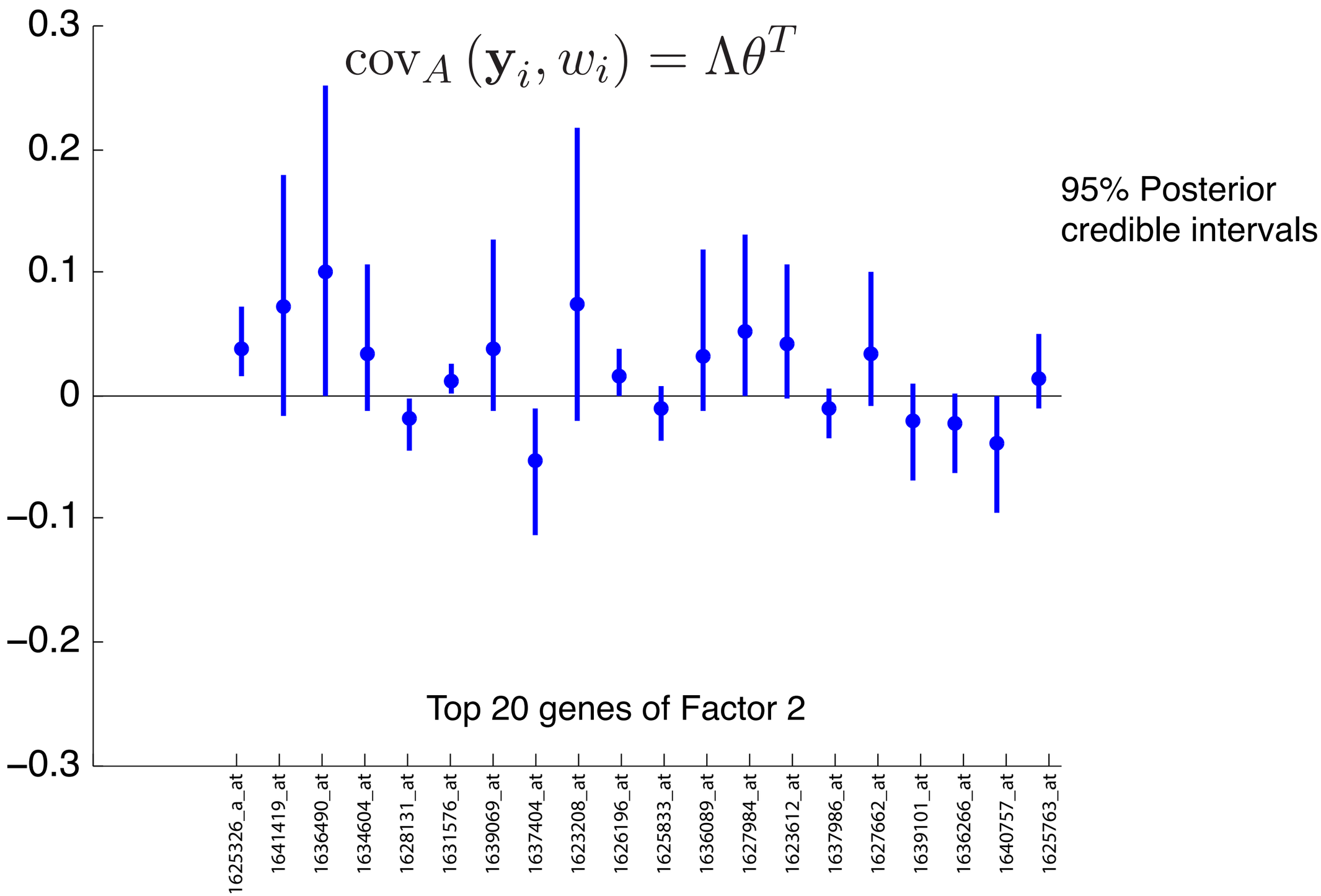
Loadings matrix



Genes related to defense and immune responses

Case study: *Drosophila* gene expression

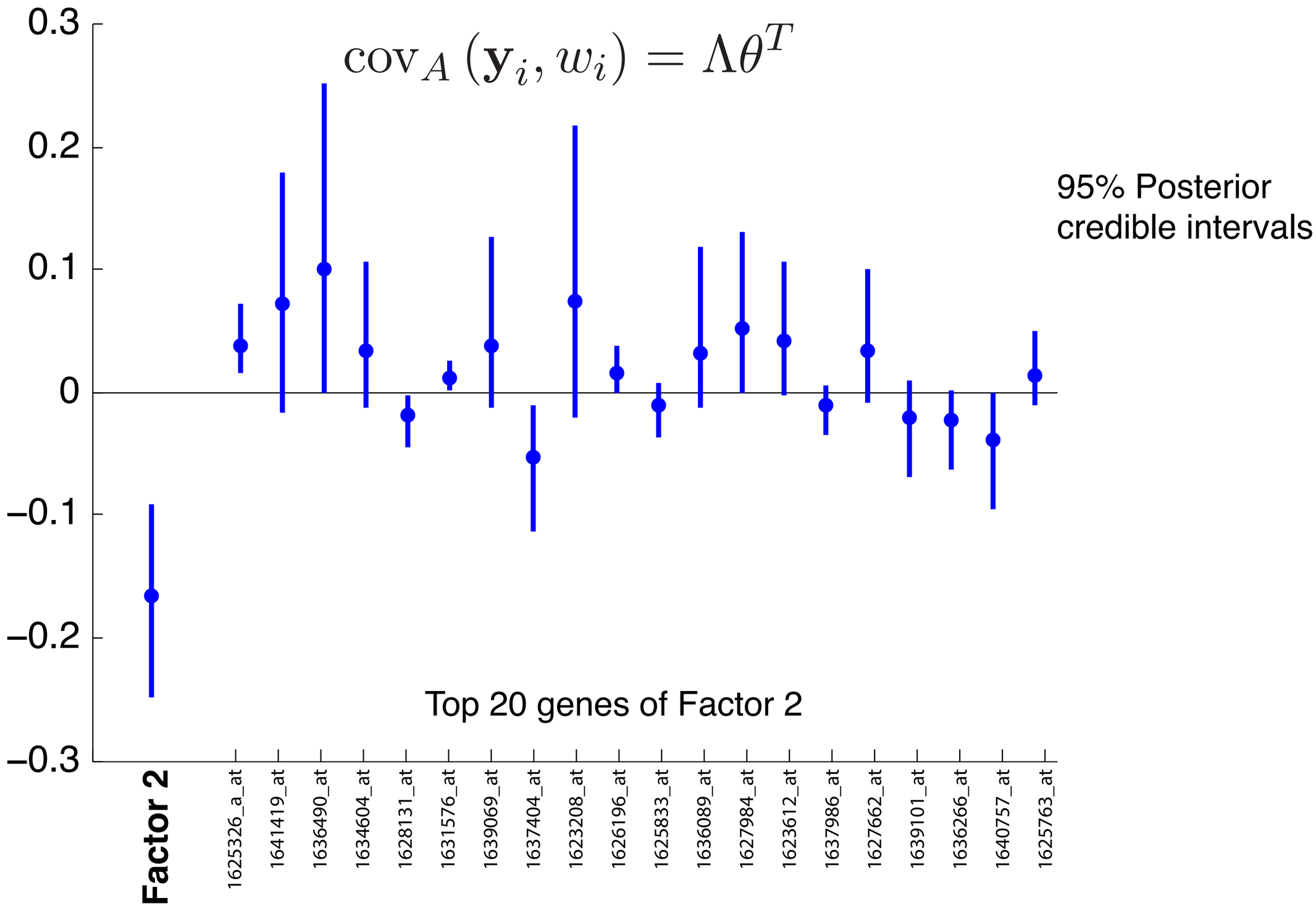
We can measure genetic covariances with Starvation Resistance



Case study: *Drosophila* gene expression

We can measure genetic covariances with Starvation Resistance

But have more power to identify covariances with underlying traits



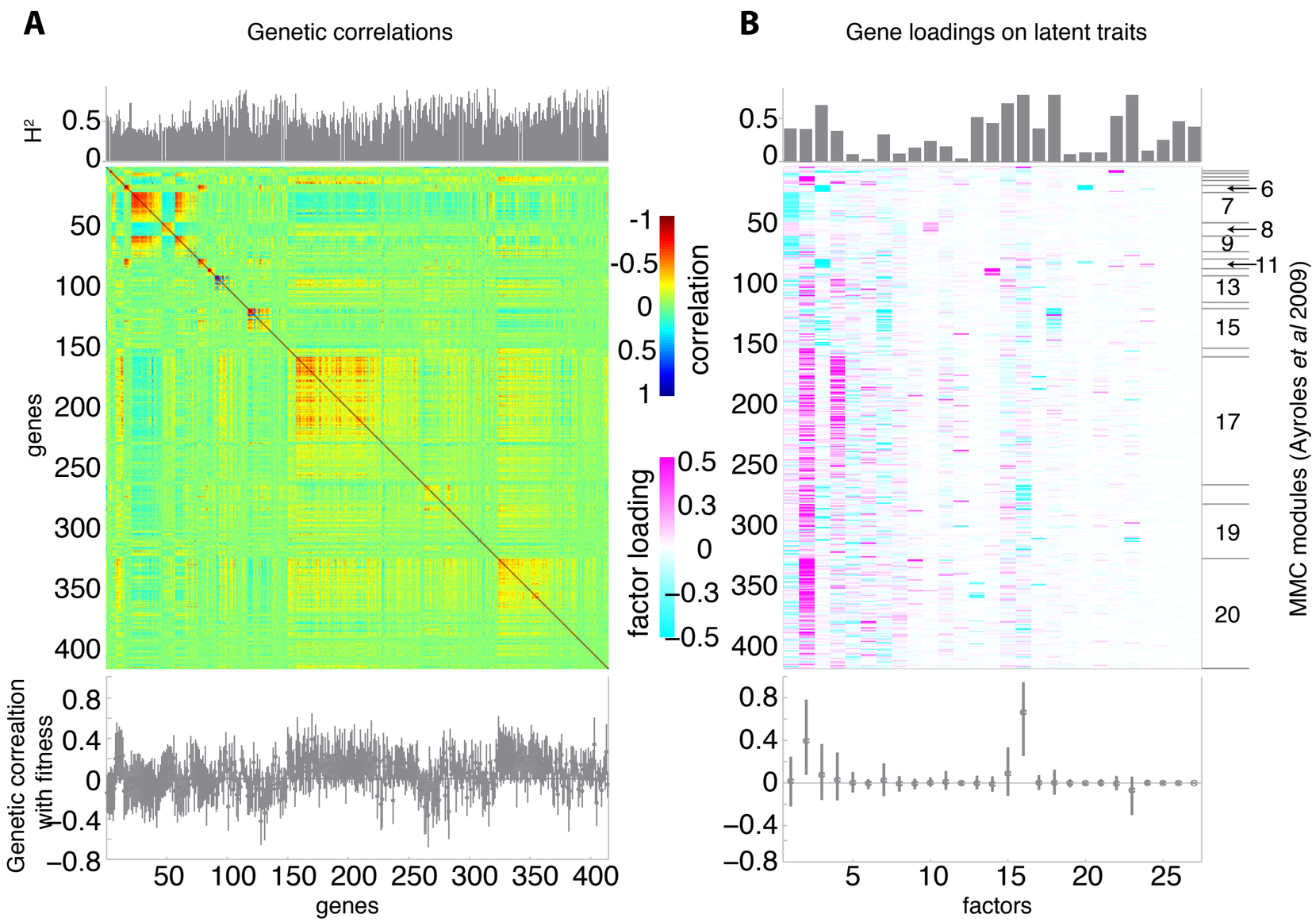
Drosophila melanogaster

Expression profiles of 414 genes and measures of competitive fitness of 40 wild-derived lines of *Drosophila melanogaster* from Ayroles et. al. 2009.

Ccompetitive fitness – percentage of offspring bearing a line's genotype given original proportion of the line.

Fixed effect of sex and random effects of the sex:line interaction were modeled.

Drosophila results



Software and paper

(1) Software:
http://www.stat.duke.edu/~sayan/bfgr/index.shtml

Software and paper

(1) Software:
http://www.stat.duke.edu/~sayan/bfgr/index.shtml

(2) Paper:
Dissecting High-Dimensional Phenotypes with Bayesian Sparse Factor Analysis of Genetic Covariance Matrices , Runcie and Mukherjee, Genetics, **194**:3, 753–767.

Extensions and open problems

(1) Problems estimating the number of factors.

Extensions and open problems

(1) Problems estimating the number of factors.

(2) Develop distance metrics for covariance matrices or subspaces.

Extensions and open problems

(1) Problems estimating the number of factors.

(2) Develop distance metrics for covariance matrices or subspaces.

(3) Response to selection

$$\Delta \bar{\mathbf{y}} = \Lambda_{u_{/p^*}} \Lambda_{u_{p^*}}^T .$$

Extensions and open problems

(1) Problems estimating the number of factors.

(2) Develop distance metrics for covariance matrices or subspaces.

(3) Response to selection

$$\Delta \bar{\mathbf{y}} = \Lambda_{u_{/p^*}} \Lambda_{u_{p^*}}^T .$$

(4) Percentage genetic variation in fitness by measured traits

$$1 - \Psi_{u_{p^*}} / \mathbf{G}_{p^*,p^*} .$$

Extensions and open problems

(1) Problems estimating the number of factors.

(2) Develop distance metrics for covariance matrices or subspaces.

(3) Response to selection

$$\Delta \bar{\mathbf{y}} = \Lambda_{u_{/p^*}} \Lambda_{u_{p^*}}^T .$$

(4) Percentage genetic variation in fitness by measured traits

$$1 - \Psi_{u_{p^*}} / \mathbf{G}_{p^*,p^*} .$$

(5) Incorporation with GWAS.

Extensions and open problems

(1) Problems estimating the number of factors.

(2) Develop distance metrics for covariance matrices or subspaces.

(3) Response to selection

$$\Delta \bar{\mathbf{y}} = \Lambda_{u_{/p^*}} \Lambda_{u_{p^*}}^T.$$

(4) Percentage genetic variation in fitness by measured traits

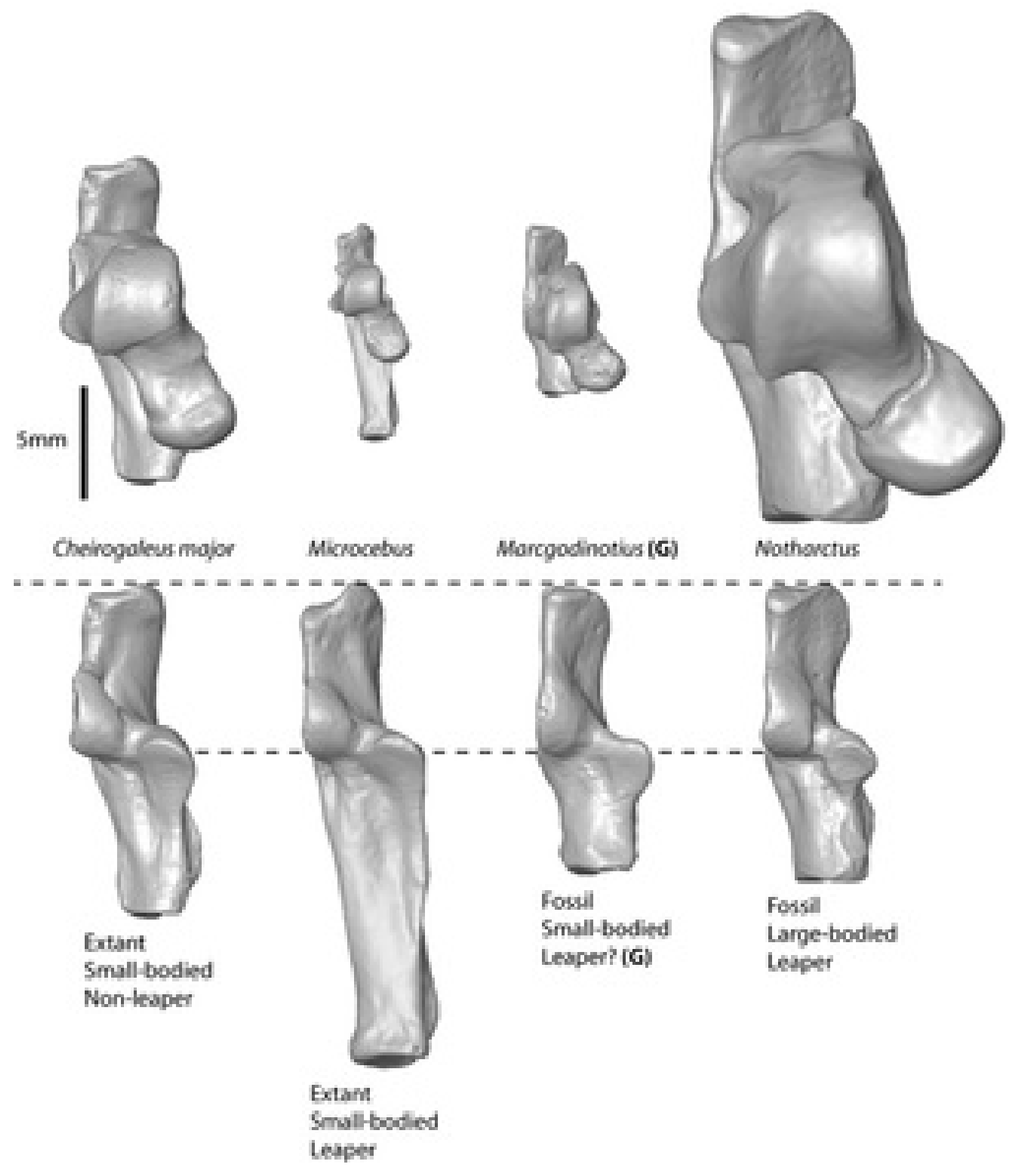
$$1 - \Psi_{u_{p^*}} / \mathbf{G}_{p^*,p^*}.$$

(5) Incorporation with GWAS.

(6) Discrete traits and time varying traits.

Homology on homology

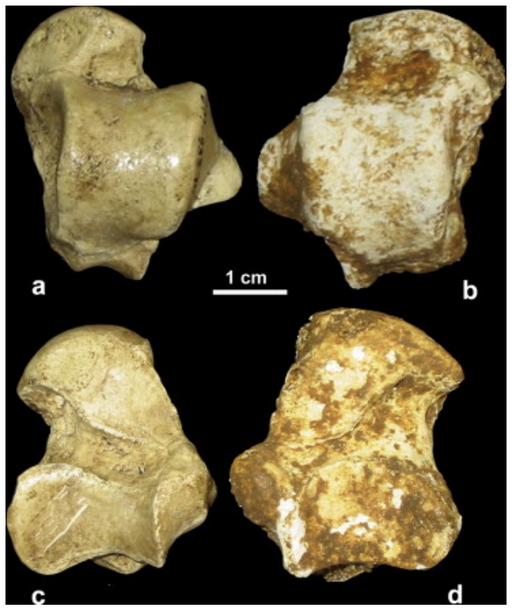
Morphology



From D. Boyer.

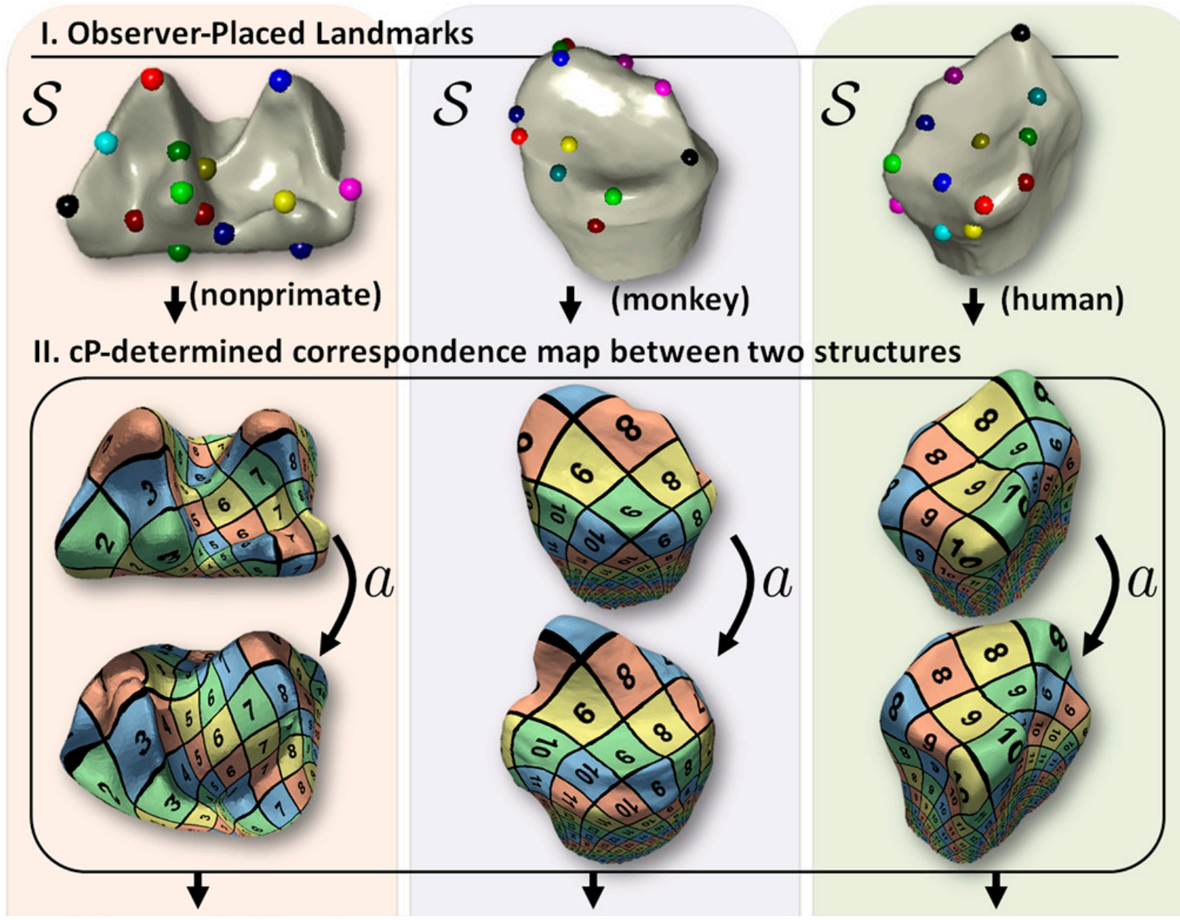
Morphology

Distance between heel bones across primates for evolutionary analysis.



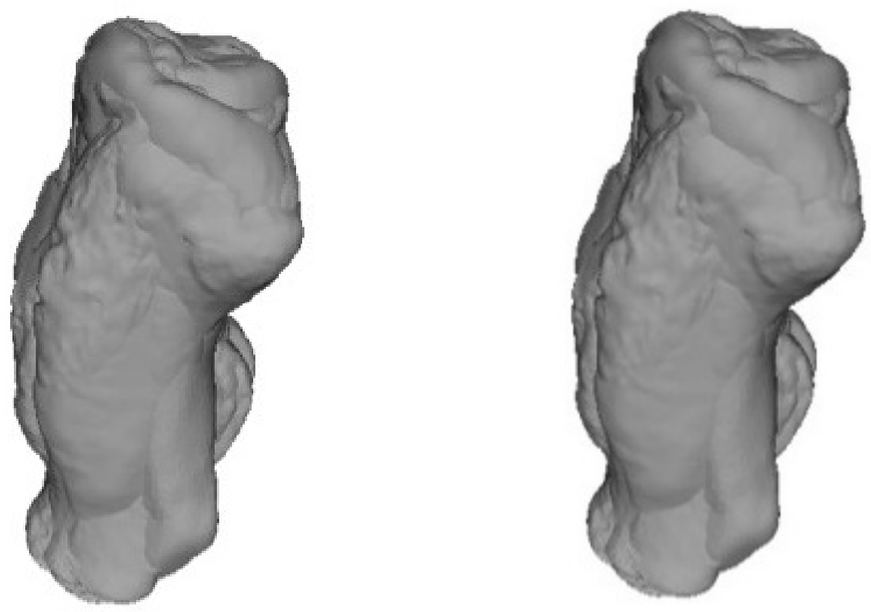
Algorithms to automatically quantify the geometric similarity of anatomical surfaces, Boyer et. al. PNAS 2011.

Geometric algorithm



Topological methods

What happens when the shapes are not isomorphic ?



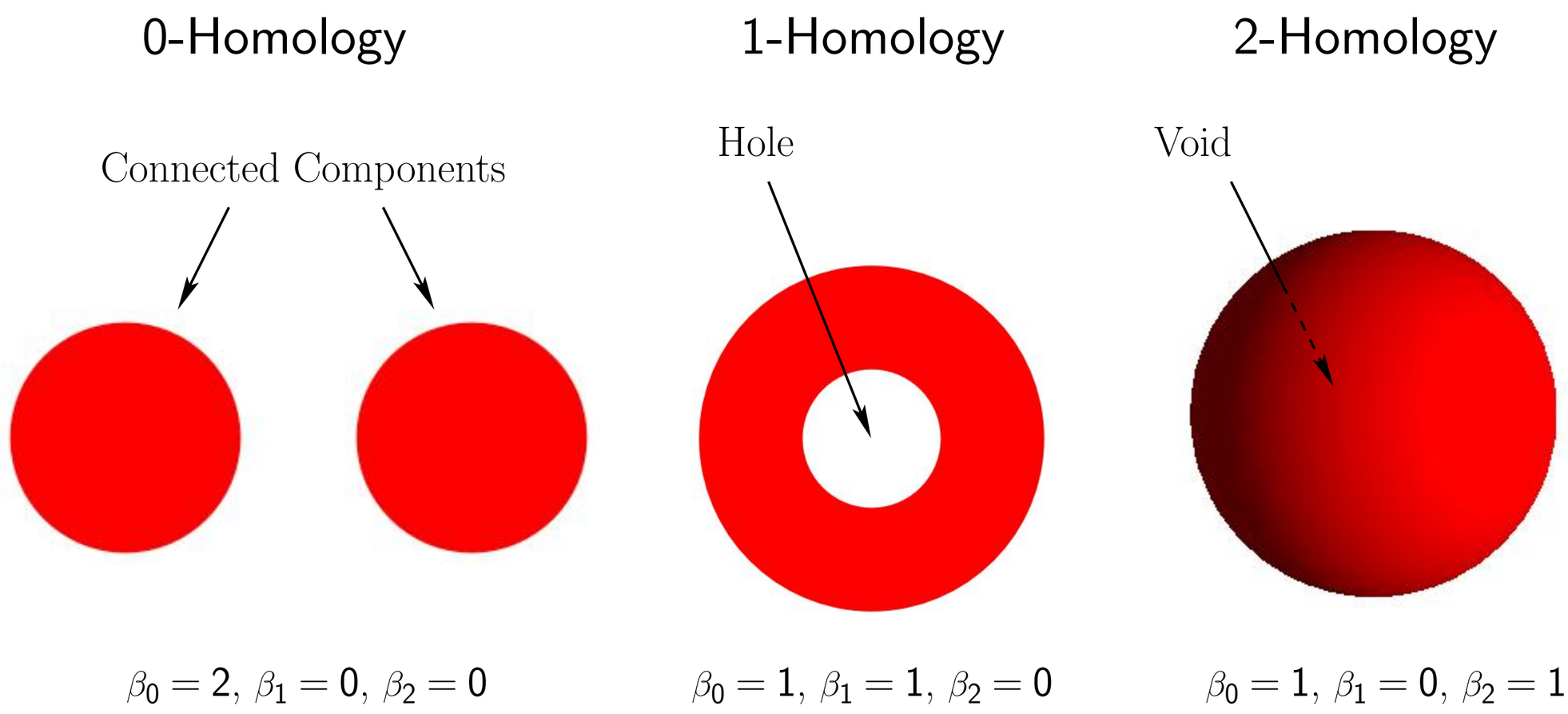
Topological methods

Broken claw tips.

Use integral geometry

(1) Hadwiger integrals
(2) Minkowski functionals
(3) Euler integration
(4) Radon transform.

Betti numbers



Euler characteristic

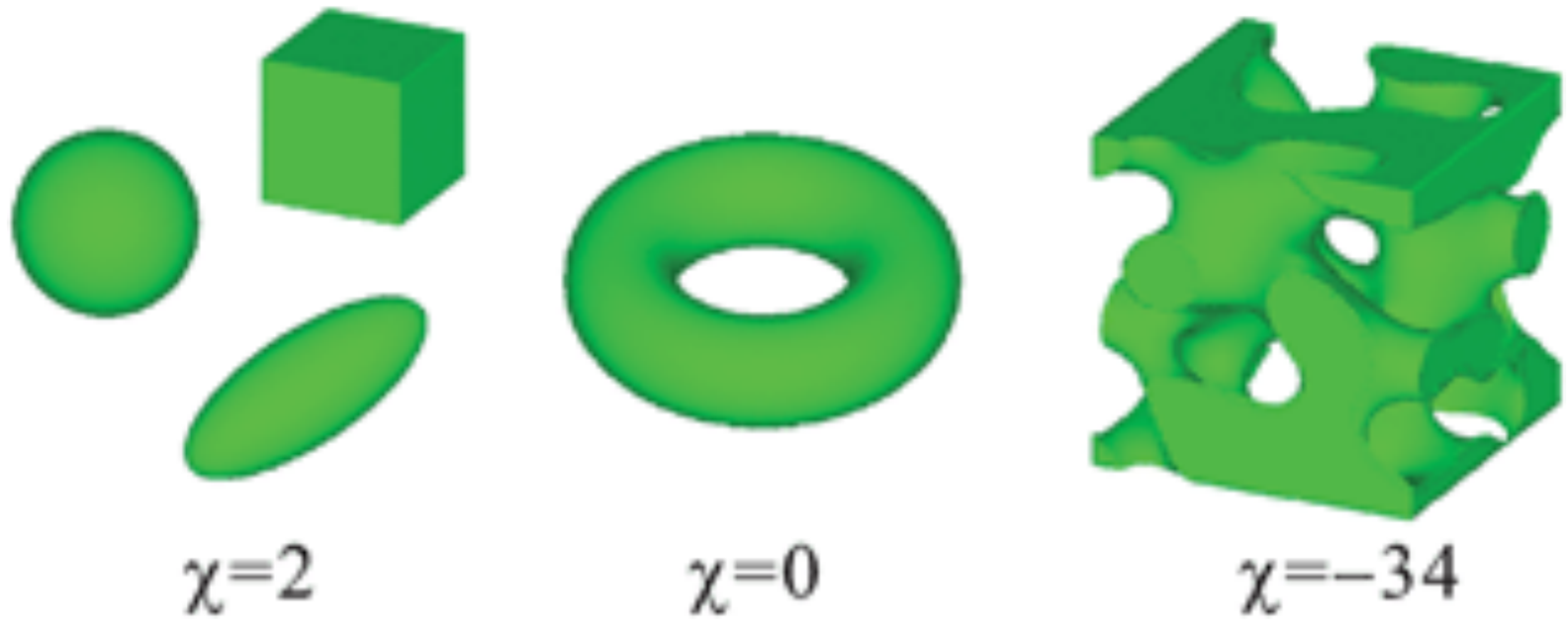
Given a shape M the Euler characteristic is

$$\chi(M) = \sum_{i=0}^{d} (-1)^i \beta_i = \#\text{vertices} - \#\text{edges} + \#\text{faces}.$$

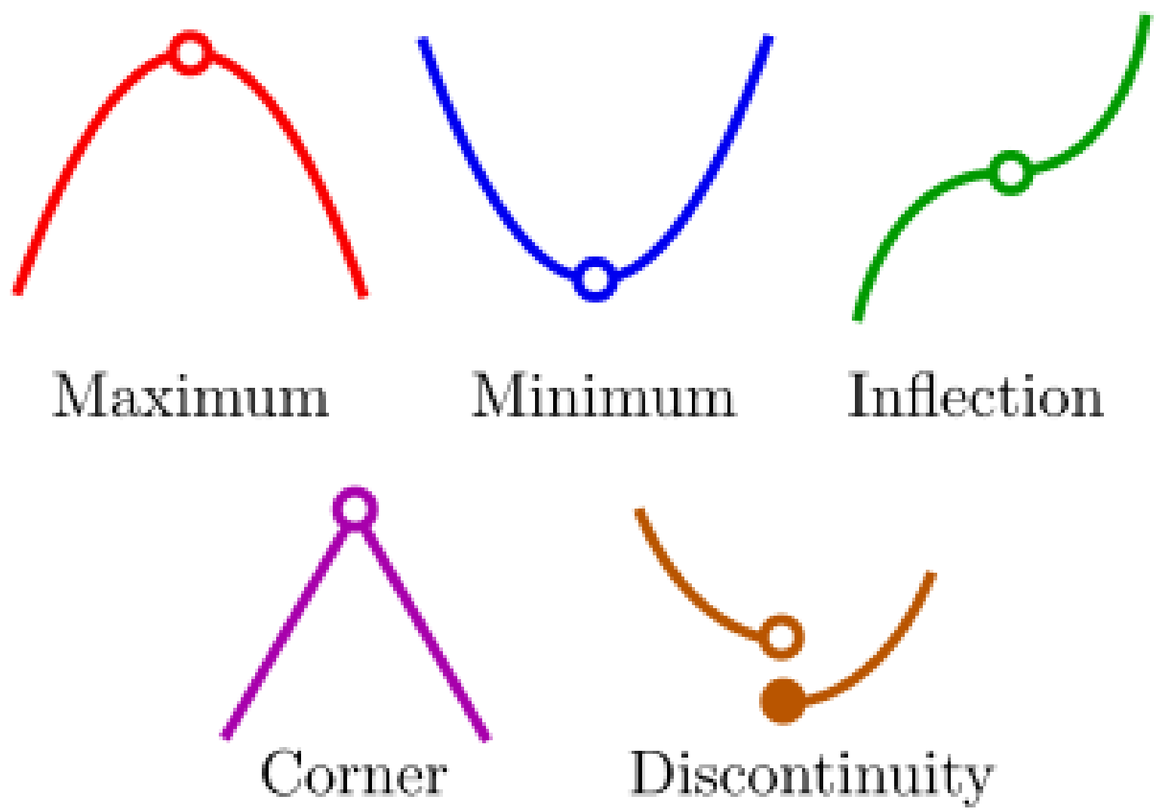
Euler characteristic

Given a shape M the Euler characteristic is

$$\chi(M) = \sum_{i=0}^{d} (-1)^i \beta_i = \#\text{vertices} - \#\text{edges} + \#\text{faces}.$$

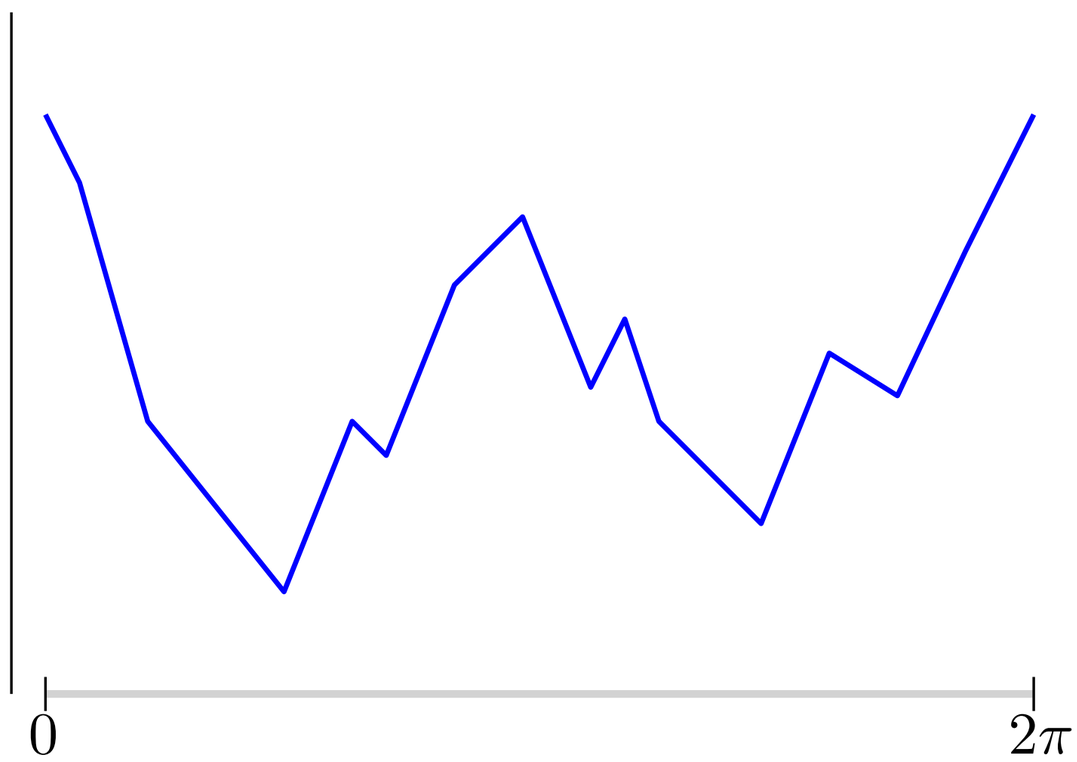


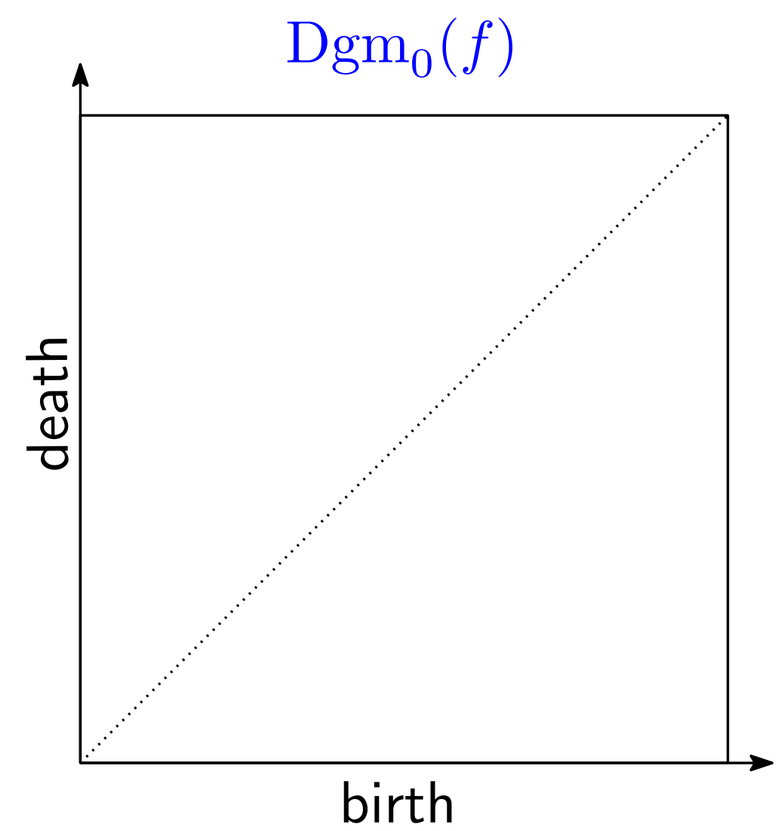
Critical points



Persistent homology: Morse theory

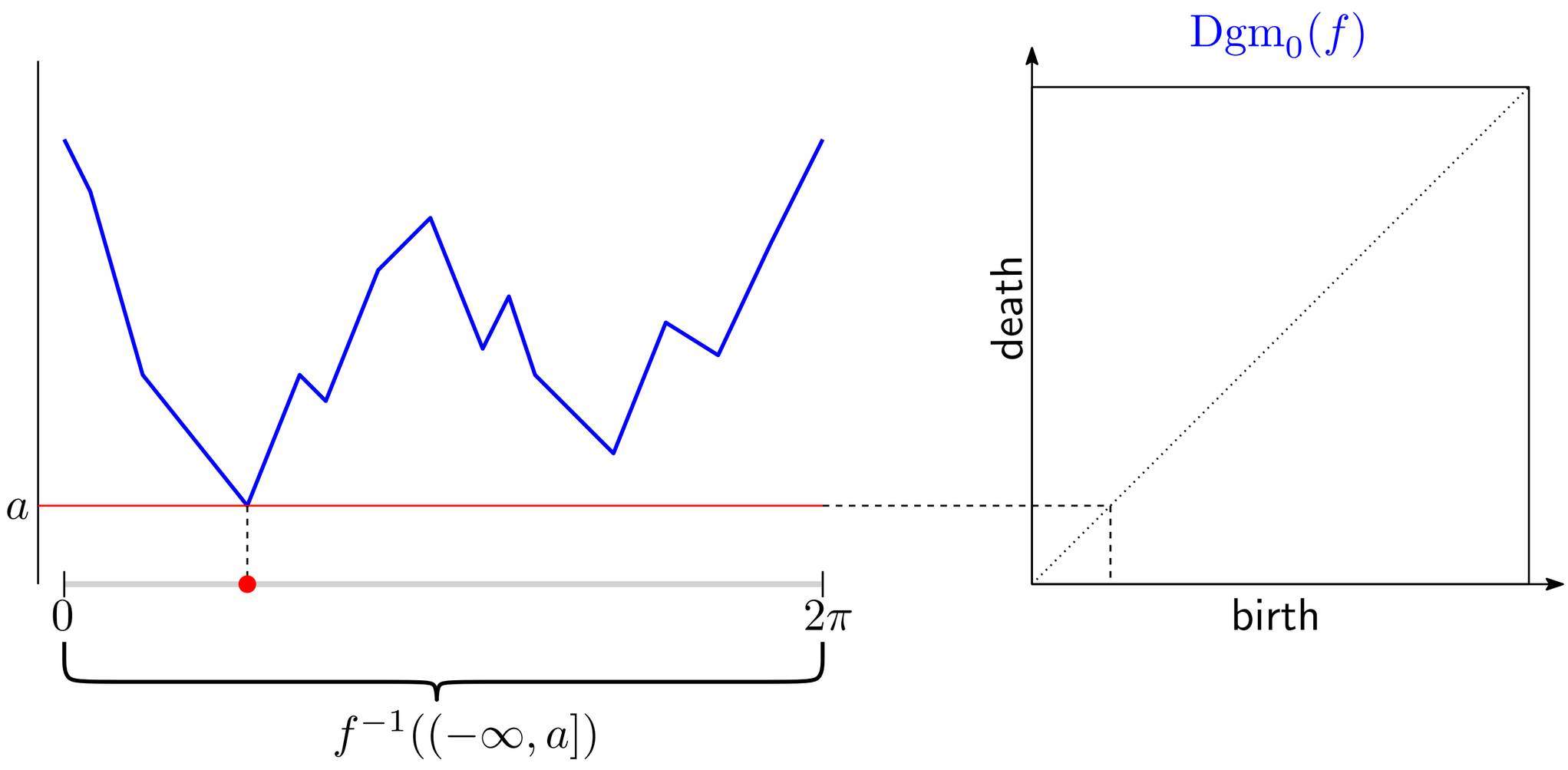
Evolution of homology as birth-death pair.





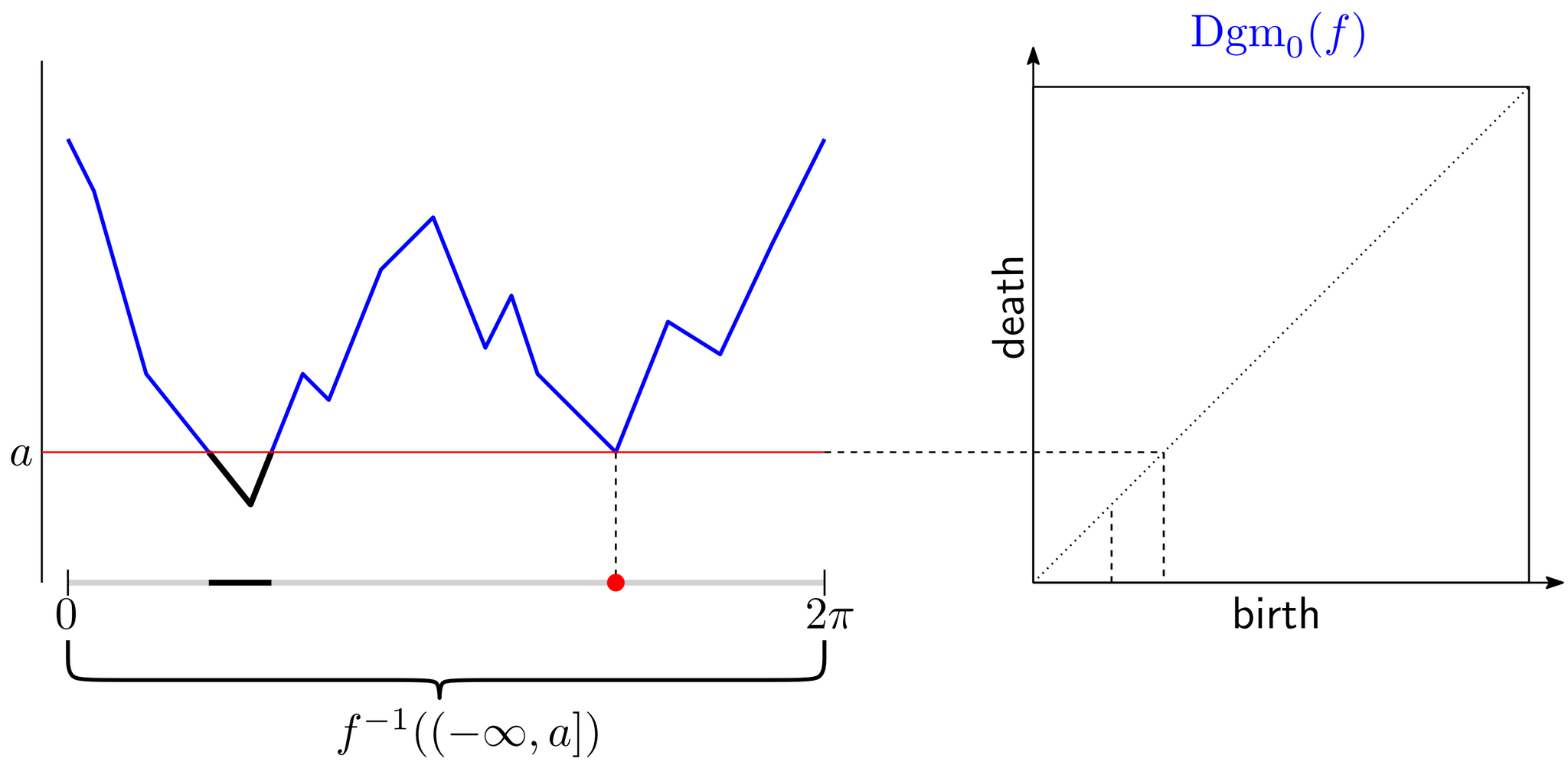
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



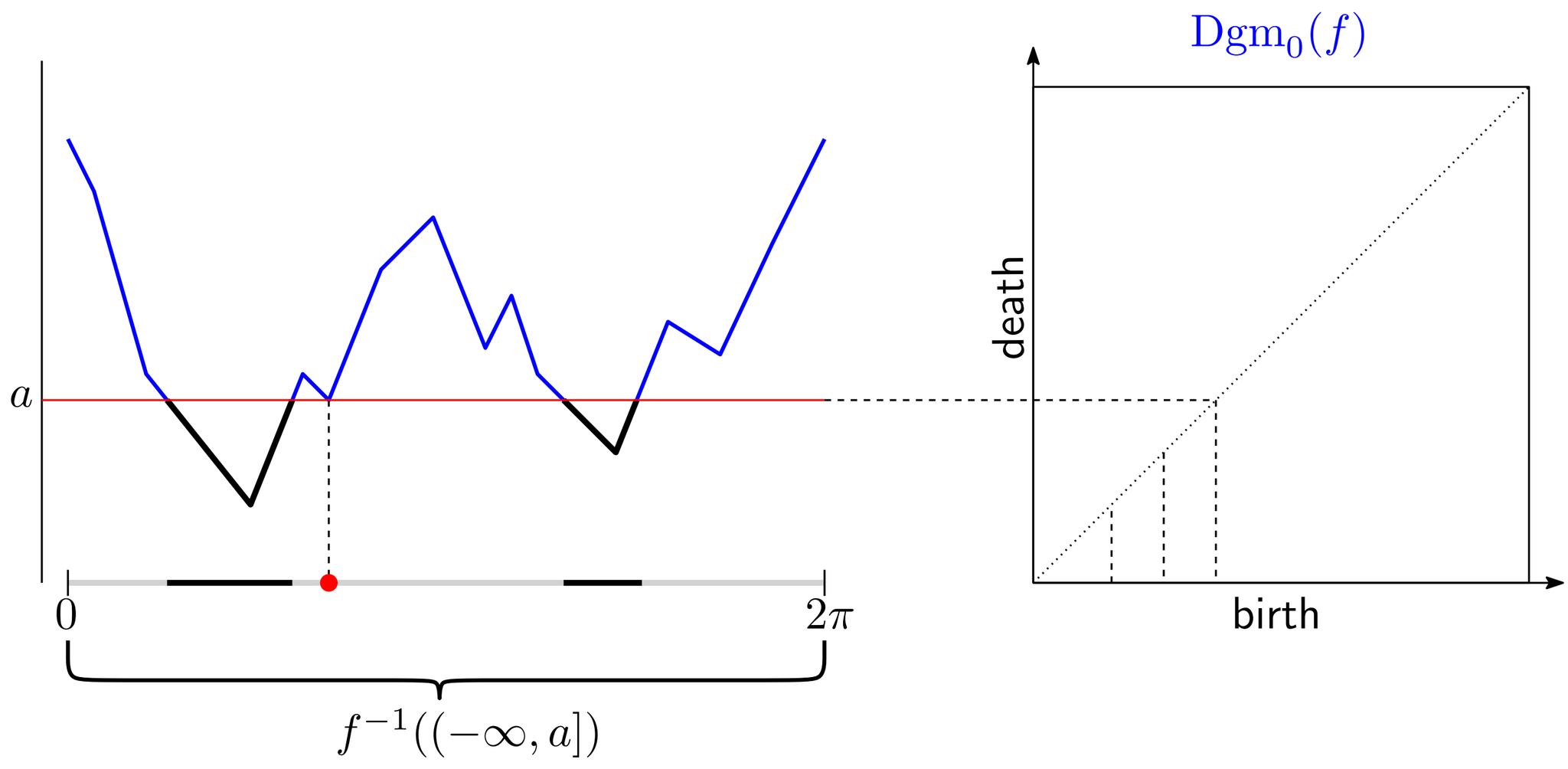
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



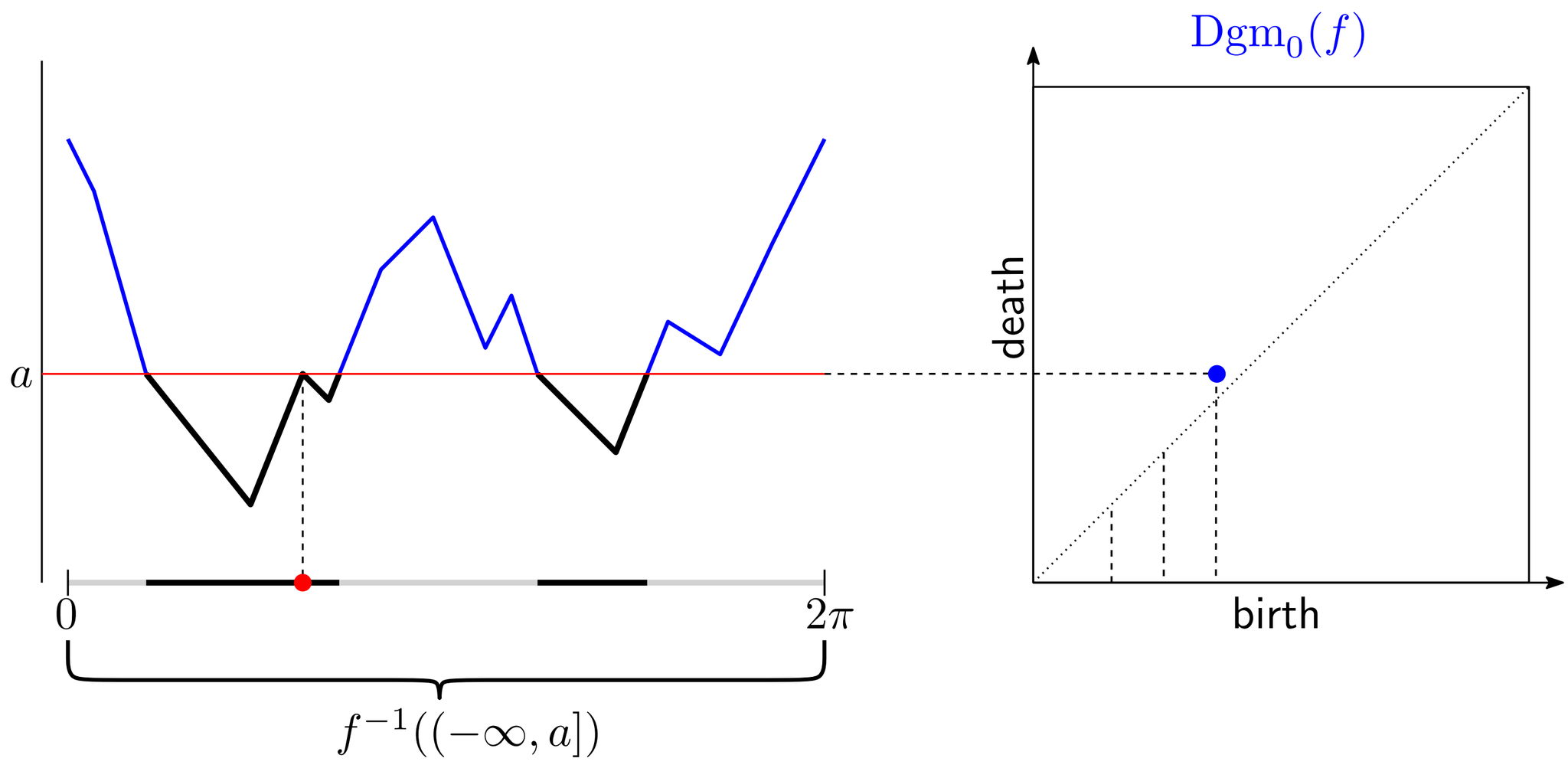
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



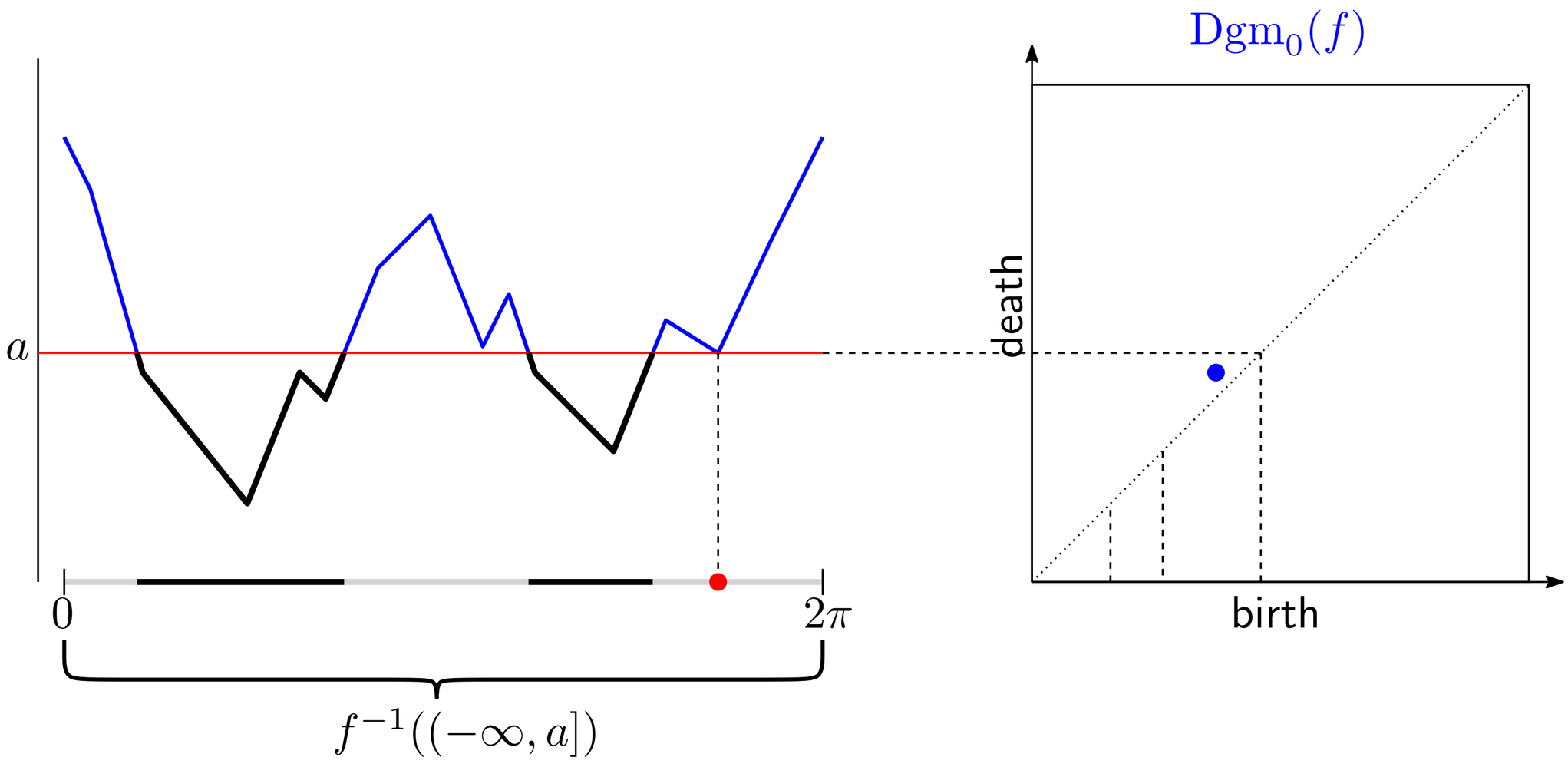
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



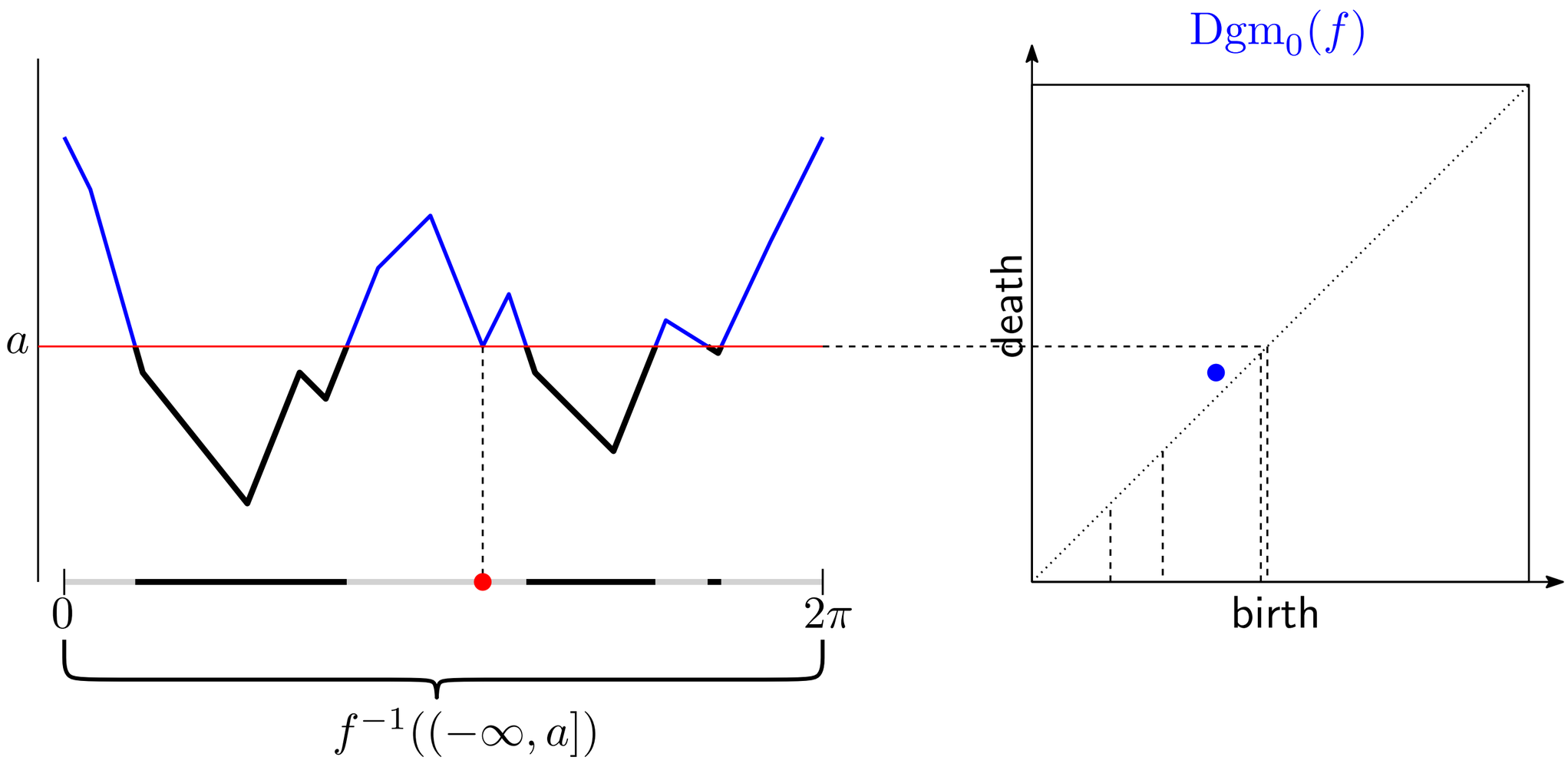
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



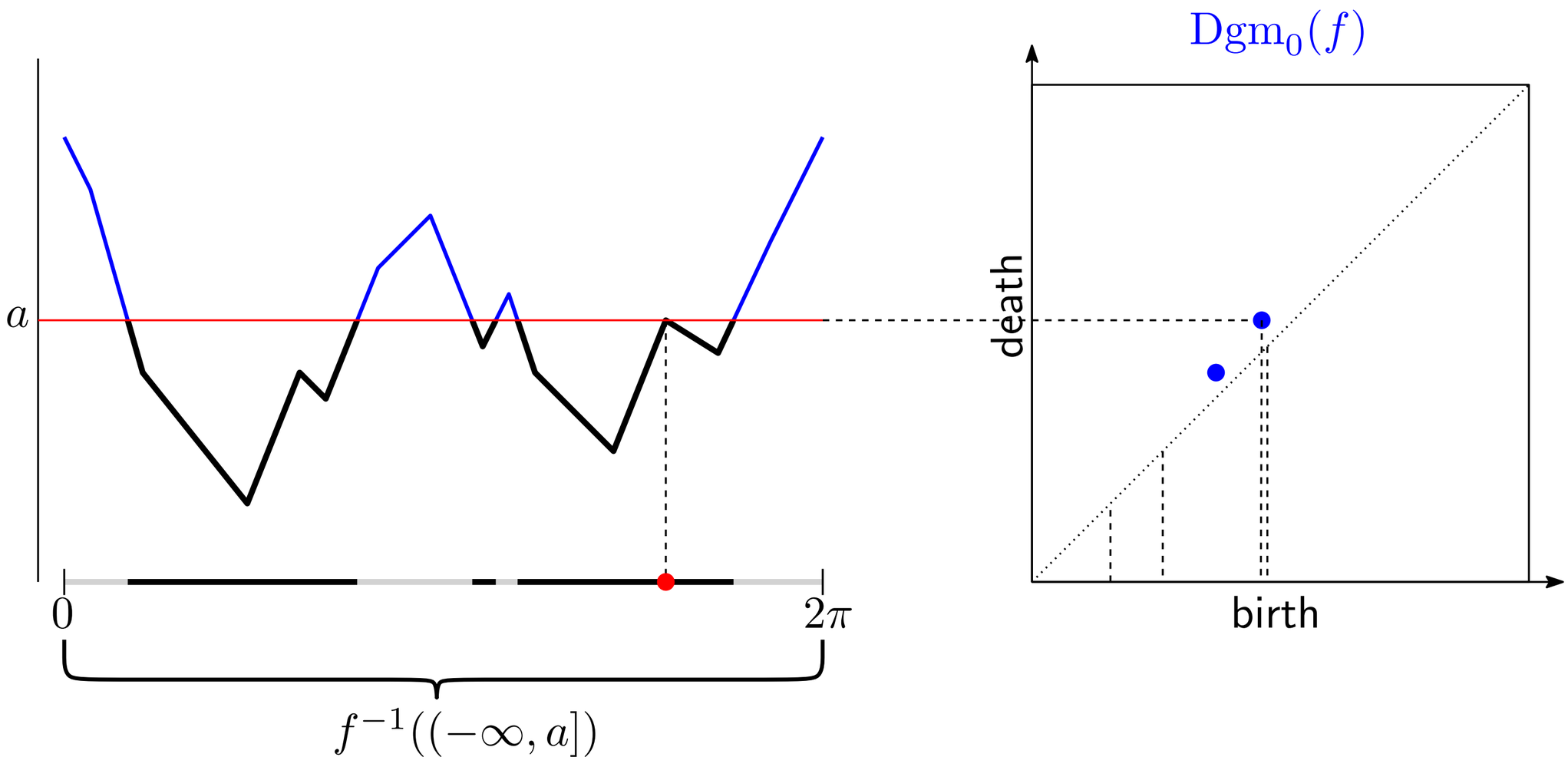
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



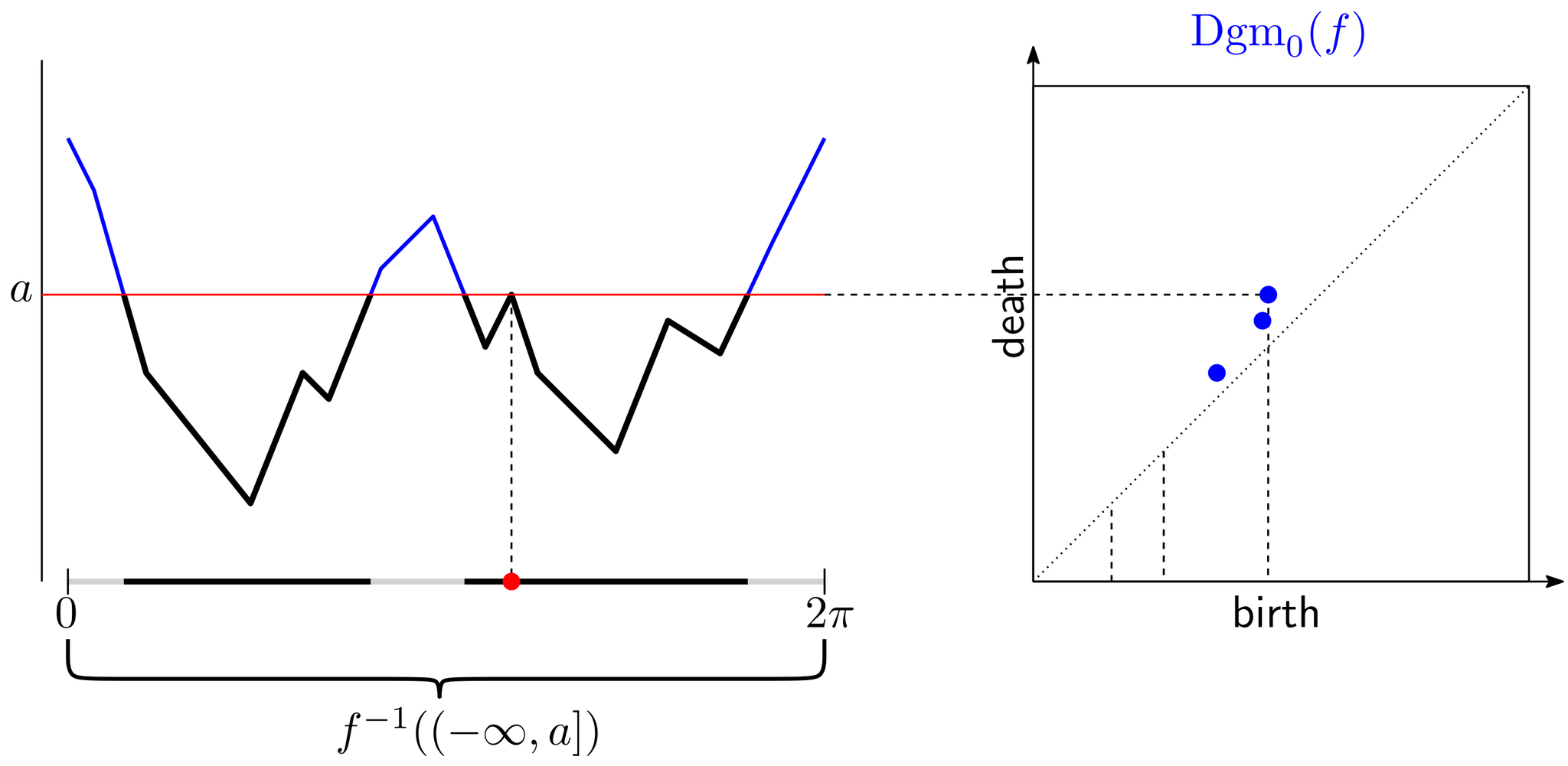
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



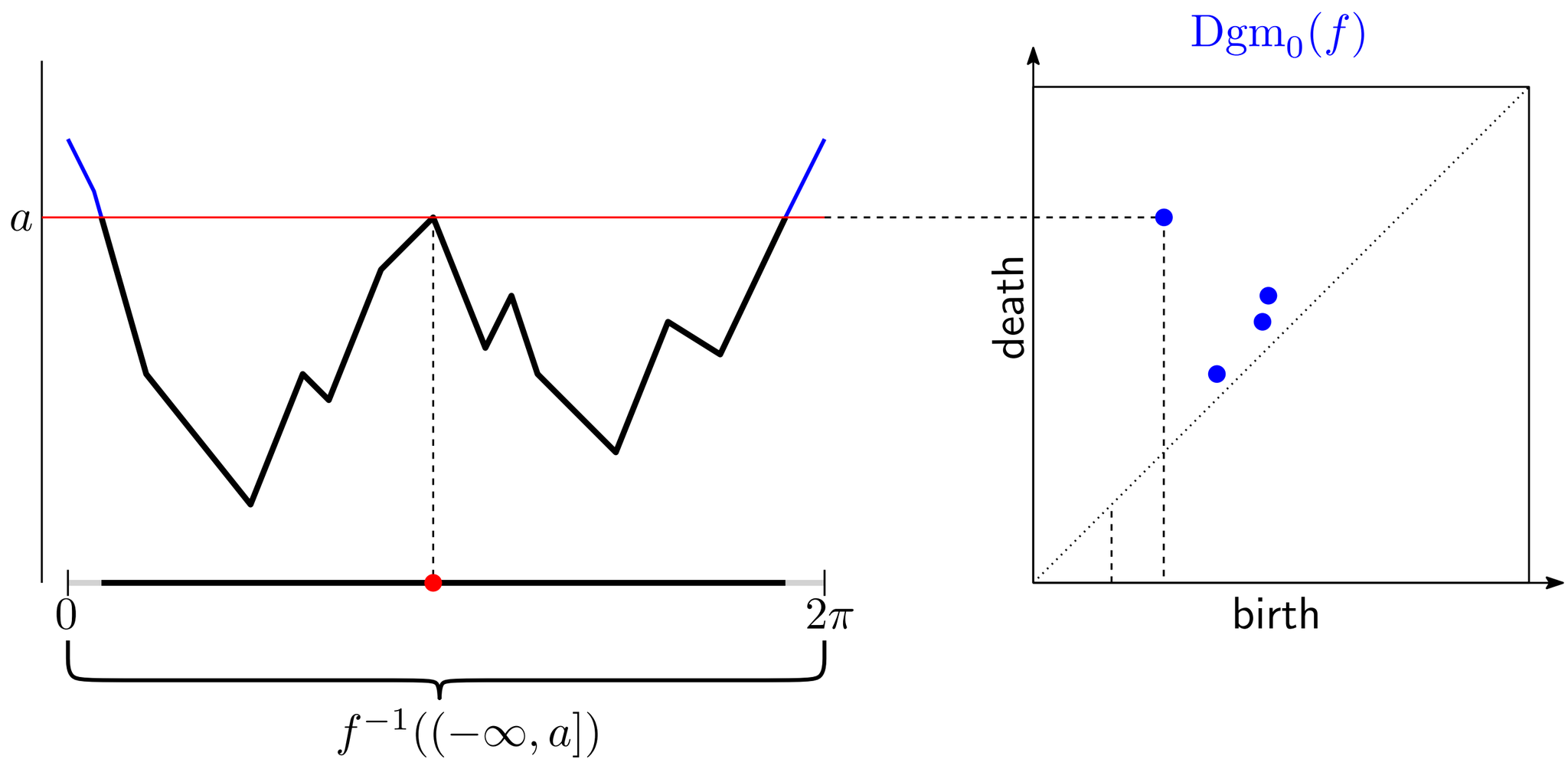
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



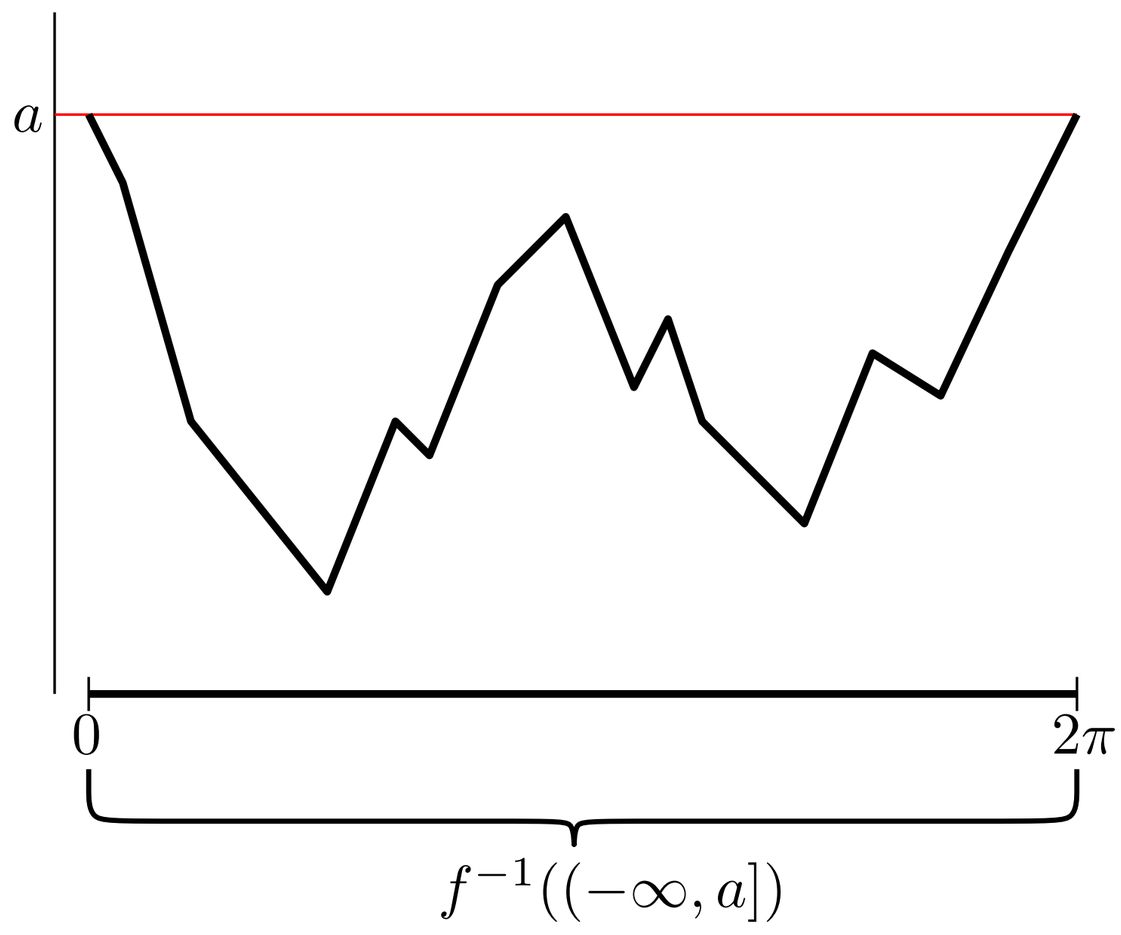
Persistent homology: Morse theory

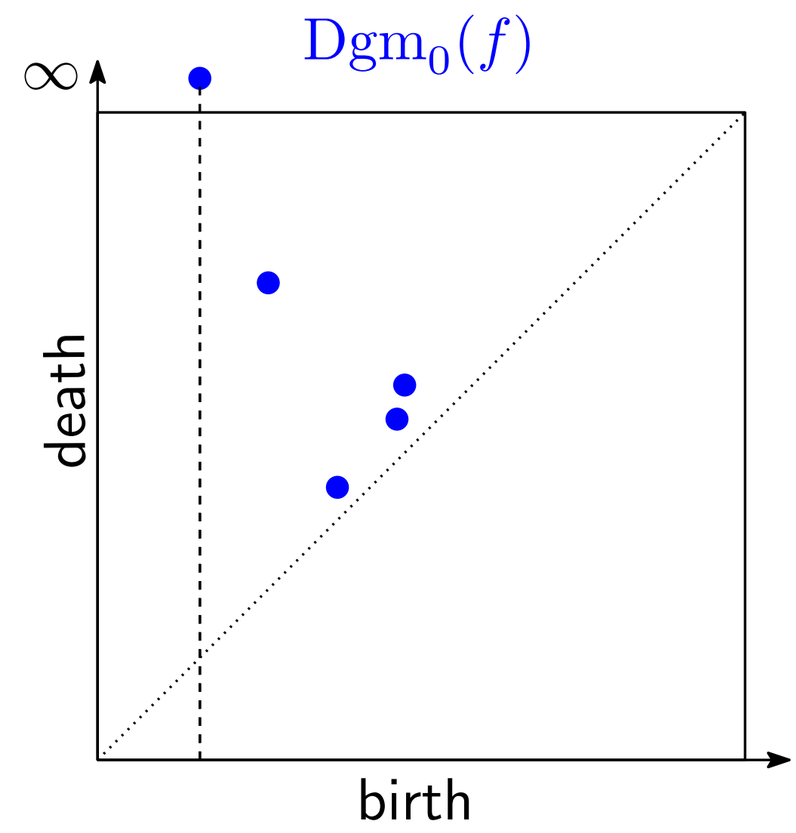
Evolution of homology as birth-death pair.



Persistent homology: Morse theory

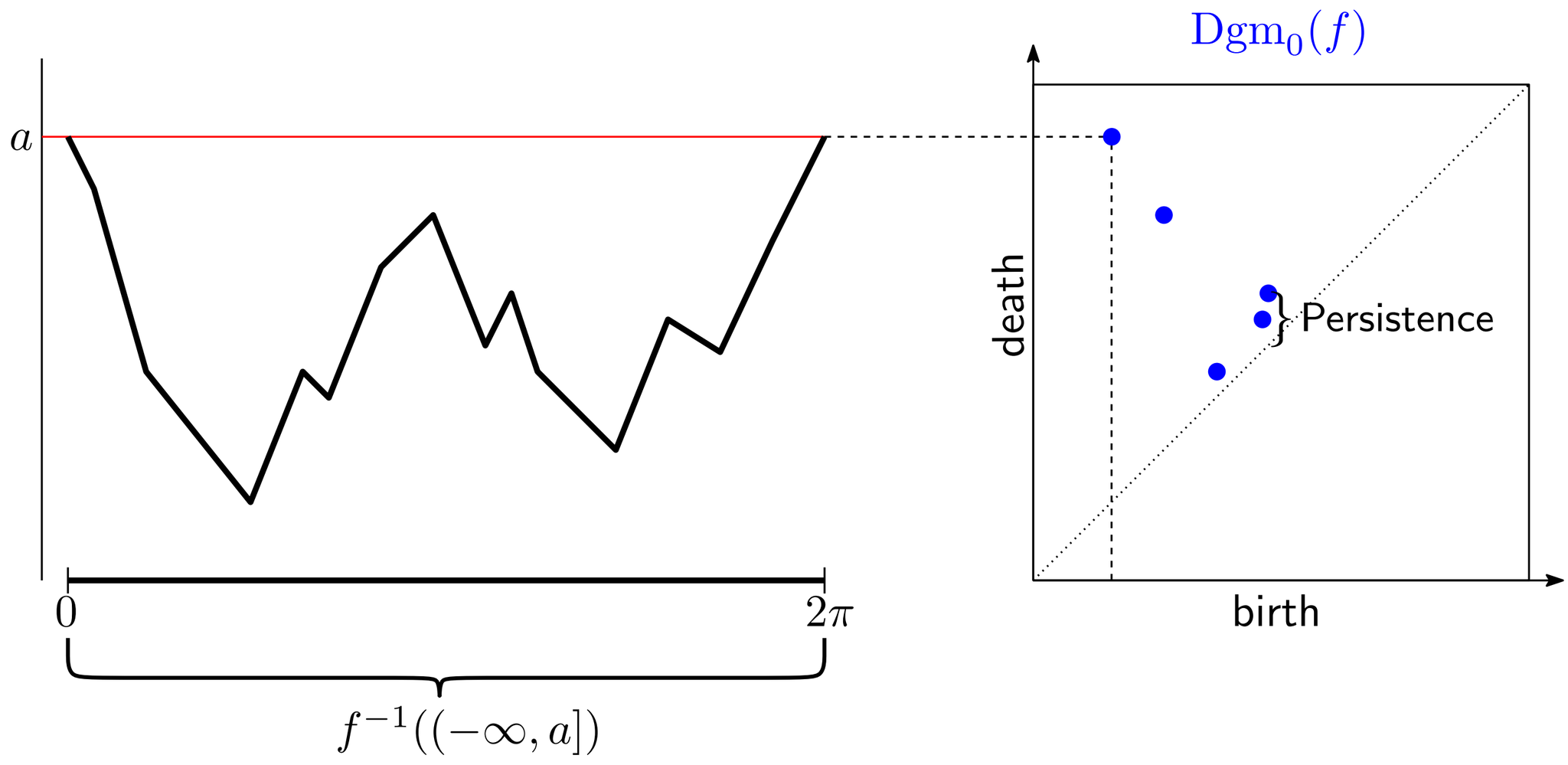
Evolution of homology as birth-death pair.



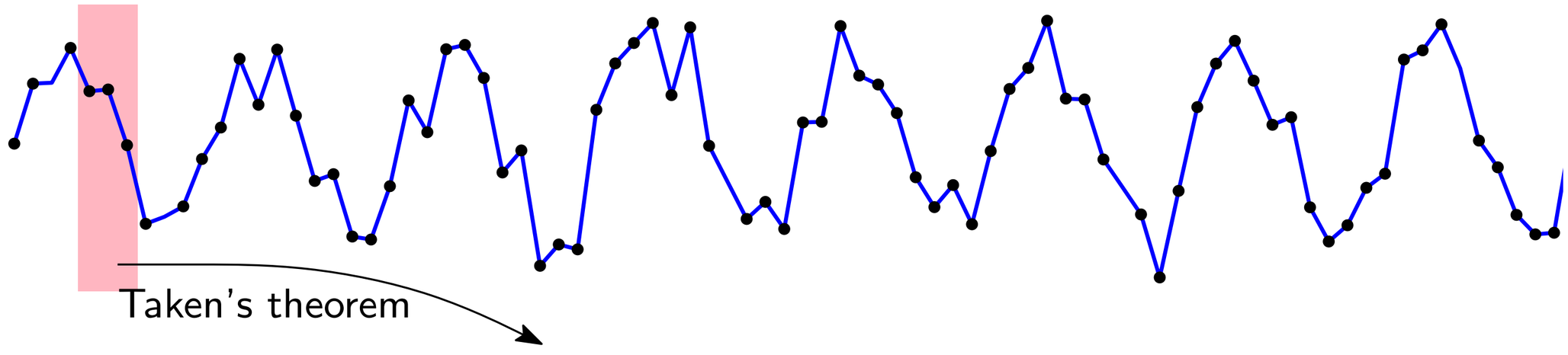


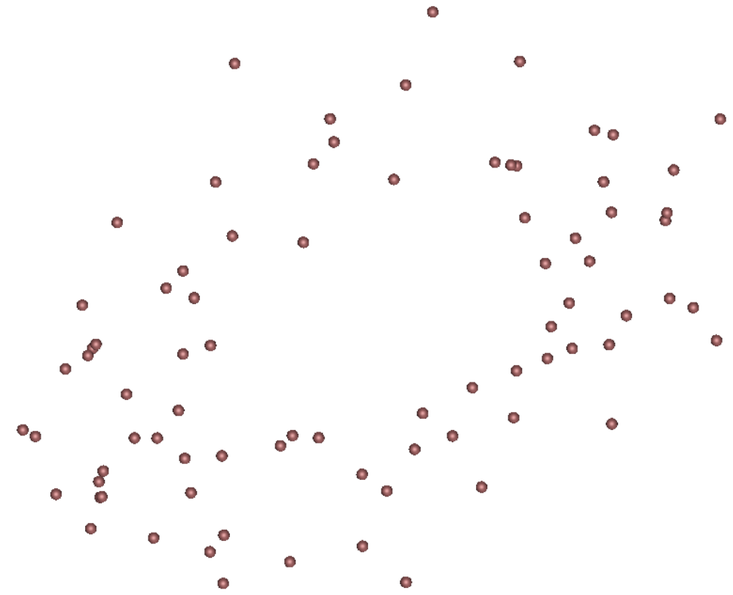
Persistent homology: Morse theory

Evolution of homology as birth-death pair.

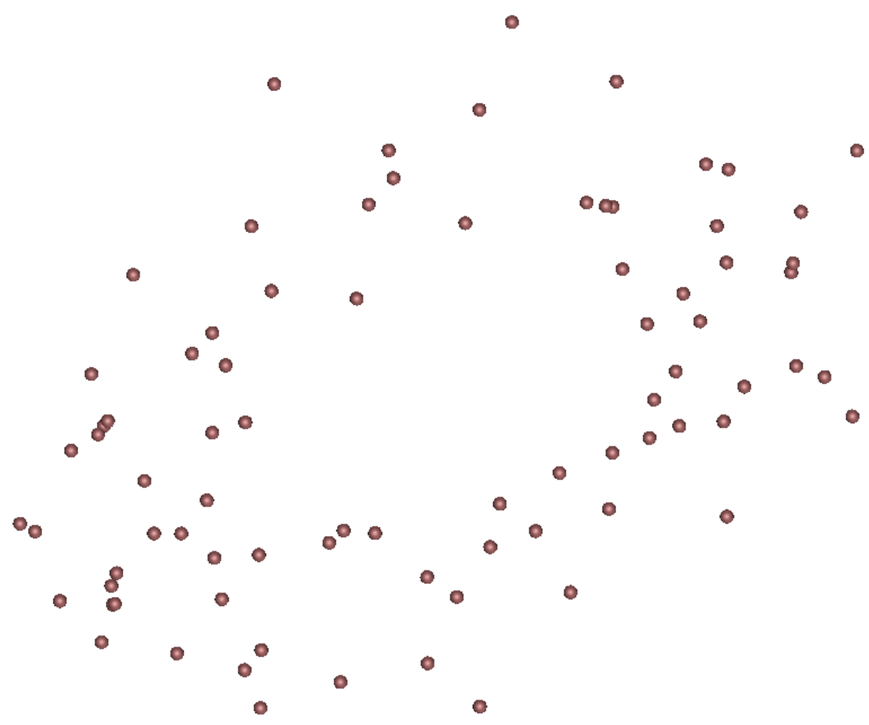


Point cloud data

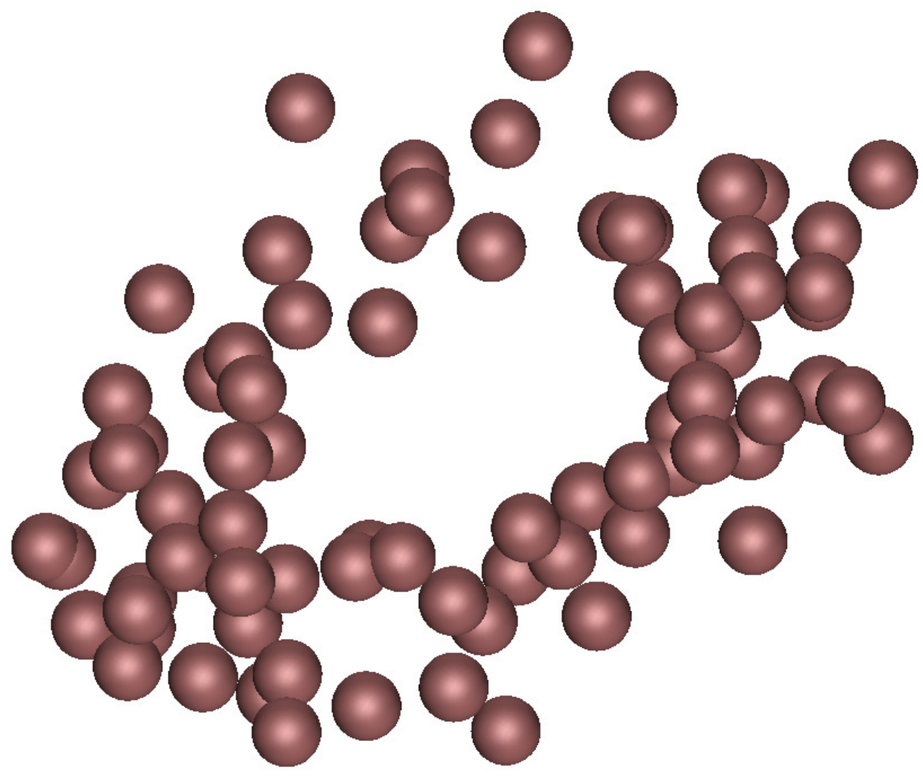




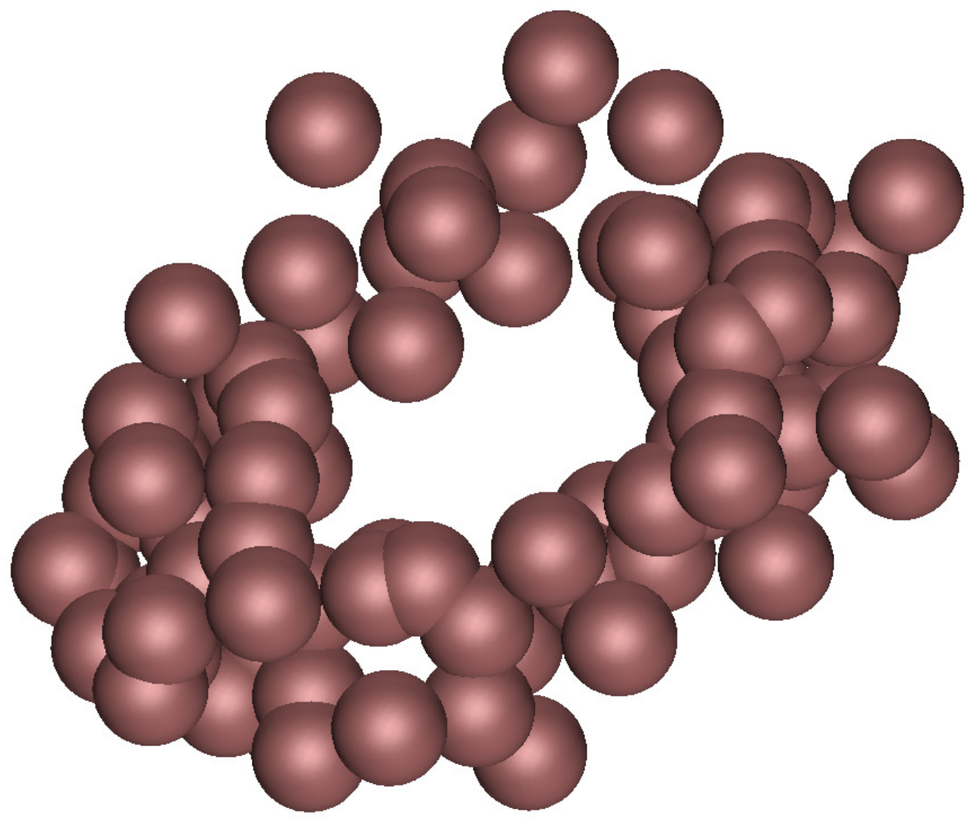
Filtration, $\mathbb{X}_0$



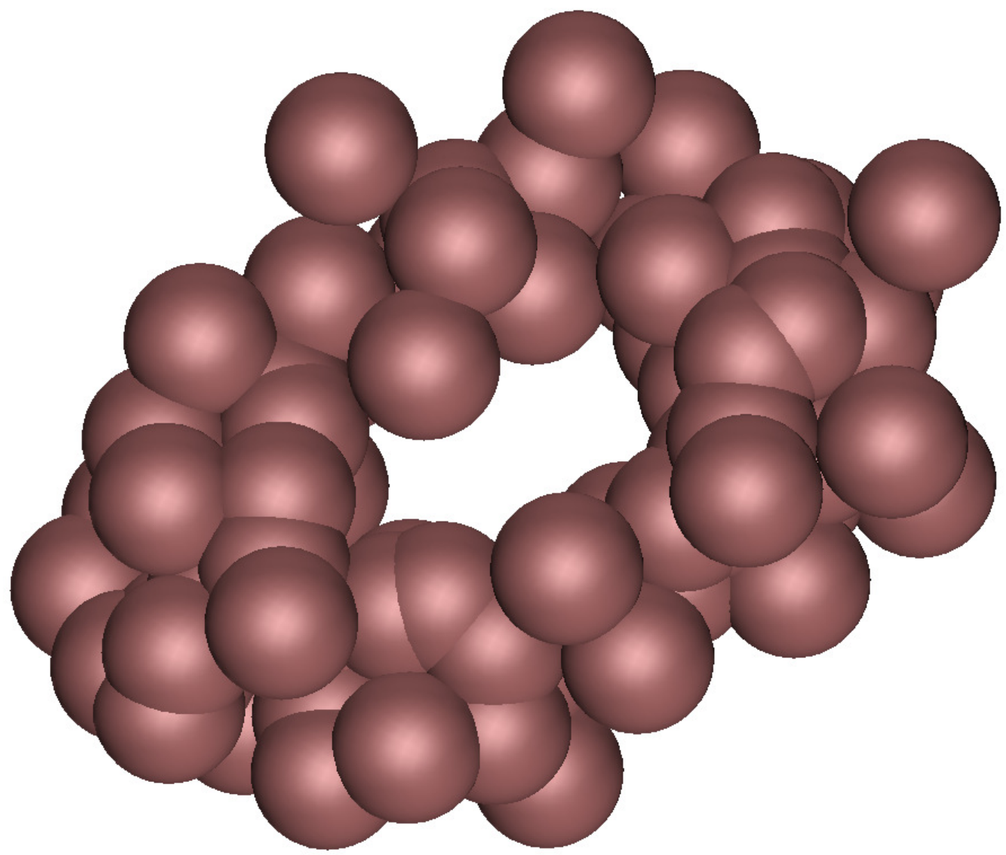
Filtration, $\mathbb{X}_1$



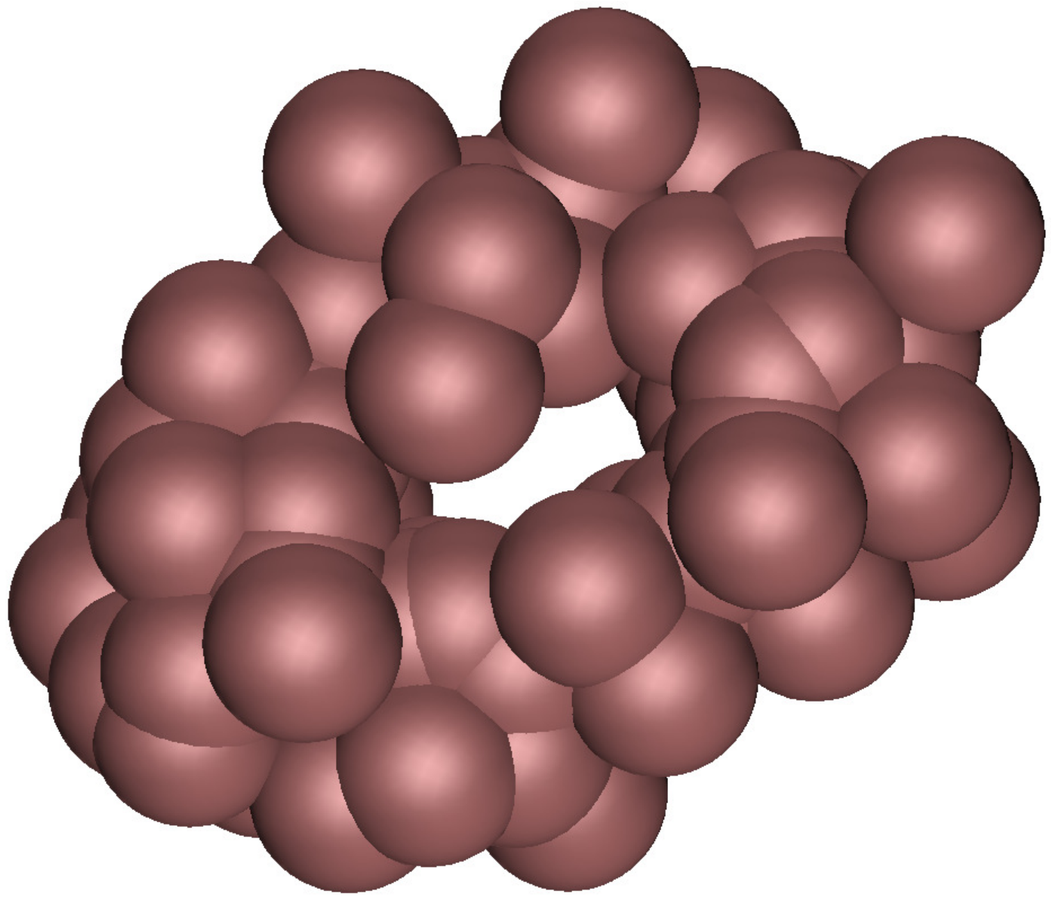
Filtration, $\mathbb{X}_2$



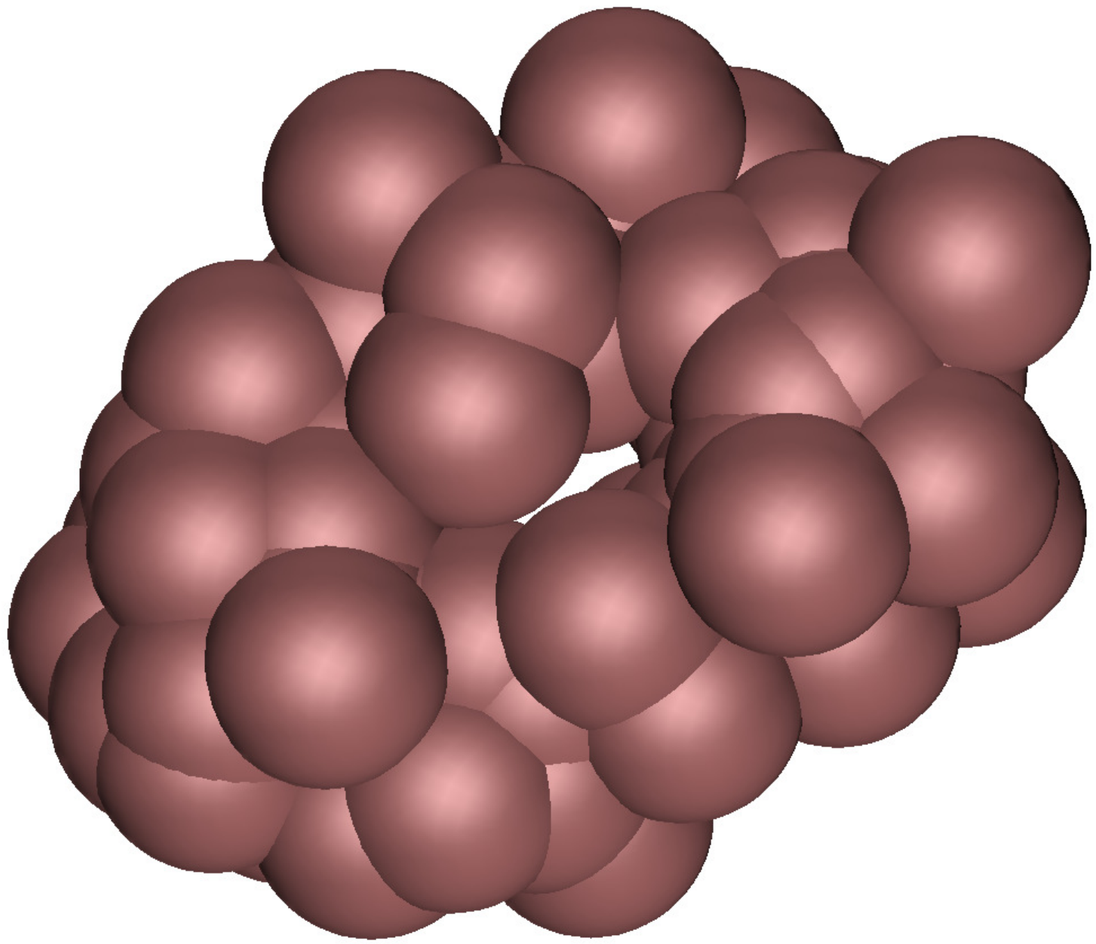
Filtration, $\mathbb{X}_3$



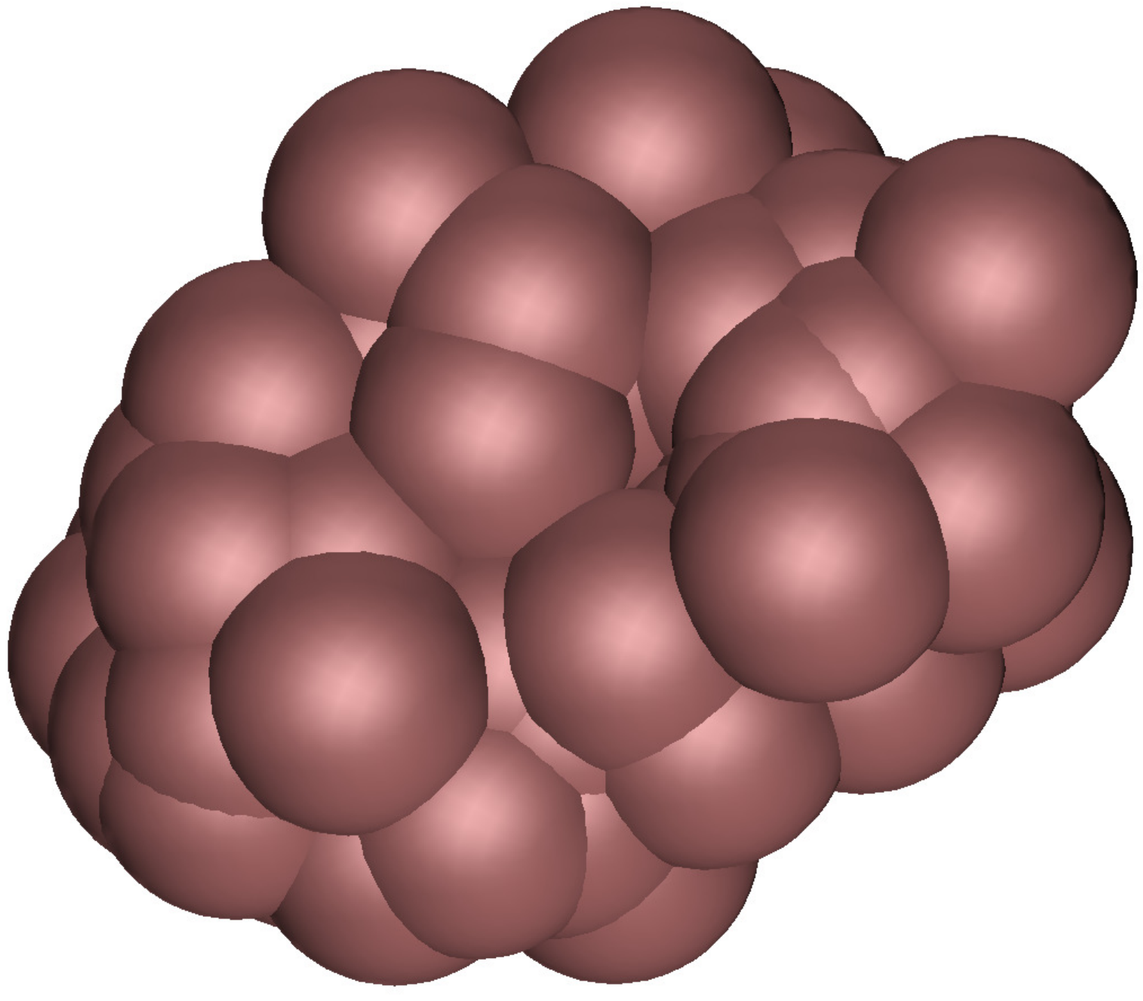
Filtration, $\mathbb{X}_4$



Filtration, $\mathbb{X}_5$

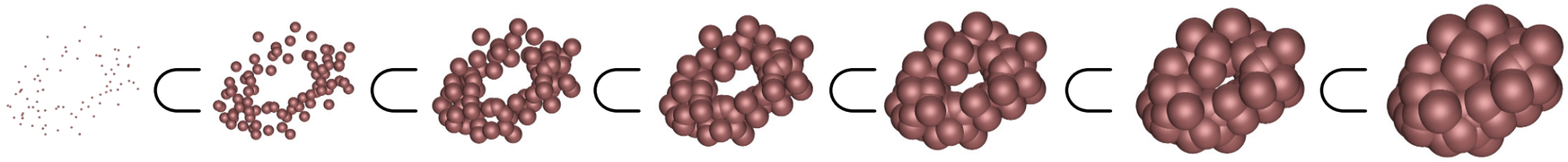


Filtration, $\mathbb{X}_6$



Persistent homology

Construct a filtration

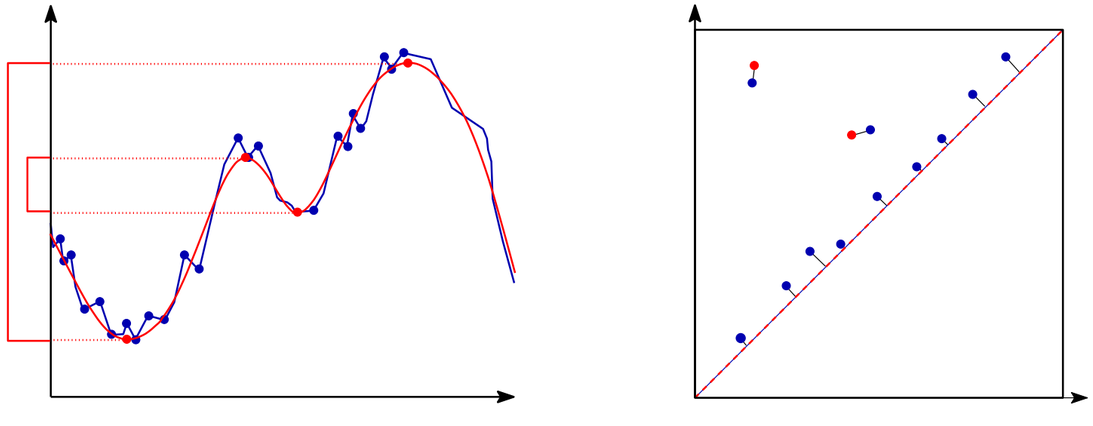


$$\mathbb{X}_0 \subset \mathbb{X}_1 \subset \mathbb{X}_2 \subset \mathbb{X}_3 \subset \mathbb{X}_4 \subset \mathbb{X}_5 \subset \mathbb{X}_6$$

$$H_p(\mathbb{X}_0) \to H_p(\mathbb{X}_1) \to H_p(\mathbb{X}_2) \to H_p(\mathbb{X}_3) \to H_p(\mathbb{X}_4) \to H_p(\mathbb{X}_5) \to H_p(\mathbb{X}_6)$$

Images of linear maps $\phi_p^{i,j} : H_p(\mathbb{X}_i) \to H_p(\mathbb{X}_j)$ induced by inclusion.
Determine when a homology class is born and when it dies.

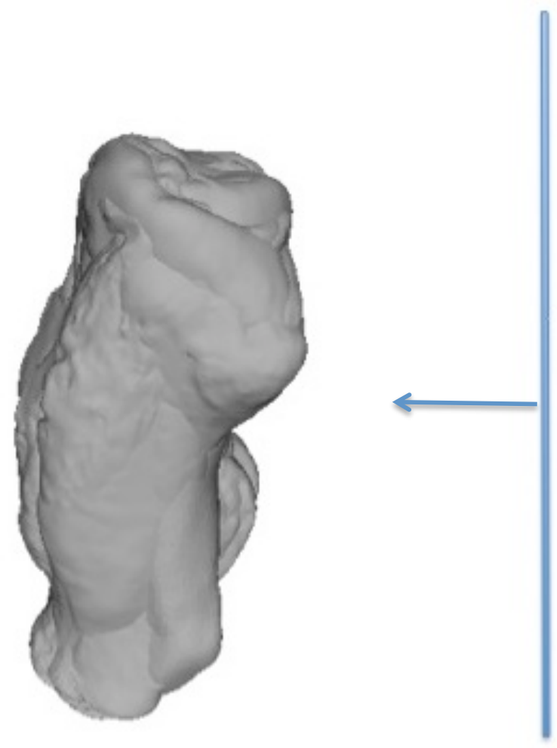
Metrics on diagrams



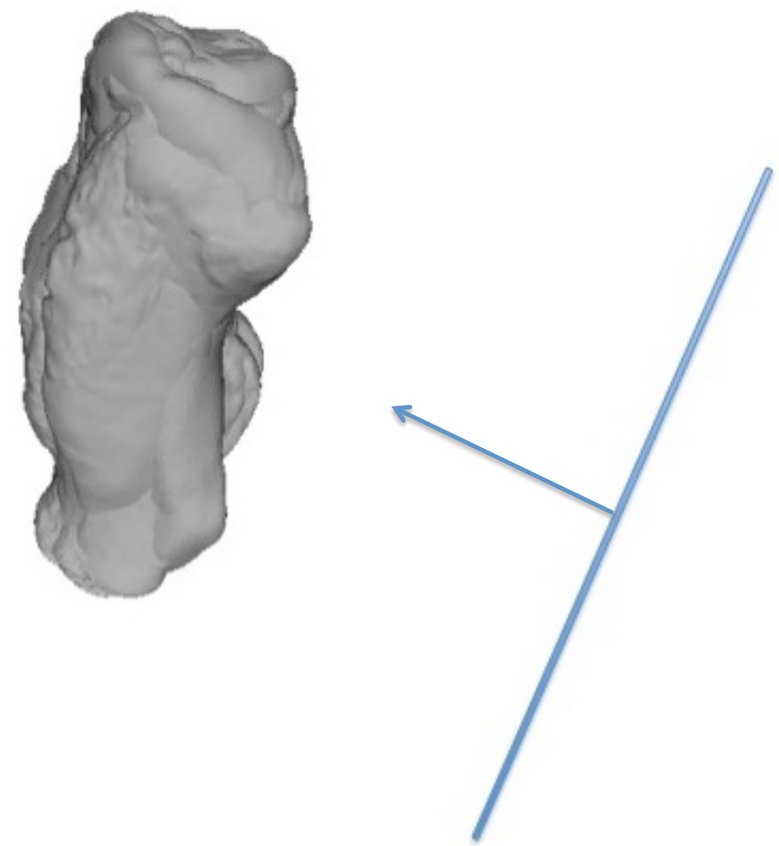
L^2-Wasserstein distance

$$d_{L^2}(X, Y)^2 = \inf_{\phi: X \to Y} \sum_{x \in X} \|x - \phi(x)\|^2$$

Height function: v_1



Height function: v_2



Persistence homology transform (PHT)

M is simplicial complex in $\mathbb{R}^d$ and $v \in S^{d-1}$ is a unit vector.
$X_k(M, v)$ captures changes in topology of

$$M(v)_r = \{\Delta \in M : x \cdot v \leq r \text{ for all } x \in \Delta\}.$$

Persistence homology transform (PHT)

M is simplicial complex in $\mathbb{R}^d$ and $v \in S^{d-1}$ is a unit vector.
$X_k(M, v)$ captures changes in topology of

$$M(v)_r = \{\Delta \in M : x \cdot v \leq r \text{ for all } x \in \Delta\}.$$

Definition

The persistent homology transform of $M \in \mathbb{R}^d$ is the function

$$\begin{aligned} \mathrm{PHT}(M) : S^{d-1} &\to \mathcal{D}^{d-1} \\ v &\mapsto (X_0(M, v), X_1, (M, v) \ldots, X_{d-1}(M, v)). \end{aligned}$$

Distances

$\mathcal{M}_d$ is the space of finite simplicial complexes in $\mathbb{R}^d$.

Distances

$\mathcal{M}_d$ is the space of finite simplicial complexes in $\mathbb{R}^d$.

The distance between two surfaces M_1, M_2 is

$$d_{\mathcal{M}_d}(M_1, M_2) := \sum_{k=0}^{d} \int_{S^{d-1}} d(X_k(M_1, v), X_k(M_2, v))dv.$$

Euler characteristic transform (ECT)

M is simplicial complex in $\mathbb{R}^d$ and $v \in S^{d-1}$ is a unit vector.
$\chi(M, v)$ captures changes in topology of

$$M(v)_r = \{\Delta \in M : x \cdot v \leq r \text{ for all } x \in \Delta\}.$$

Euler characteristic transform (ECT)

M is simplicial complex in $\mathbb{R}^d$ and $v \in S^{d-1}$ is a unit vector.
$\chi(M, v)$ captures changes in topology of

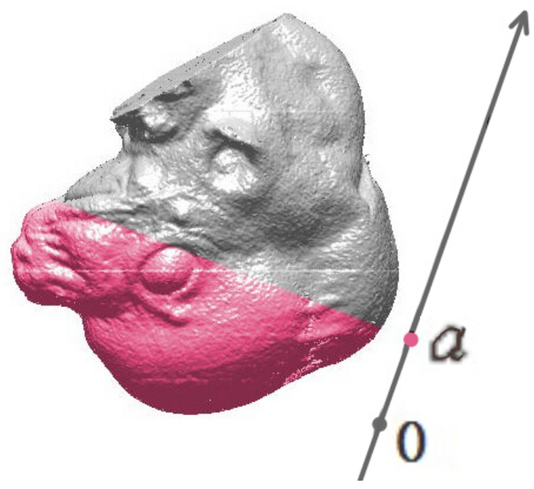
$$M(v)_r = \{\Delta \in M : x \cdot v \leq r \text{ for all } x \in \Delta\}.$$

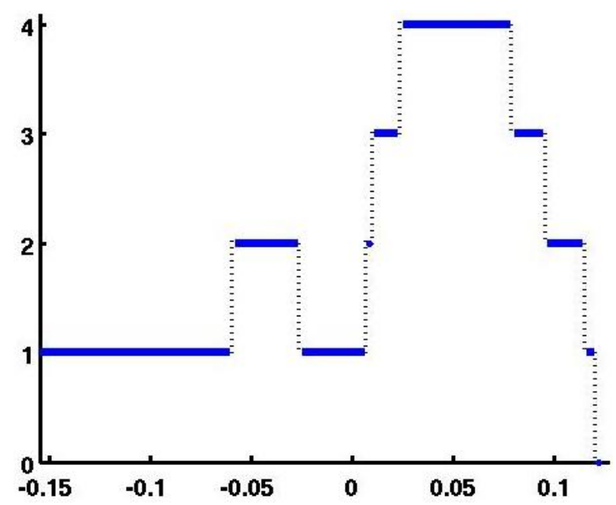
Definition

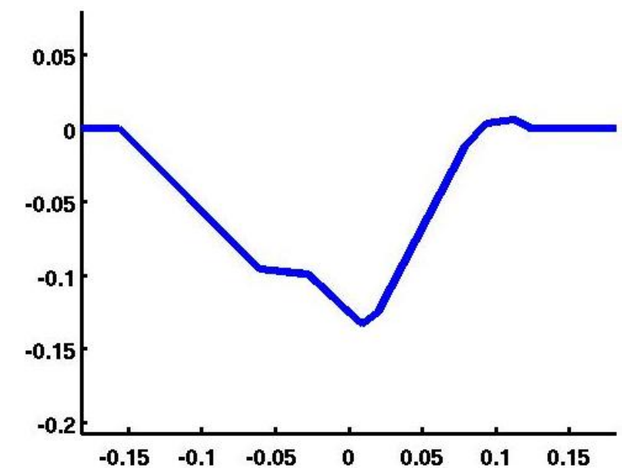
The Euler characteristic transform of $M \in \mathbb{R}^d$ is the function

$$\begin{aligned} \mathrm{ECT}(M) : S^{d-1} &\to L_2(\mathbb{R}) \\ v &\mapsto \chi(M, v). \end{aligned}$$

Euler characteristic curve







Mao Li

Sufficient statistic

Given $X \sim f_\theta \in \mathcal{F}$, a statistic $T = T(X)$ is sufficient if for the parameter θ if for all sets B the probability $\mathbb{P}[X \in B \mid T(X) = t]$ does not depend on θ

$$\mathbb{P}[X \mid T(X) = t, \theta] = \mathbb{P}[X \mid T(X) = t].$$

Sufficient statistic

Given $X \sim f_\theta \in \mathcal{F}$, a statistic $T = T(X)$ is sufficient if for the parameter θ if for all sets B the probability $\mathbb{P}[X \in B \mid T(X) = t]$ does not depend on θ

$$\mathbb{P}[X \mid T(X) = t, \theta] = \mathbb{P}[X \mid T(X) = t].$$

For the normal distribution with known variance $\hat{\mu} = \frac{1}{n} \sum_i x_i$ is a sufficient statistic.

Sufficiency of the PHT

Theorem (Turner-M-Boyer)

The persistent homology transform is injective when the domain is $\mathcal{M}_d$ for $d = 2, 3$.

Sufficiency of the PHT

Theorem (Turner-M-Boyer)

The persistent homology transform is injective when the domain is $\mathcal{M}_d$ *for* $d = 2, 3$.

Corollary (Turner-M-Boyer)

Consider the subspace of shapes $\mathcal{M}_k^N$ *(for* $k = 2$ *or* 3*), piecewise linear simplicial complexes with at most* N *vertices. Let* $f(x; \theta)$ *be a density function over* $\mathcal{M}_k$ *with parameters* $\theta \in \Theta$ *and* $x \in \mathcal{M}_k$ *whose support is contained in some* $\mathcal{M}_k^N$*. The persistence homology transform* $t(X) \in C(S^2, \mathcal{D}^3)$ *is a sufficient statistic.*

Sufficiency of the ECT

Theorem (Turner-M-Boyer)

The Euler characteristic transform is injective when the domain is $\mathcal{M}_d$ for $d = 2, 3$.

Sufficiency of the ECT

Theorem (Turner-M-Boyer)

The Euler characteristic transform is injective when the domain is $\mathcal{M}_d$ *for* $d = 2, 3$.

Corollary (Turner-M-Boyer)

Consider the subspace of shapes $\mathcal{M}_k^N$ *(for* $k = 2$ *or* 3*), piecewise linear simplicial complexes with at most* N *vertices. Let* $f(x; \theta)$ *be a density function over* $\mathcal{M}_k$ *with parameters* $\theta \in \Theta$ *and* $x \in \mathcal{M}_k$ *whose support is contained in some* $\mathcal{M}_k^N$*. The Euler characteristic transform* $t(X) \in C(S^2, \mathcal{D}^3)$ *is a sufficient statistic.*

Exponential family

Given sufficient statistic $T(z) = (T_1(x), ..., T_d(x))^T$ the exponential family takes the form

$$p_\theta(x) = a(\theta)\, h(x)\, \exp(-\langle \theta, T(x) \rangle),$$

with $\langle \cdot, \cdot \rangle$ standard inner product.

Exponential family

Given sufficient statistic $T(z) = (T_1(x), ..., T_d(x))^T$ the exponential family takes the form

$$p_\theta(x) = a(\theta)\, h(x)\, \exp(-\langle \theta, T(x) \rangle),$$

with $\langle \cdot, \cdot \rangle$ standard inner product.

Likelihood model for surfaces Data $\equiv (X_1, ..., X_n) \overset{iid}{\sim} p_\theta$, stated as

$$\text{Lik}(\text{Data} \mid \theta) = \prod_{i=1}^{n} a(\theta)\, h(x_i)\, \exp(-\langle \theta, T(x_i) \rangle).$$

Exponential family and ECT

Denote the Euler characteristic curve for each direction:
$f(y) = \chi(M, v)$ Define the integral of $f(y)$ as $F(x) = \int_0^x f(y)dy$.

Exponential family and ECT

Denote the Euler characteristic curve for each direction:
$f(y) = \chi(M, v)$ Define the integral of $f(y)$ as $F(x) = \int_0^x f(y)dy$.

This results in K smooth curves $\{F_1, ..., F_K\}$.

Exponential family and ECT

Denote the Euler characteristic curve for each direction:
$f(y) = \chi(M, v)$ Define the integral of $f(y)$ as $F(x) = \int_0^x f(y)dy$.

This results in K smooth curves $\{F_1, ..., F_K\}$.

Exponential family model

$$p_\theta(x) = a(\theta)\, h(x)\, \exp\left(-\sum_{k=1}^{K} \langle \theta, F_k(x) \rangle\right).$$

The matrix variate normal

Define $\mathbf{F} = [F_1 F_2 \cdots F_K]$ as a $K \times T$ matrix and

$$p(\mathbf{F} \mid \mathbf{A}, \mathbf{U}, \mathbf{V}) = \frac{\exp\left(-\frac{1}{2}\mathrm{tr}[\mathbf{V}^{-1}(\mathbf{F}-\mathbf{A})^T\mathbf{U}^{-1}(\mathbf{F}-\mathbf{A})]\right)}{(2\pi)^{KT/2}|\mathbf{V}|^{L/2}|\mathbf{U}|^{K/2}},$$

A models mean
U models covariance between curves
V models covariance between points in a curve.

The matrix variate normal

Define $\mathbf{F} = [F_1 F_2 \cdots F_K]$ as a $K \times T$ matrix and

$$p(\mathbf{F} \mid \mathbf{A}, \mathbf{U}, \mathbf{V}) = \frac{\exp\left(-\frac{1}{2}\mathrm{tr}[\mathbf{V}^{-1}(\mathbf{F}-\mathbf{A})^T\mathbf{U}^{-1}(\mathbf{F}-\mathbf{A})]\right)}{(2\pi)^{KT/2}|\mathbf{V}|^{L/2}|\mathbf{U}|^{K/2}},$$

A models mean
U models covariance between curves
V models covariance between points in a curve.

The given n meshes $(M_1, ..., M_n)$ we can define a likelihood model

$$\mathrm{Lik}(M_1, ..., M_n \mid \mathbf{A}, \mathbf{U}, \mathbf{V}) = \prod_{i=1}^{n} p(\mathbf{F}(M_i) \mid \mathbf{A}, \mathbf{U}, \mathbf{V}), \tag{4}$$

Distances without alignment

Theorem (Turner-M)

Let $f : S^2 \to L_2(\mathbb{R})$ *and* $g : S^2 \to L_2(\mathbb{R})$ *be the ECT for two finite simplicial complexes* M_f *and* M_g *respectively. Both* f *and* g *are generically injective. Let* μ *be the measure on* S^2*. If* $f_*(\mu) = g_*(\mu)$*, the push forwards of the measure are equal, then there is some* $X \in O(3)$ *such that* $M_g = X(M_f)$*.*

The distributions of the Euler characteristic curves are sufficient statistics.

Picture of heel bone

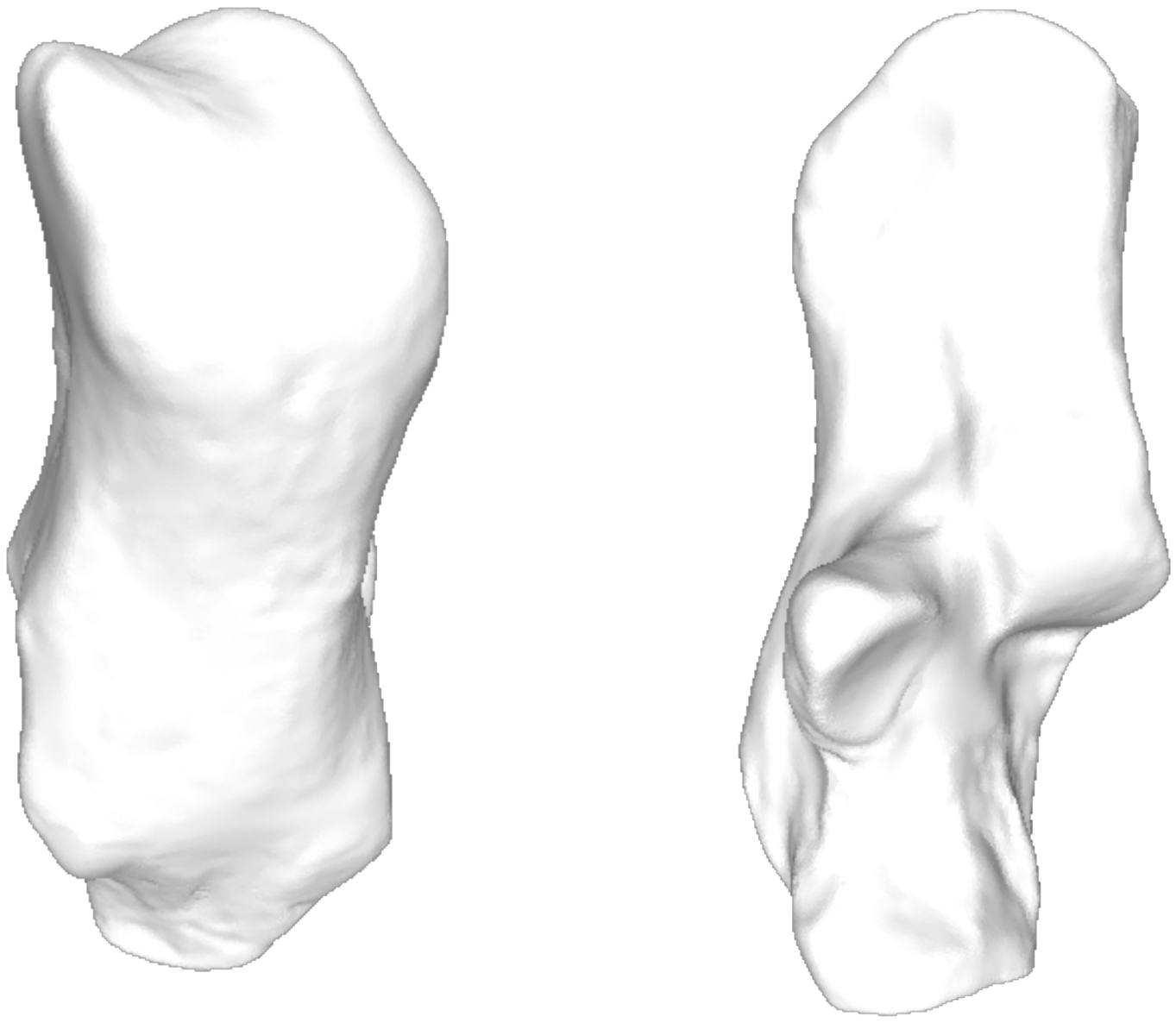
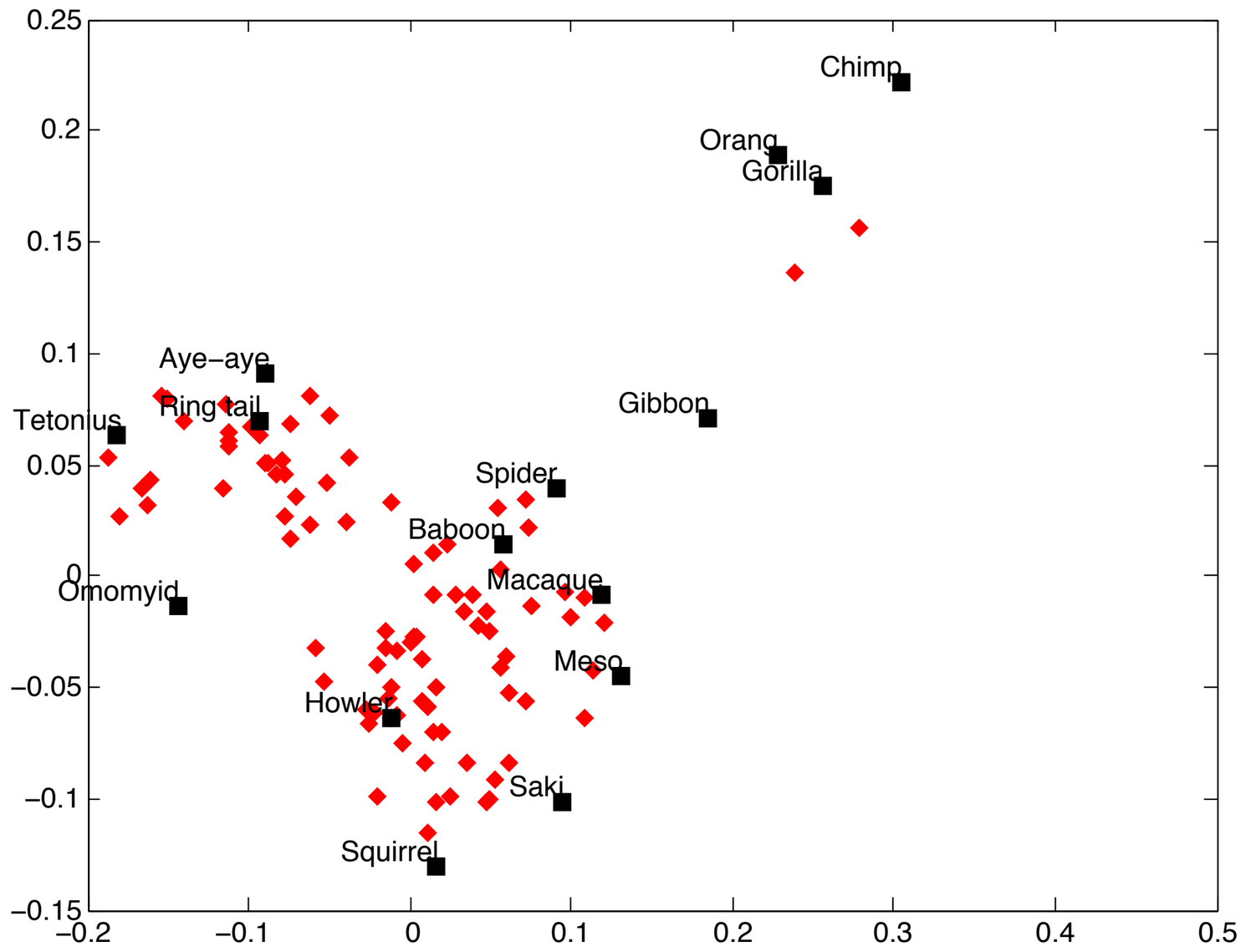


Figure : Images of a calcaneus from two different angles.

106 primates



Primate calcanei

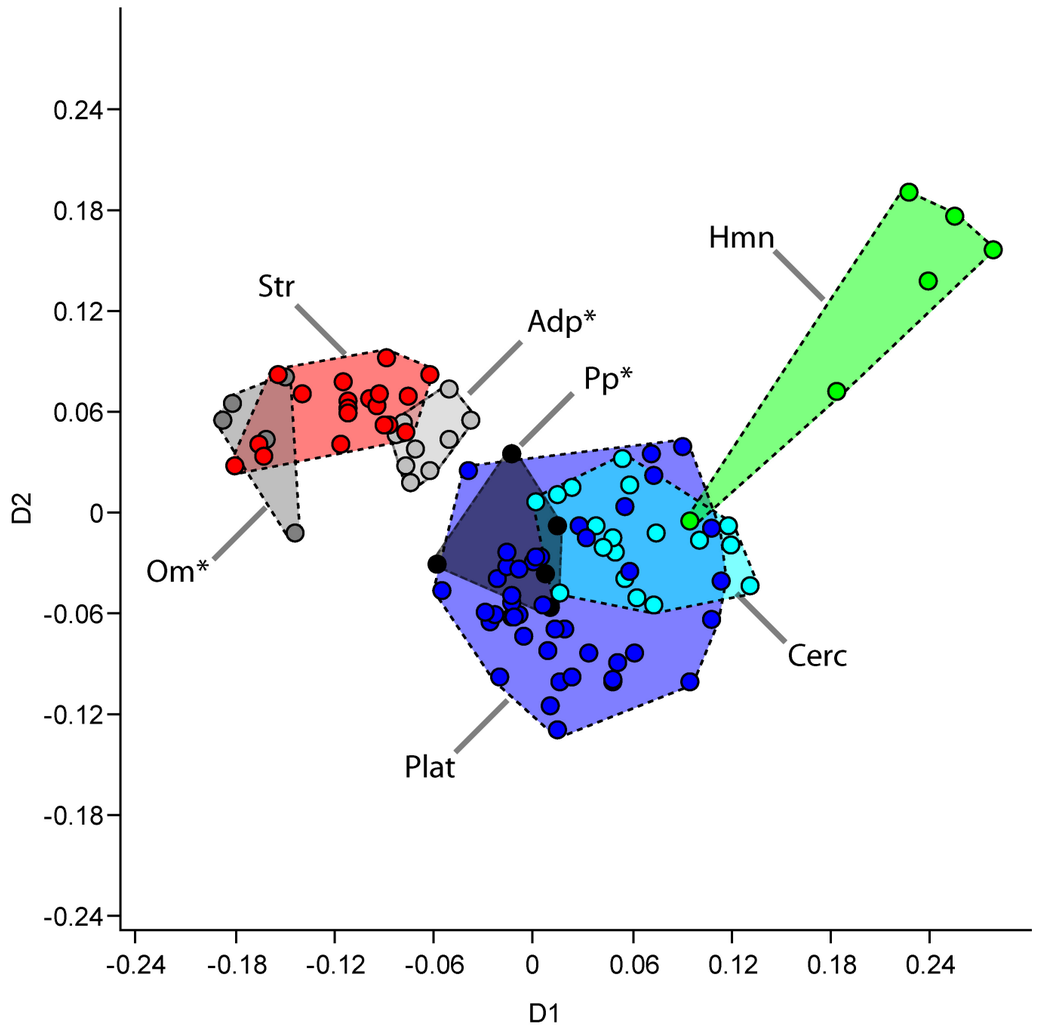
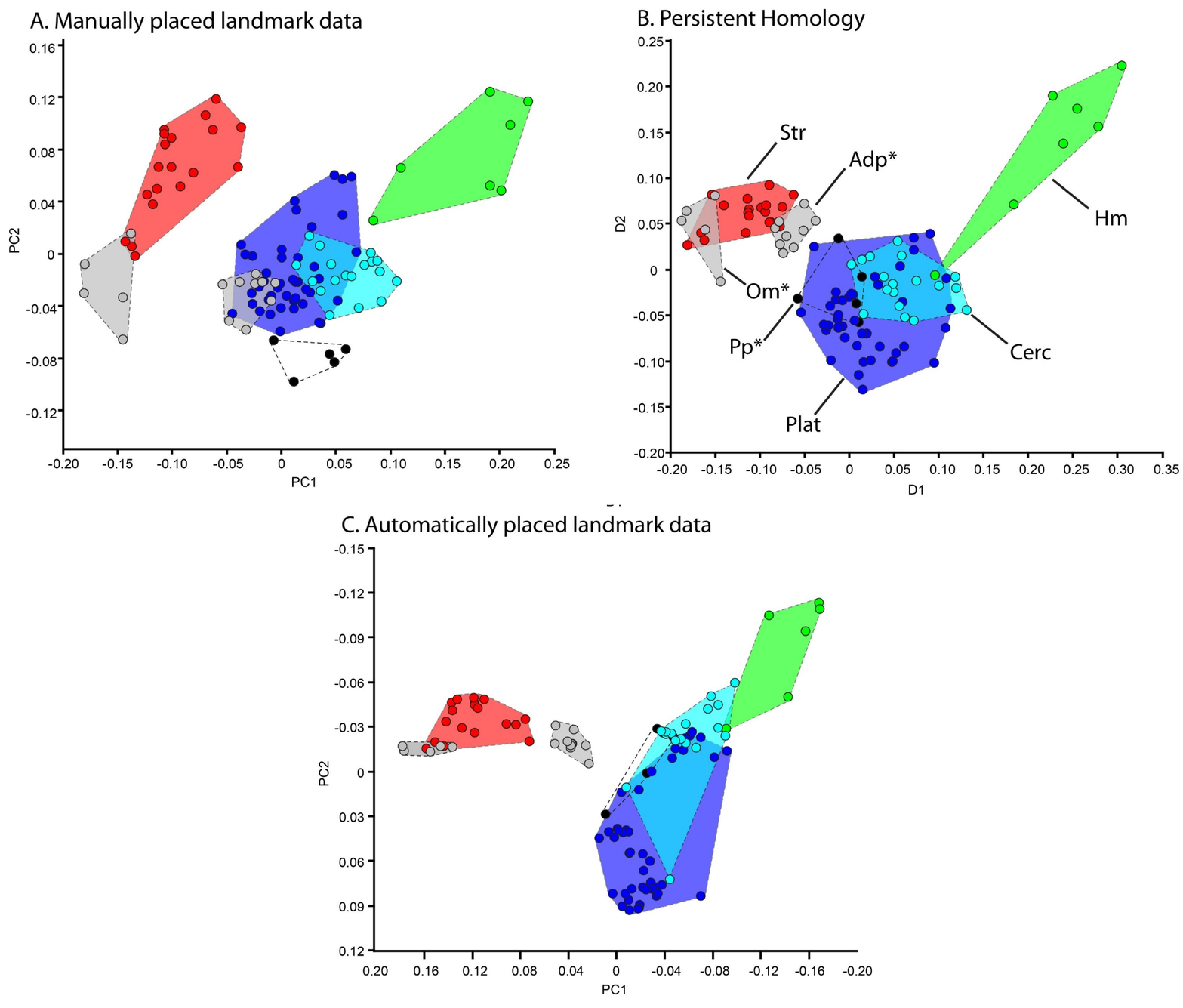


Figure : Phenetic clustering of phylogenetic groups of primate calcanei ($n = 106$). 67 genera are represented. Asterisks indicate groups of extinct taxa. Abbreviations: Str, Strepsirrhines; Plat, platyrrhines; Cerc, Cercopithecoids; Om, Omomyiforms; Adp, Adapiforms; Pp, parapithecids; Hmn, Hominoids. Note that more primitive prosimian taxa cluster separately from simians (Om, Adp, Str.). Also note that monkeys (Plat, Cerc, Pp) cluster mainly separately from apes (Hmn).

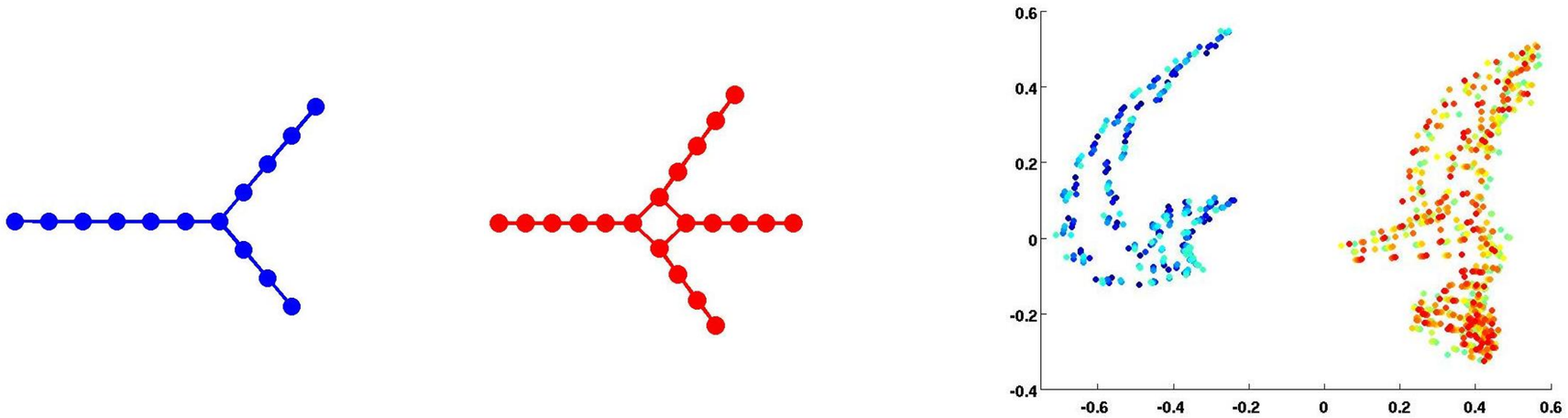
Comment from Doug

”In at least one way the method matched shapes with family groups better than any of the other previous methods... it linked a Hylobates specimen with the the other ape specimens (pan, gorilla, pongo, and oreopithecus). Previous both hylobatids (which ARE apes) always ended up closest to some Alouatta specimens.”

Comparing methods

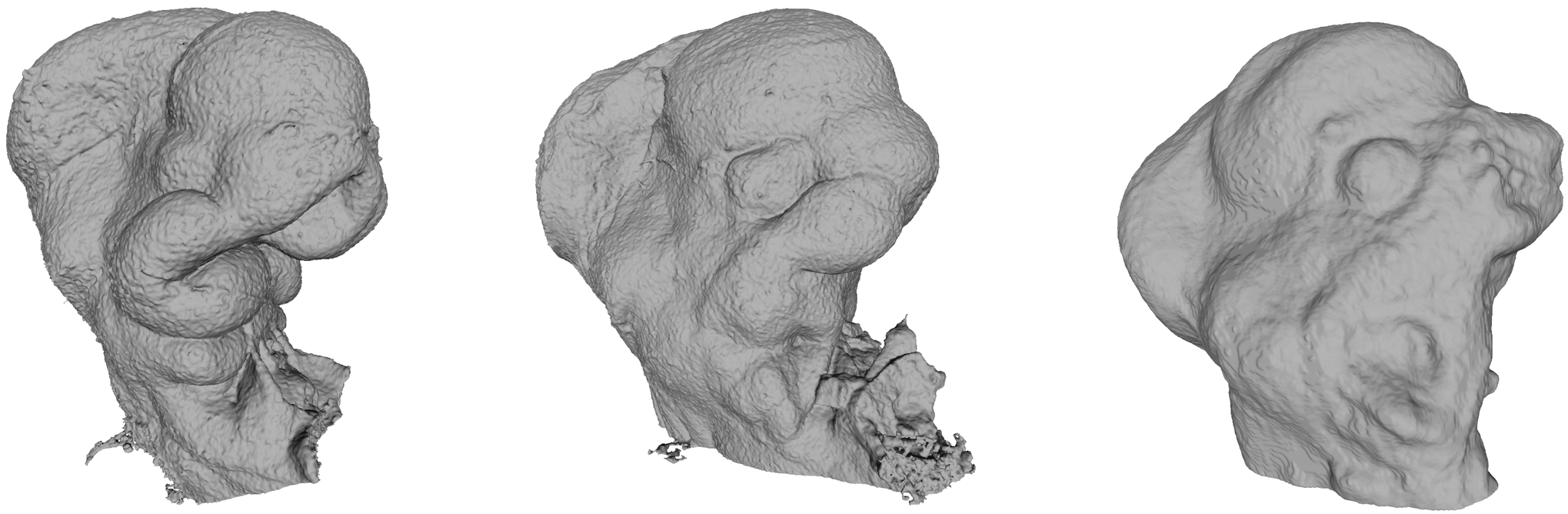


Can you hear the shape of a drum ?

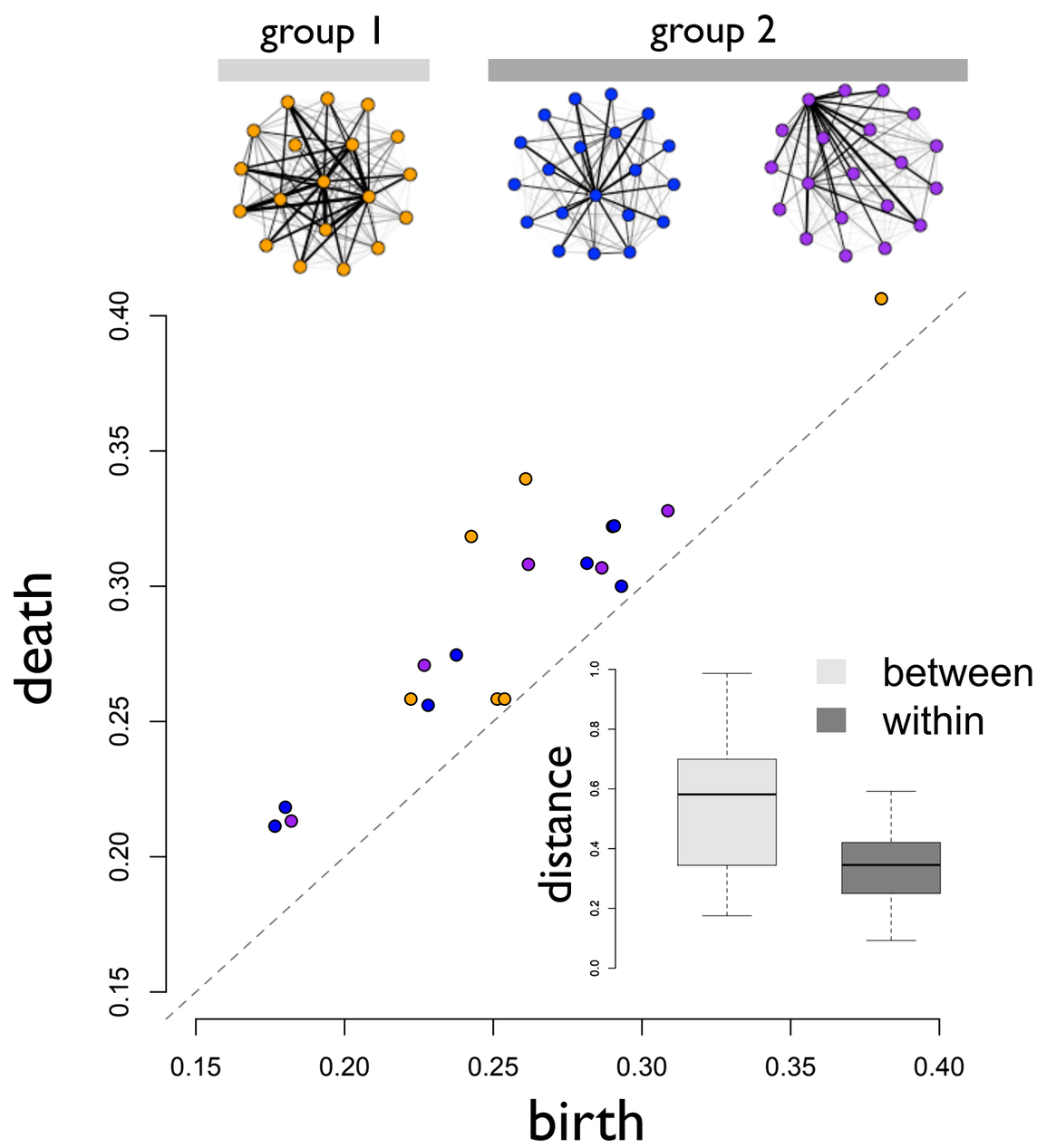


Mao Li

Association studies of shape phenotypes



Variation in baboon microbiome networks



Open problems

(1) Automatic alignment.

Open problems

(1) Automatic alignment.

(2) Correspondence.

Open problems

(1) Automatic alignment.

(2) Correspondence.

(3) Signal processing theory for surfaces based on Euler integration.

Open problems

(1) Automatic alignment.

(2) Correspondence.

(3) Signal processing theory for surfaces based on Euler integration.

(5) Maps between networks - relation between behavioral networks and genetic networks.

Open problems

(1) Automatic alignment.

(2) Correspondence.

(3) Signal processing theory for surfaces based on Euler integration.

(5) Maps between networks - relation between behavioral networks and genetic networks.

(6) Combine the two parts of this talk.

Acknowledgements

Thanks !!

Acknowledgements

Thanks !!

Funding:

- Center for Systems Biology at Duke
- NSF DMS and CCF
- DARPA
- AFOSR
- NIH

Simulation procedure

Simulate from

$$\mathbf{Y} = \mathbf{XB} + \mathbf{ZU} + \mathbf{E},$$

with $\mathbf{Z} = \mathbf{I}_n$, $\mathbf{B} = \mathbf{0}_p$, and $\mathbf{X} = \mathbf{1}$.

Simulation procedure

Simulate from

$$\mathbf{Y} = \mathbf{XB} + \mathbf{ZU} + \mathbf{E},$$

with $\mathbf{Z} = \mathbf{I}_n$, $\mathbf{B} = \mathbf{0}_p$, and $\mathbf{X} = \mathbf{1}$.

Effect of **G** and **R** in above equation on inference.

Traits measured on the off-spring of a balanced paternal half-sib breeding design.

Simulation parameters

	# factors			**R** type		# traits		Sample size		
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
G and **R**										
# traits	100	100	100	100	100	20	1000	100	100	100
Residual type	SF[a]	SF	SF	F[b]	Wishart[c]	SF	SF	SF	SF	SF
# factors	10	25	50	10	5	10	10	10	10	10
h^2 of factors[d]	0.5(5) 0.0(5)	0.5(15) 0.0(10)	0.5(30) 0.0(20)	0.5(5) 0.0(5)	1.0(5)	0.5(5) 0.0(5)		0.9-0.1(5) 0.0(5)		
Sample Size										
# sires	100	100	100	100	100	100	100	50	100	500
# offspring/sire	10	10	10	10	10	10	10	5	10	10

[a] **R** – sparse factor.

[b] **R** – factor.

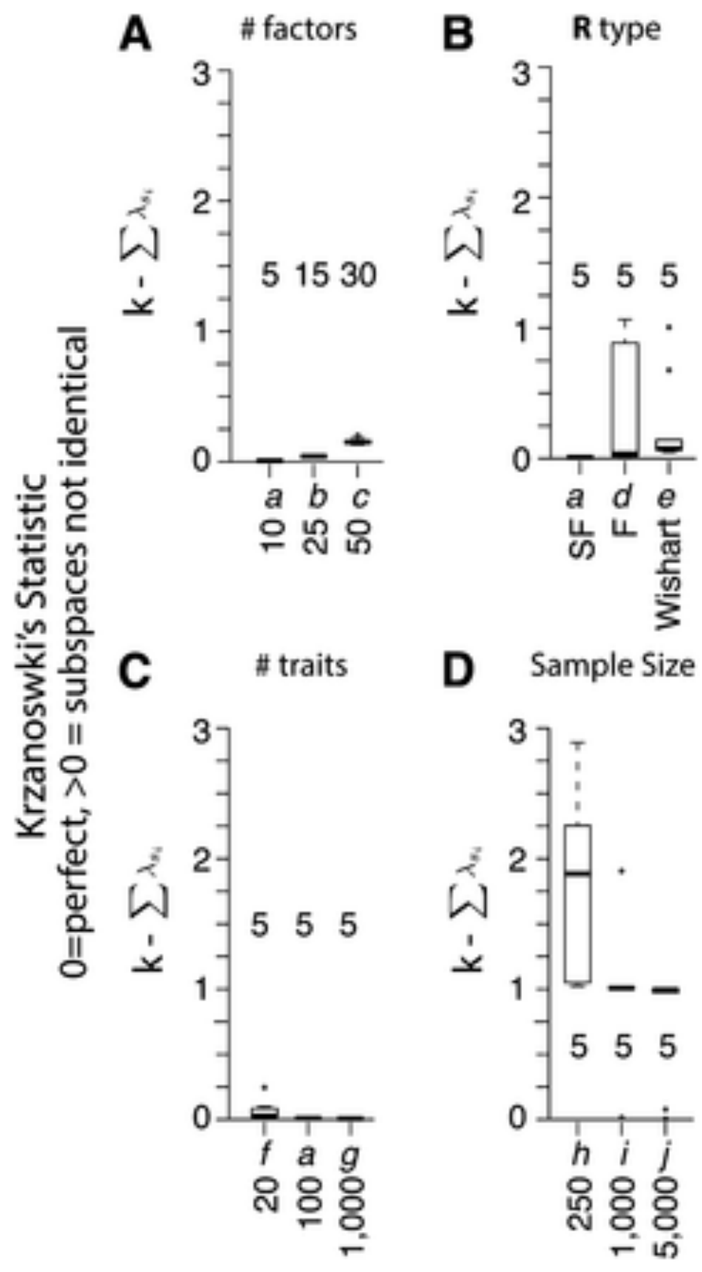
. [c] **R** – Wishart

[d] number of heritable factors.

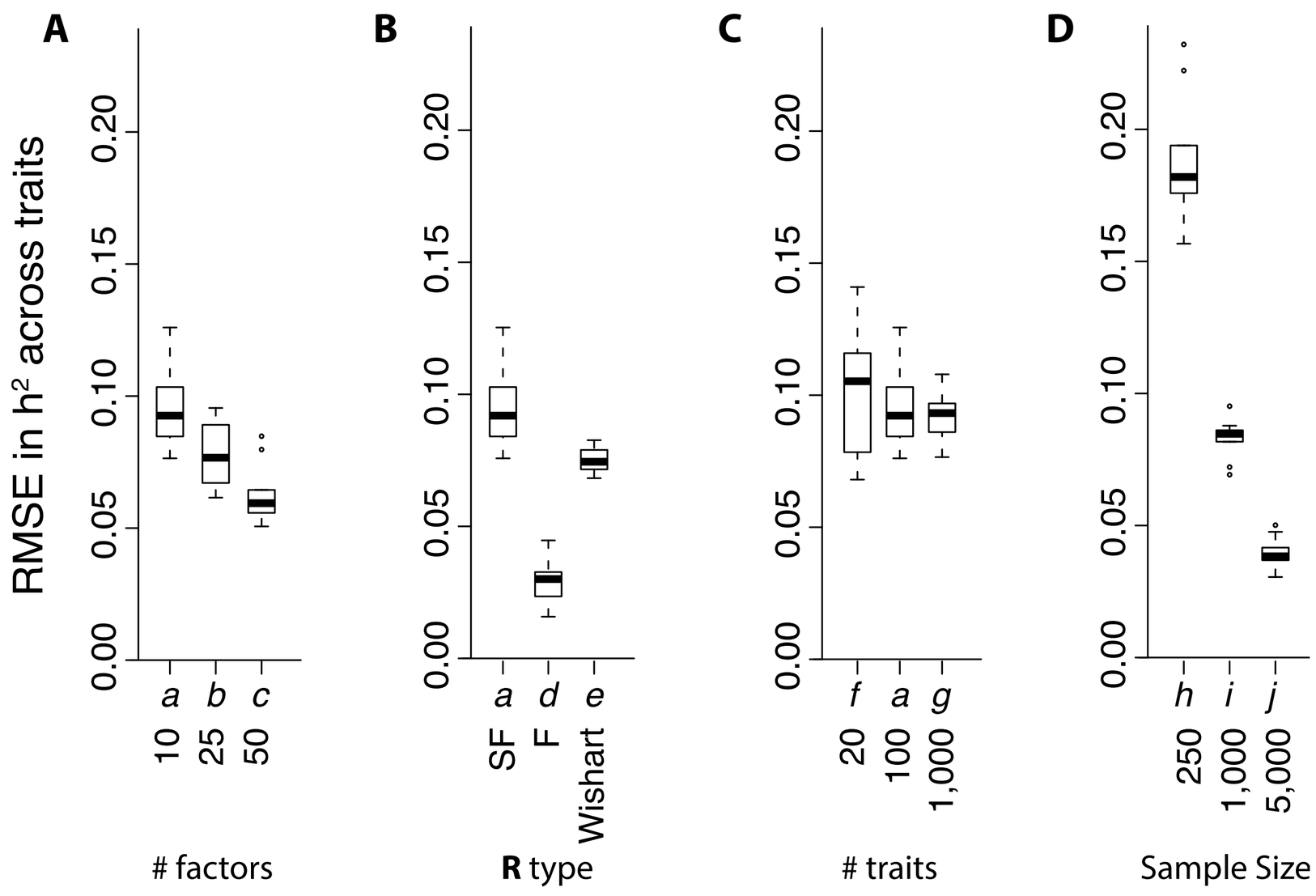
Recovering factors

Scenario		Expected	Median	Range
# factors	*a*	10	10	(10,10)
	b	25	25	(23,25)
	c	50	49	(48,50)
R type	*d*	10	10	(10,10)
	e	NA	56	(44,66)
# traits	*f*	10	9	(8,11)
	g	10	10	(10,10)
Sample size	*h*	10	10	(10,10)
	i	10	10	(10,10)
	j	10	10	(10,10)

Factor heritability



Trait heritability



$\mathcal{D}$ as a metric space

Alexandrov space bounded from below: Given a geodesic space $\mathbb{X}$ with metric d' for any geodesic $\gamma : [0,1] \to \mathbb{X}$ from X to Y and any $Z \in \mathbb{X}$

$$d'(Z,\gamma(t))^2 \geq t d'(Z,Y)^2 + (1-t)d'(Z,X)^2 - t(1-t)d'(X,Y)^2.$$

$\mathcal{D}$ as a metric space

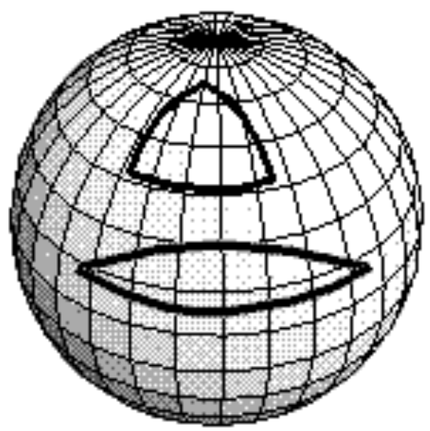
Alexandrov space bounded from below: Given a geodesic space $\mathbb{X}$ with metric d' for any geodesic $\gamma : [0,1] \to \mathbb{X}$ from X to Y and any $Z \in \mathbb{X}$

$$d'(Z,\gamma(t))^2 \geq t d'(Z,Y)^2 + (1-t)d'(Z,X)^2 - t(1-t)d'(X,Y)^2.$$

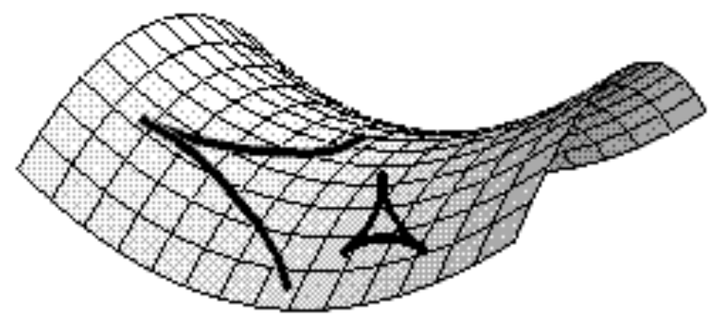
Theorem (Turner-Milyeko-M-Harer)

$(\mathcal{D}, d_{L^2})$ is a geodesic space and is a non-negatively curved Alexandrov space.

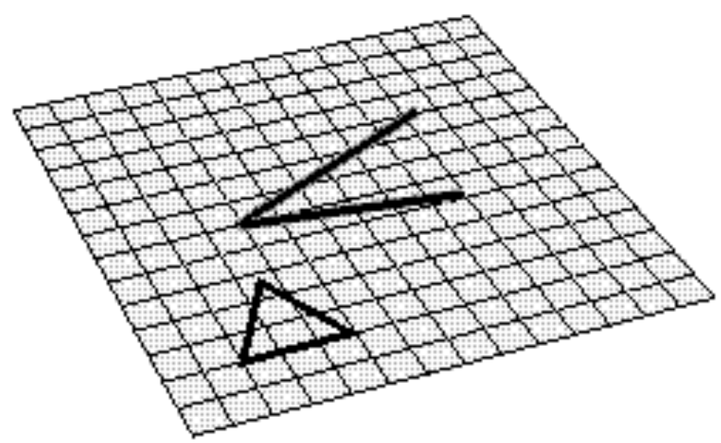
Comparison triangles



Universe with *positive* curvature. Diverging line converge at great distances. Triangle angles add to more than 180°.



Universe with *negative* curvature. Lines diverge at ever increasing angles. Triangle angles add to less than 180°.



Universe with no curvature. Lines diverge at constant angle. Triangle angles add to 180°.