Modeling Complex Phenotypes

USF, Mathematics & Statistics Colloquim

Sayan Mukherjee

Departments of Statistical Science, Mathematics, Computer Science
Biostatistics & Bioinformatics
Duke University

www.stat.duke.edu/~sayan

Mapping epsitasis	—	L. Crawford (Brown) | K Wood (Duke) | X. Zhou (U Michigan)
Digital phenotyping	—	K. Turner (ANU) | D. Boyer (Duke)
	—	A. Monod, R. Rabadán, A. Chen (Columbia) | L. Crawford (Brown)

Two stories

(1) Modeling epistatic effects.

Two stories

(1) Modeling epistatic effects.

(2) Digitizing shape phenotypes.

Modeling epistasis

Epistasis

Epistasis–interactions between genes or interaction between alleles at different loci.

Epistasis

Epistasis–interactions between genes or interaction between alleles at different loci.

Epistasis first defined by Bateson (1908) is an overloaded term:

(1) *Compositional epistasis*: Bateson observed some mutations seemed to be stopping or standing on the effects of other mutations. Such mutations were said to be epistatic (the ones being blocked, hypostatic).

Epistasis

Epistasis–interactions between genes or interaction between alleles at different loci.

Epistasis first defined by Bateson (1908) is an overloaded term:

(1) *Compositional epistasis*: Bateson observed some mutations seemed to be stopping or standing on the effects of other mutations. Such mutations were said to be epistatic (the ones being blocked, hypostatic).

(2) *Statistical epistasis*: Fisher coined the term epistacy to mean any statistical deviation from the additive combination of two loci in their effects on a phenotype.

Two important problems

Mapping traits: associating genetic loci to phenotypes. Relevant to genetic architecture of complex (disease) phenotypes.

Two important problems

Mapping traits: associating genetic loci to phenotypes. Relevant to genetic architecture of complex (disease) phenotypes.

Genomic selection: predicting phenotypic variation based on genotypic/genetic variation—relevant to animal and crop breeding programs.

The data

Trait: $\mathbf{Y}$ is $n \times 1$ with each $Y_i \in \mathbb{R}$ (crop yield):

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

The data

Trait: $\mathbf{Y}$ is $n \times 1$ with each $Y_i \in \mathbb{R}$ (crop yield):

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Genotype: $\mathbf{X}$ is $n \times p$ with each $X_{ij} \in \{0, 1, 2\}$ (# minor alleles):

$$\mathbf{X} = \begin{array}{c} \text{loci} \\ \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix} \end{array} \begin{array}{c} \\ \\ \\ \\ \\ \text{individuals} \end{array}$$

Trait mapping–linear models

Linear models used for GWAS and eQTL mapping

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \Sigma_n),$$

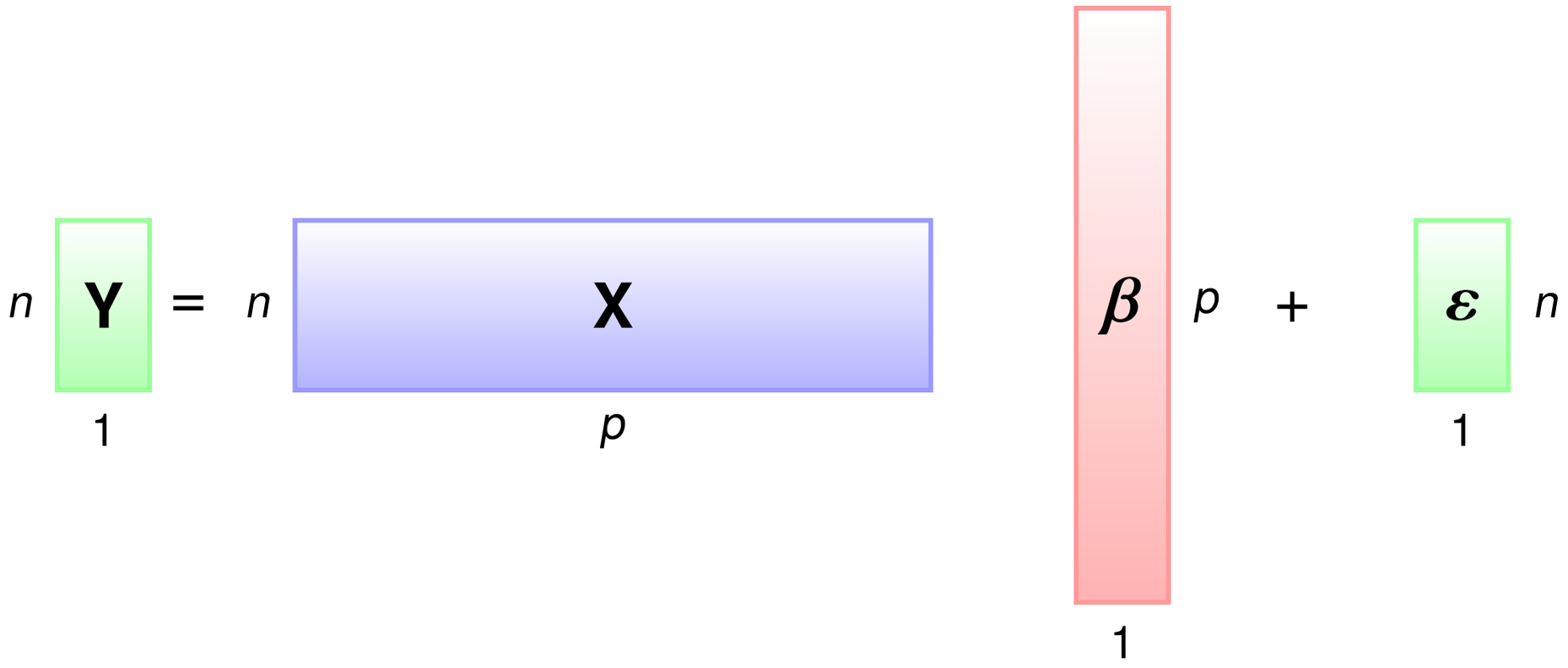
magnitude of $|\boldsymbol{\beta}_j|$ is evidence for association of locus j.

Trait mapping–linear models

Linear models used for GWAS and eQTL mapping

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \Sigma_n),$$

magnitude of $|\boldsymbol{\beta}_j|$ is evidence for association of locus *j*.



Genomic selection–nonlinear models

Nonlinear functions perform better for genomic selection

$$Y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0, \quad f \in \mathcal{H}.$$

Genomic selection–nonlinear models

Nonlinear functions perform better for genomic selection

$$Y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0, \quad f \in \mathcal{H}.$$

Kernel or Gaussian process models: state-of-the-art nonlinear regression functions of the form

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i),$$

where $\mathbf{x}_i$ are the genotypes and the kernel function $k(\mathbf{u}, \mathbf{v})$ is a similarity measure, for example $k(\mathbf{u}, \mathbf{v}) = \exp(-\kappa^2 \|\mathbf{u} - \mathbf{v}\|^2)$.

Genomic selection–nonlinear models

Nonlinear functions perform better for genomic selection

$$Y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0, \quad f \in \mathcal{H}.$$

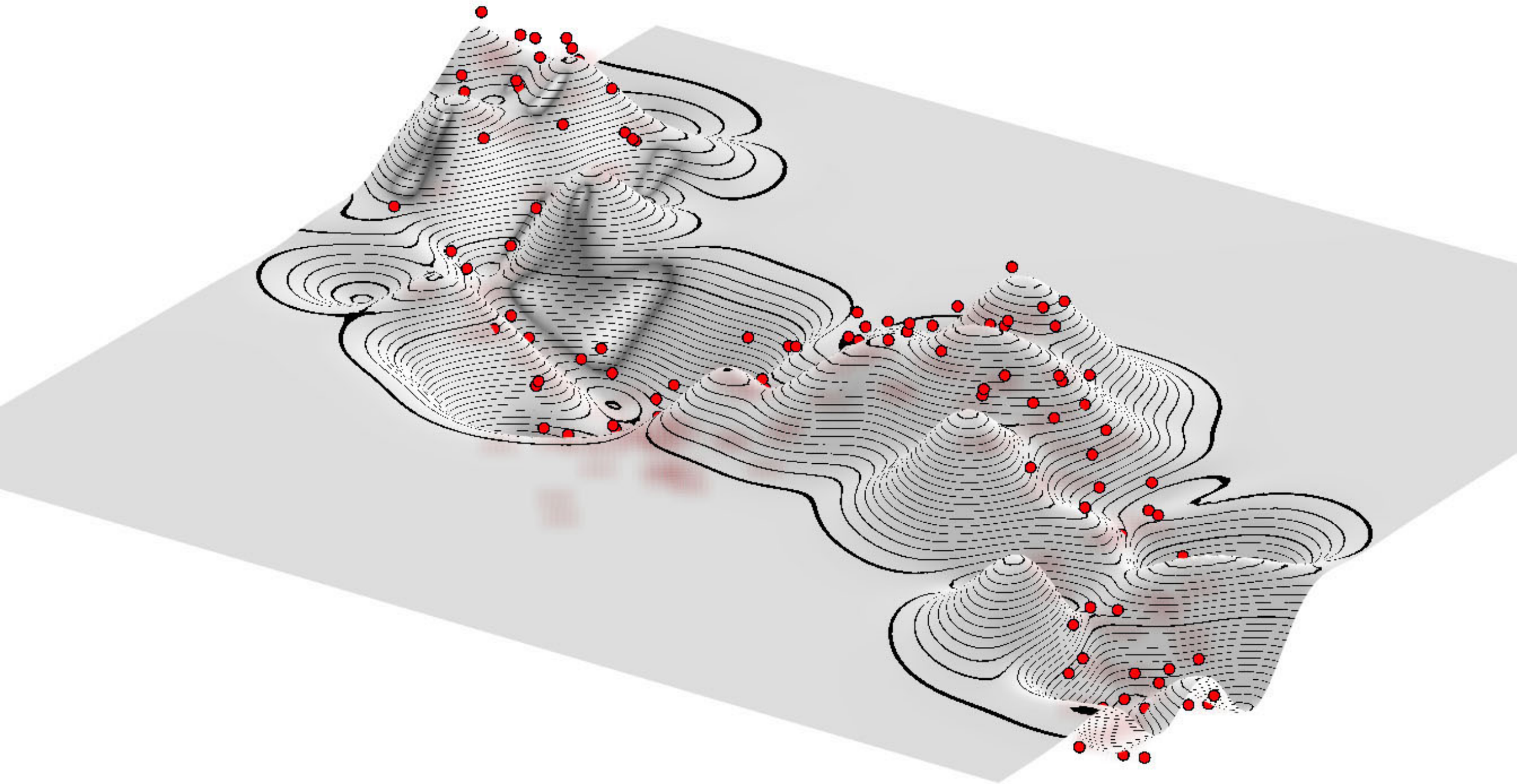
Kernel or Gaussian process models: state-of-the-art nonlinear regression functions of the form

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i),$$

where $\mathbf{x}_i$ are the genotypes and the kernel function $k(\mathbf{u}, \mathbf{v})$ is a similarity measure, for example $k(\mathbf{u}, \mathbf{v}) = \exp(-\kappa^2 \|\mathbf{u} - \mathbf{v}\|^2)$.

Effect of a locus (β) ?

Kernel models



Linear vs. nonlinear

Statistical conventional wisdom:

(1) In high-dimensional regression smooth nonlinear functions are more predictive than linear functions.

(2) Variable selection is much easier for linear regression.

Linear vs. nonlinear

Statistical conventional wisdom:

(1) In high-dimensional regression smooth nonlinear functions are more predictive than linear functions.

(2) Variable selection is much easier for linear regression.

Genomic conventional wisdom:

(1) Use nonlinear models for genomic selection.

(2) Use linear models for QTL mapping.

Our objective

Develop a single model for trait mapping and genomic selection which

(1) Predicts as well as state-of-the-art in genomic selection.

(2) Scalable trait mapping or variable selection.

Two observations

High-dimensional $p \gg n$ setting

Two observations

High-dimensional $p \gg n$ setting

(1) Smooth nonlinear functions predict more accurately than linear functions and sharply varying nonlinear functions.

Two observations

High-dimensional $p \gg n$ setting

(1) Smooth nonlinear functions predict more accurately than linear functions and sharply varying nonlinear functions.

(2) A smooth nonlinear function can be reasonably approximated by a linear function.

Reproducing kernel Hilbert space

Key idea: a penalized loss function for estimation

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 + \lambda ||f||_K^2 \right], \quad \lambda > 0.$$

Reproducing kernel Hilbert space

Key idea: a penalized loss function for estimation

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 + \lambda ||f||_K^2 \right], \quad \lambda > 0.$$

Representer theorem

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i).$$

∞ dimensional optimization problem $\longrightarrow$ n dimensional problem.

Reproducing kernel Hilbert space

Key idea: a penalized loss function for estimation

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 + \lambda ||f||_K^2 \right], \quad \lambda > 0.$$

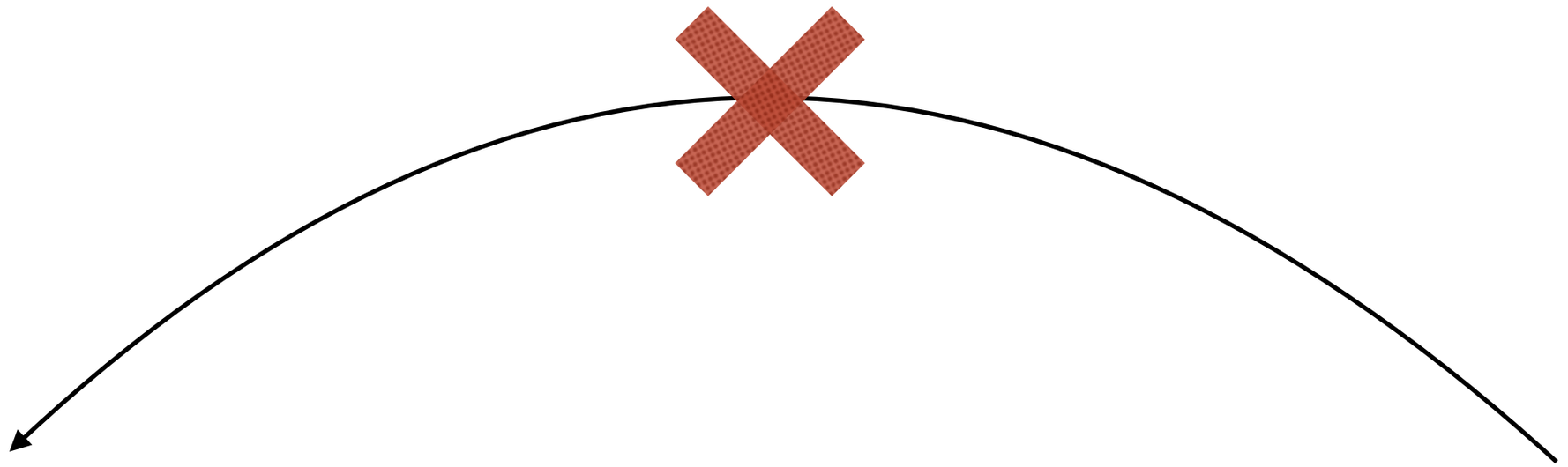
Representer theorem

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i).$$

∞ dimensional optimization problem $\longrightarrow$ n dimensional problem.

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \dots & k(\mathbf{x}_1, \mathbf{x}_n) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \dots & k(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & k(\mathbf{x}_n, \mathbf{x}_2) & \dots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

Kernel methods



Original p-dimensional space

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Downside: Classic idea of variable selection is lost

n-dimensional kernel space

$$\mathbf{y} = \mathbf{K}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

Proposal

A Bayesian kernel regression model that provides an analogue for the effect size of each explanatory variable in the original high dimensional space.

Subspace of RKHS

The subspace of the RKHS realized by the representer theorem

$$\mathcal{H}_{\mathbf{X}} = \left\{ f \,|\, f(\mathbf{x}) = \sum_{j=1}^{n} \alpha_j k(\mathbf{x}_j, \mathbf{x}), \quad \{\alpha_j\} \in \mathbb{R}^n, \quad \|f\|_K < \infty \right\}.$$

Subspace of RKHS

The subspace of the RKHS realized by the representer theorem

$$\mathcal{H}_{\mathbf{X}} = \left\{ f \,|\, f(\mathbf{x}) = \sum_{j=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}), \quad \{\alpha_i\} \in \mathbb{R}^n, \quad \|f\|_K < \infty \right\}.$$

An alternate parameterization

$$\mathcal{H}_{\mathbf{X}} = \left\{ f \,|\, f(\mathbf{x}) = \boldsymbol{\Psi}_{\mathbf{x}}^T \mathbf{c}, \quad \|f\|_K < \infty \right\}.$$

with eigenvalues and eigenfunctions

$$\lambda_j \psi_j(\mathbf{x}) = \int k(\mathbf{u}, \mathbf{x}) \psi_j(\mathbf{u}) \mathrm{d}\mathbf{u},$$

$\boldsymbol{\Psi}_{\mathbf{x},1} = \{ \sqrt{\lambda_i} \psi_j(\mathbf{x}_1) \}_{j=1}^{\infty}$ and $\boldsymbol{\Psi}_{\mathbf{x}} = \left[\boldsymbol{\Psi}_{\mathbf{x},1} \cdots \boldsymbol{\Psi}_{\mathbf{x},n} \right]^T$ and $\mathbf{c} = \{ \mathrm{c}_j \}_{j=1}^{\infty}$.

Subspace of RKHS

The subspace of the RKHS realized by the representer theorem

$$\mathcal{H}_{\mathbf{X}} = \left\{ f \,|\, f(\mathbf{x}) = \sum_{j=1}^{n} \alpha_j k(\mathbf{x}_j, \mathbf{x}), \quad \{\alpha_j\} \in \mathbb{R}^n, \quad \|f\|_K < \infty \right\}.$$

An alternate parameterization

$$\mathcal{H}_{\mathbf{X}} = \left\{ f \,|\, f(\mathbf{x}) = \mathbf{\Psi}_{\mathbf{x}}^T \mathbf{c}, \quad \|f\|_K < \infty \right\}.$$

where $\mathbf{\Psi}_{\mathbf{x}} = \{ \sqrt{\lambda_j} \psi_j(\mathbf{x}) \}_{j=1}^{\infty}$ and $\mathbf{c} = \{c_j\}_{j=1}^{\infty}$

$$\mathbf{K} = \mathbf{\Psi}_{\mathbf{x}}^T \mathbf{\Psi}_{\mathbf{x}} \Longrightarrow \mathbf{c} = \mathbf{\Psi}_{\mathbf{x}} \alpha.$$

Maps between spaces

This exists in theory, but is very hard to model in practice

Original p-dimensional space

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \varepsilon,$$

Bases Representation

$$\mathbf{Y} = \boldsymbol{\Psi}_{\mathbf{x}}^{T}\mathbf{c} + \varepsilon,$$

n-dimensional kernel space

$$\mathbf{Y} = \mathbf{K}\,\boldsymbol{\alpha} + \varepsilon,$$

Effect size for a smooth nonlinear functions

Can we define a reasonable β ?

Effect size for a smooth nonlinear functions

Can we define a reasonable β ?

How about projecting the nonlinear function onto the data

Effect size for a smooth nonlinear functions

Can we define a reasonable β ?

How about projecting the nonlinear function onto the data

$$\begin{aligned} \mathbf{X}\beta &\approx \boldsymbol{\Psi}_{\mathbf{x}}^{T}\mathbf{c} \\ \text{“}\beta &= \mathbf{X}^{-1}\boldsymbol{\Psi}_{\mathbf{x}}^{T}\mathbf{c}.\text{”} \end{aligned}$$

Random Fourier bases

We can specify useful maps between the feature space and predictor space using a randomized feature maps.

Bochner's theorem

Assume the kernel is shift-invariant and integrates to one:

$$k(\mathbf{u}, \mathbf{v}) = k(\mathbf{u} - \mathbf{v}), \quad \int k(\mathbf{z})\, \mathrm{d}\mathbf{z} = 1,\ \mathbf{z} = \mathbf{u} - \mathbf{v}.$$

Bochner's theorem

Assume the kernel is shift-invariant and integrates to one:

$$k(\mathbf{u}, \mathbf{v}) = k(\mathbf{u} - \mathbf{v}), \quad \int k(\mathbf{z})\, \mathrm{d}\mathbf{z} = 1,\ \mathbf{z} = \mathbf{u} - \mathbf{v}.$$

Bochner's theorem

$$k\left(\mathbf{x}_i - \mathbf{x}_j\right) = \int_{\mathbb{R}^p} f(\omega) \exp\left\{\iota\, \omega^T (\mathbf{x}_i - \mathbf{x}_j)\right\} d\omega = \mathbb{E}_\omega[\eta_\omega(\mathbf{x}_i)\, \eta_\omega(\mathbf{x}_j)^*],$$

$$\eta_\omega(\mathbf{x}_i) = \exp\left(\iota\, \omega^T \mathbf{x}_i\right).$$

Bochner's theorem

Assume the kernel is shift-invariant and integrates to one:

$$k(\mathbf{u}, \mathbf{v}) = k(\mathbf{u} - \mathbf{v}), \quad \int k(\mathbf{z})\, \mathrm{d}\mathbf{z} = 1,\ \mathbf{z} = \mathbf{u} - \mathbf{v}.$$

Bochner's theorem

$$k\left(\mathbf{x}_i - \mathbf{x}_j\right) = \int_{\mathbb{R}^p} f(\omega) \exp\left\{\iota\, \omega^T (\mathbf{x}_i - \mathbf{x}_j)\right\} d\omega = \mathbb{E}_\omega[\eta_\omega(\mathbf{x}_i)\, \eta_\omega(\mathbf{x}_j)^*],$$

$$\eta_\omega(\mathbf{x}_i) = \exp\left(\iota\, \omega^T \mathbf{x}_i\right).$$

The Fourier transform of the kernel function is a probability density

$$f(\omega) = \int_{\mathcal{X}} k(\mathbf{x}) e^{-\iota 2\pi \omega^T \mathbf{x}}\, \mathrm{d}\mathbf{x}.$$

Approximate kernel

Given a sample $(\omega_1, ..., \omega_p) \overset{iid}{\sim} f(\omega)$ consider

$$k(\mathbf{u}, \mathbf{v}) = \mathbb{E}_\omega[\eta_\omega(\mathbf{u})\, \eta_\omega(\mathbf{v})^*] \approx \frac{1}{p} \sum_{j=1}^{p} \left[\eta_{\omega_j}(\mathbf{u})\, \eta_{\omega_j}(\mathbf{v})^*\right].$$

Approximate kernel

A sampling procedure

$$\omega_\ell \overset{iid}{\sim} f(\omega), \quad \mathrm{b}_\ell \overset{iid}{\sim} U[0, 2\pi], \quad \ell = 1, ..., p$$

$$\mathbf{\Omega} = [\omega_1, \ldots, \omega_p] \in \mathbb{R}^{p \times p}, \quad \mathbf{b} = [\mathrm{b}_1, \ldots, \mathrm{b}_p]^T \in \mathbb{R}^d,$$

$$\mathbf{z}(\mathbf{x}_i) = \sqrt{\frac{2}{p}} \cos\left(\mathbf{x}_i^T \mathbf{\Omega} + \mathbf{b}\right), \quad \tilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{z}(\mathbf{x}_i)^T \mathbf{z}(\mathbf{x}_j)$$

Approximate kernel

A sampling procedure

$$\omega_\ell \overset{iid}{\sim} f(\omega), \quad \mathrm{b}_\ell \overset{iid}{\sim} U[0, 2\pi], \quad \ell = 1, ..., p$$

$$\mathbf{\Omega} = [\omega_1, \ldots, \omega_p] \in \mathbb{R}^{p \times p}, \quad \mathbf{b} = [\mathrm{b}_1, \ldots, \mathrm{b}_p]^T \in \mathbb{R}^d,$$

$$\mathbf{z}(\mathbf{x}_i) = \sqrt{\frac{2}{p}} \cos\left(\mathbf{x}_i^T \mathbf{\Omega} + \mathbf{b}\right), \quad \tilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{z}(\mathbf{x}_i)^T \mathbf{z}(\mathbf{x}_j)$$

Analog of a design matrix $\mathbf{Z} = [\mathbf{z}(\mathbf{x}_1), \ldots, \mathbf{z}(\mathbf{x}_n)]$ and approximate kernel matrix $\widetilde{\mathbf{K}} = \mathbf{Z}^T\mathbf{Z}$, and coefficients $\mathbf{c}$ where

$$\sum_{j=1}^{n} \alpha_i \tilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{z}(\mathbf{x}_i)^T \mathbf{c}, \quad \forall i = 1, ..., n.$$

Effect size for a smooth nonlinear functions

The nonlinear regression is

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \tilde{k}(\mathbf{x}, \mathbf{x}_i), \quad f(\mathbf{x}) = \mathbf{z}(\mathbf{x})^T \mathbf{c}$$

Effect size for a smooth nonlinear functions

The nonlinear regression is

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \tilde{k}(\mathbf{x}, \mathbf{x}_i), \quad f(\mathbf{x}) = \mathbf{z}(\mathbf{x})^T \mathbf{c}$$

Project the nonlinear function onto the data

Effect size for a smooth nonlinear functions

The nonlinear regression is

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \tilde{k}(\mathbf{x}, \mathbf{x}_i), \quad f(\mathbf{x}) = \mathbf{z}(\mathbf{x})^T \mathbf{c}$$

Project the nonlinear function onto the data

$$\begin{aligned} \mathbf{X}\beta &\approx \mathbf{Z}^T \mathbf{c} \\ \beta &= \mathbf{X}^{-1} \mathbf{Z}^T \mathbf{c}. \end{aligned}$$

Maps between spaces

Original p-dimensional space

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \varepsilon,$$

Basis Representation

$$\mathbf{Y} = \boldsymbol{\Psi}_{\mathbf{x}}^{T}\mathbf{c} + \varepsilon, \longrightarrow$$

n-dimensional kernel space

$$\mathbf{Y} = \mathbf{K}\,\boldsymbol{\alpha} + \varepsilon,$$

Maps between spaces

Original p-dimensional space

Basis Representation

n-dimensional approx. kernel space

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \varepsilon, \quad \longleftrightarrow \quad \mathbf{Y} = \mathbf{Z}^T\mathbf{c} + \varepsilon \quad \longleftrightarrow \quad \mathbf{Y} = \mathbf{K}\,\boldsymbol{\alpha} + \varepsilon,$$

Generalized kernel model

We can specify the generalized kernel model as

$$y_i \sim p(y \mid \mu_i), \qquad \mu_i = g(\mathbf{z}_i^T \mathbf{c}).$$

Generalized kernel model

We can specify the generalized kernel model as

$$y_i \sim p(y \mid \mu_i), \qquad \mu_i = g(\mathbf{z}_i^T \mathbf{c}).$$

In terms of the approximate kernel

$$y_i \sim p(y \mid \mu_i), \qquad \mu_i = g(\mathbf{k}_i^T \boldsymbol{\alpha}).$$

where

$$\mathbf{k}_i = \mathbf{K}_{i\cdot} = \left[\tilde{k}(\mathbf{x}_i, \mathbf{x}_1), ..., \tilde{k}(\mathbf{x}_i, \mathbf{x}_n)\right]^T.$$

Generalized factored kernel model

K is symmetric and posiitve (semi) definite:

$$\mathbf{K} = \mathbf{U}\,\mathbf{\Lambda}\,\mathbf{U}^T,$$

U is an $n \times n$ orthogonal matrix of eigenvectors
$\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix of eigenvalues.

Generalized factored kernel model

K is symmetric and posiitve (semi) definite:

$$\mathbf{K} = \mathbf{U}\,\mathbf{\Lambda}\,\mathbf{U}^T,$$

U is an $n \times n$ orthogonal matrix of eigenvectors
$\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix of eigenvalues.

Factor kernel regression model

$$y_i \sim p(y \mid \mu_i), \qquad \mu_i = g(\mathbf{u}_i^T \boldsymbol{\theta}),$$

where $\mathbf{u}_i = \mathbf{U}_{i\cdot}$ and $\boldsymbol{\theta} = \mathbf{\Lambda}\mathbf{U}^T\boldsymbol{\alpha}$.

Projection onto original covariates

Recall that

$$\alpha = \mathbf{K}\mathbf{Z}^T\mathbf{c}, \qquad \theta = \mathbf{\Lambda}\mathbf{U}^T\alpha.$$

Projection onto original covariates

Recall that

$$\alpha = \mathbf{K}\mathbf{Z}^T\mathbf{c}, \qquad \theta = \mathbf{\Lambda}\mathbf{U}^T\alpha.$$

This implies a “general effect size”

$$\mathbf{c} = (\mathbf{\Lambda}\mathbf{U}^T\mathbf{K}^{-1}\mathbf{Z}^T)^{-1}, \qquad \beta = \mathbf{X}^\dagger\mathbf{Z}^T\mathbf{c}.$$

Capturing nonlinear interactions

A warm up example

$$
\begin{aligned}
K(\mathbf{u},\mathbf{v}) &= (\mathbf{u}^T\mathbf{v}+c)^2 \\
&= \sum_{k=1}^{p}\left(\mathbf{u}_k^2\right)\left(\mathbf{v}_k^2\right) + \sum_{k=2}^{p}\sum_{j=1}^{k-1}\left(\sqrt{2}\mathbf{u}_k\mathbf{u}_j\right)\left(\sqrt{2}\mathbf{v}_k\mathbf{v}_j\right) \\
&\quad + \sum_{k=1}^{p}\left(\sqrt{2c}\mathbf{u}_k\right)\left(\sqrt{2c}\mathbf{v}_k\right) + c^2
\end{aligned}
$$

Capturing nonlinear interactions

Taylor expand $\mathbf{z}(\mathbf{x}_i)$ and $\mathbf{z}(\mathbf{x}_j)$

$$\mathbf{z}(\mathbf{x}_i)^T\mathbf{z}(\mathbf{x}_j) = \frac{2}{p}[\cos(\mathbf{\Omega}^T\mathbf{x}_i + \mathbf{b})\cos(\mathbf{x}_j^T\mathbf{\Omega} + \mathbf{b})]$$

$$= \frac{2}{p}\sum_{r=0}^{\infty}\sum_{t=0}^{\infty}(-1)^{r+t}\frac{(\mathbf{\Omega}^T\mathbf{x}_i + \mathbf{b})^{2r}(\mathbf{x}_j'\mathbf{\Omega} + \mathbf{b})^{2t}}{(2r!)(2t!)}.$$

Bayesian modeling and inference

The model parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ are a priori specified as $\pi(\boldsymbol{\theta})$.
The likelihood of the data $\mathbf{D}$ is specified as $\mathrm{Lik}(\mathbf{D} \mid \boldsymbol{\theta})$.
We want to compute the posterior of $\boldsymbol{\theta}$ given data $\mathbf{D}$

$$\mathrm{Post}(\boldsymbol{\theta} \mid \mathbf{D}) = \frac{\pi(\boldsymbol{\theta})\,\mathrm{Lik}(\mathbf{D} \mid \boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})\,\mathrm{Lik}(\mathbf{D} \mid \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} = \frac{\text{prior} \times \text{likelihood}}{\text{marginal likelihood}}.$$

Bayesian modeling and inference

The model parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ are a priori specified as $\pi(\boldsymbol{\theta})$.
The likelihood of the data $\mathbf{D}$ is specified as $\mathrm{Lik}(\mathbf{D} \mid \boldsymbol{\theta})$.
We want to compute the posterior of $\boldsymbol{\theta}$ given data $\mathbf{D}$

$$\mathrm{Post}(\boldsymbol{\theta} \mid \mathbf{D}) = \frac{\pi(\boldsymbol{\theta})\,\mathrm{Lik}(\mathbf{D} \mid \boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})\,\mathrm{Lik}(\mathbf{D} \mid \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} = \frac{\text{prior} \times \text{likelihood}}{\text{marginal likelihood}}.$$

Typically we cannot compute $\mathrm{Post}(\boldsymbol{\theta} \mid \mathbf{D})$ but we can sample

$$\boldsymbol{\theta} \sim \mathrm{Post}(\boldsymbol{\theta} \mid \mathbf{D}).$$

Model specification: nonlinear regression

Hierarchical factor model specification

$$y_i = \mathbf{u}_i^T \boldsymbol{\theta} + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right),$$

Model specification: nonlinear regression

Hierarchical factor model specification

$$y_i = \mathbf{u}_i^T \boldsymbol{\theta} + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right),$$

$$\boldsymbol{\theta} \sim \mathrm{MVN}(\mathbf{0}_q, \sigma_\theta^2 \boldsymbol{\Lambda}),$$

Model specification: nonlinear regression

Hierarchical factor model specification

$$
\begin{aligned}
y_i &= \mathbf{u}_i^T \boldsymbol{\theta} + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right), \\
\boldsymbol{\theta} &\sim \mathrm{MVN}(\mathbf{0}_q, \sigma_\theta^2 \boldsymbol{\Lambda}), \\
\sigma_\theta^2 &\sim \text{Scale-inv-}\chi^2(\nu_\theta, \phi_\theta),
\end{aligned}
$$

Model specification: nonlinear regression

Hierarchical factor model specification

$$
\begin{aligned}
y_i &= \mathbf{u}_i^T \boldsymbol{\theta} + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right), \\
\boldsymbol{\theta} &\sim \mathrm{MVN}(\mathbf{0}_q, \sigma_\theta^2 \boldsymbol{\Lambda}), \\
\sigma_\theta^2 &\sim \text{Scale-inv-}\chi^2(\nu_\theta, \phi_\theta), \\
\sigma_\varepsilon^2 &\sim \text{Scale-inv-}\chi^2(\nu_\varepsilon, \phi_\varepsilon).
\end{aligned}
$$

Model specification: nonlinear regression

Hierarchical factor model specification

$$
\begin{aligned}
y_i &= \mathbf{u}_i^T \boldsymbol{\theta} + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right), \\
\boldsymbol{\theta} &\sim \mathrm{MVN}(\mathbf{0}_q, \sigma_\theta^2 \boldsymbol{\Lambda}), \\
\sigma_\theta^2 &\sim \text{Scale-inv-}\chi^2(\nu_\theta, \phi_\theta), \\
\sigma_\varepsilon^2 &\sim \text{Scale-inv-}\chi^2(\nu_\varepsilon, \phi_\varepsilon).
\end{aligned}
$$

Implicit normal prior on $\boldsymbol{\beta}$

$$
\boldsymbol{\beta} \propto \exp\left\{ -\frac{1}{2\sigma_\theta^2} \boldsymbol{\beta}^T (\mathbf{B}\boldsymbol{\Lambda}\mathbf{B}^T)^{-1} \boldsymbol{\beta} \right\}, \quad \mathbf{B} = \mathbf{X}^\dagger \mathbf{Z}^T (\boldsymbol{\Lambda}\mathbf{U}^T \mathbf{K}^{-1} \mathbf{Z}^T)^{-1}.
$$

Model specification: nonlinear binary regression

Specify the following hierarchical model

$$y_i = \begin{cases} 1 & \text{if} \quad s_i > 0 \\ 0 & \text{if} \quad s_i \leq 0, \end{cases}$$

$$s_i = \mathbf{u}_i^T \boldsymbol{\theta} + \varepsilon_i, \quad \varepsilon_i \sim \mathsf{N}(0, 1),$$

Model specification: nonlinear binary regression

Specify the following hierarchical model

$$
\begin{aligned}
y_i &= \begin{cases} 1 & \text{if} \quad s_i > 0 \\ 0 & \text{if} \quad s_i \leq 0, \end{cases} \\
s_i &= \mathbf{u}_i^T \boldsymbol{\theta} + \varepsilon_i, \quad \varepsilon_i \sim \mathsf{N}(0, 1), \\
\boldsymbol{\theta} &\sim \mathsf{MVN}(\mathbf{0}_q, \sigma_\theta^2 \boldsymbol{\Lambda}),
\end{aligned}
$$

Model specification: nonlinear binary regression

Specify the following hierarchical model

$$
\begin{aligned}
y_i &= \begin{cases} 1 & \text{if} \quad s_i > 0 \\ 0 & \text{if} \quad s_i \leq 0, \end{cases} \\
s_i &= \mathbf{u}_i^T \boldsymbol{\theta} + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, 1), \\
\boldsymbol{\theta} &\sim \mathrm{MVN}(\mathbf{0}_q, \sigma_\theta^2 \boldsymbol{\Lambda}), \\
\sigma_\theta^2 &\sim \text{Scale-inv-}\chi^2(\nu_\theta, \phi_\theta).
\end{aligned}
$$

Model specification: nonlinear mixed model

Specify the following hierarchical model

$$\mathrm{y}_i = \mathbf{u}_i^T \boldsymbol{\theta} + \varphi_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right), \quad \varphi_i \perp\!\!\!\perp \varepsilon_i,$$

Model specification: nonlinear mixed model

Specify the following hierarchical model

$$y_i = \mathbf{u}_i^T \boldsymbol{\theta} + \varphi_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right), \quad \varphi_i \perp\!\!\!\perp \varepsilon_i,$$

$$\boldsymbol{\theta} \sim \mathrm{MVN}(\mathbf{0}_q, \sigma_\theta^2 \boldsymbol{\Lambda}),$$

$$\sigma_\theta^2 \sim \mathrm{Scale} - \mathrm{inv}{-}\chi^2(\nu_\theta, \phi_\theta),$$

Model specification: nonlinear mixed model

Specify the following hierarchical model

$$
\begin{aligned}
y_i &= \mathbf{u}_i^T \boldsymbol{\theta} + \varphi_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right), \quad \varphi_i \perp\!\!\!\perp \varepsilon_i, \\
\boldsymbol{\theta} &\sim \mathrm{MVN}(\mathbf{0}_q, \sigma_\theta^2 \boldsymbol{\Lambda}), \\
\sigma_\theta^2 &\sim \mathrm{Scale-inv}{-}\chi^2(\nu_\theta, \phi_\theta), \\
\sigma_\varepsilon^2 &\sim \mathrm{Scale-inv}{-}\chi^2(\nu_\varepsilon, \phi_\varepsilon), \\
\boldsymbol{\varphi} &\sim \mathrm{MVN}(\mathbf{0}_n, \boldsymbol{\Delta}).
\end{aligned}
$$

Model specification: nonlinear mixed model

Specify the following hierarchical model

$$
\begin{aligned}
y_i &= \mathbf{u}_i^T \boldsymbol{\theta} + \varphi_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right), \quad \varphi_i \perp\!\!\!\perp \varepsilon_i, \\
\boldsymbol{\theta} &\sim \mathrm{MVN}(\mathbf{0}_q, \sigma_\theta^2 \boldsymbol{\Lambda}), \\
\sigma_\theta^2 &\sim \mathrm{Scale} - \mathrm{inv}{-}\chi^2(\nu_\theta, \phi_\theta), \\
\sigma_\varepsilon^2 &\sim \mathrm{Scale} - \mathrm{inv}{-}\chi^2(\nu_\varepsilon, \phi_\varepsilon), \\
\boldsymbol{\varphi} &\sim \mathrm{MVN}(\mathbf{0}_n, \boldsymbol{\Delta}).
\end{aligned}
$$

use $\boldsymbol{\Delta}$ to model batch effects or relatedness (kinship matrix).

Posterior sampling

Use Gibb's sample to obtain posterior

(1) $\boldsymbol{\theta} \mid \sigma^2_\theta, \sigma^2_\varepsilon, \mathbf{D} \sim \text{MVN}(\mathbf{m}^*_\theta, \mathbf{V}^*_\theta)$ with
$\mathbf{V}^*_\theta = \sigma^2_\varepsilon \sigma^2_\theta (\sigma^2_\varepsilon \, \boldsymbol{\Lambda}^{-1} + \sigma^2_\theta \mathbf{I}_q)^{-1}$ and $\mathbf{m}^*_\theta = \frac{1}{\sigma^2_\varepsilon} \mathbf{V}^*_\theta \mathbf{U}^T \mathbf{Y}$;

Posterior sampling

Use Gibb's sample to obtain posterior

(1) $\boldsymbol{\theta} \mid \sigma^2_\theta, \sigma^2_\varepsilon, \mathbf{D} \sim \text{MVN}(\mathbf{m}^*_\theta, \mathbf{V}^*_\theta)$ with
$\mathbf{V}^*_\theta = \sigma^2_\varepsilon \sigma^2_\theta (\sigma^2_\varepsilon \, \boldsymbol{\Lambda}^{-1} + \sigma^2_\theta \mathbf{I}_q)^{-1}$ and $\mathbf{m}^*_\theta = \frac{1}{\sigma^2_\varepsilon} \mathbf{V}^*_\theta \mathbf{U}^T \mathbf{Y}$;

(2) $\boldsymbol{\beta} = \mathbf{X}^\dagger \mathbf{Z}^T (\boldsymbol{\Lambda} \mathbf{U}^T \mathbf{K}^{-1} \mathbf{Z}^T)^{-1} \boldsymbol{\theta}$;

Posterior sampling

Use Gibb's sample to obtain posterior

(1) $\boldsymbol{\theta} \mid \sigma^2_\theta, \sigma^2_\varepsilon, \mathbf{D} \sim \text{MVN}(\mathbf{m}^*_\theta, \mathbf{V}^*_\theta)$ with
$\mathbf{V}^*_\theta = \sigma^2_\varepsilon \sigma^2_\theta (\sigma^2_\varepsilon \, \boldsymbol{\Lambda}^{-1} + \sigma^2_\theta \mathbf{I}_q)^{-1}$ and $\mathbf{m}^*_\theta = \frac{1}{\sigma^2_\varepsilon} \mathbf{V}^*_\theta \mathbf{U}^T \mathbf{Y}$;

(2) $\boldsymbol{\beta} = \mathbf{X}^\dagger \mathbf{Z}^T (\boldsymbol{\Lambda} \mathbf{U}^T \mathbf{K}^{-1} \mathbf{Z}^T)^{-1} \boldsymbol{\theta}$;

(3) $\sigma^2_\theta \mid \boldsymbol{\theta}, \sigma^2_\varepsilon, \mathbf{D} \sim \text{Scale-inv} - \chi^2(\nu^*_\theta, \phi^*_\theta)$ where $\nu^*_\theta = \nu_\theta + q$ and
$\phi^*_\theta = \frac{1}{\nu^*_\theta} (\nu_\theta \phi_\theta + \boldsymbol{\theta}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\theta})$;

Posterior sampling

Use Gibb's sample to obtain posterior

(1) $\boldsymbol{\theta} \mid \sigma^2_\theta, \sigma^2_\varepsilon, \mathbf{D} \sim \text{MVN}(\mathbf{m}^*_\theta, \mathbf{V}^*_\theta)$ with
$\mathbf{V}^*_\theta = \sigma^2_\varepsilon \sigma^2_\theta (\sigma^2_\varepsilon \boldsymbol{\Lambda}^{-1} + \sigma^2_\theta \mathbf{I}_q)^{-1}$ and $\mathbf{m}^*_\theta = \frac{1}{\sigma^2_\varepsilon} \mathbf{V}^*_\theta \mathbf{U}^T \mathbf{Y}$;

(2) $\boldsymbol{\beta} = \mathbf{X}^\dagger \mathbf{Z}^T (\boldsymbol{\Lambda} \mathbf{U}^T \mathbf{K}^{-1} \mathbf{Z}^T)^{-1} \boldsymbol{\theta}$;

(3) $\sigma^2_\theta \mid \boldsymbol{\theta}, \sigma^2_\varepsilon, \mathbf{D} \sim \text{Scale-inv} - \chi^2(\nu^*_\theta, \phi^*_\theta)$ where $\nu^*_\theta = \nu_\theta + q$ and
$\phi^*_\theta = \frac{1}{\nu^*_\theta}(\nu_\theta \phi_\theta + \boldsymbol{\theta}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\theta})$;

(4) $\sigma^2_\varepsilon \mid \boldsymbol{\theta}, \sigma^2_\theta, \mathbf{D} \sim \text{Scale-inv} - \chi^2(\nu^*_\varepsilon, \phi^*_\varepsilon)$ where $\nu^*_\varepsilon = \nu_\varepsilon + n$ and
$\phi^*_\varepsilon = \frac{1}{\nu^*_\varepsilon}(\nu_\varepsilon \phi_\varepsilon + \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon})$, with $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{U}\boldsymbol{\theta}$.

Obtain MCMC samples $\left\{\boldsymbol{\beta}^{(t)}\right\}_{t=1}^{T}$.

Posterior probability of association analog (PPAA)

- Hard-thresholding, while controlling for specific error rates.
- **Posterior Probability of Association Analogue (PPAA):**

$$\Pr[\hat{\gamma}_j = 1 \mid \mathbf{y}] \Longleftrightarrow \Pr[|\hat{\beta}_j| > z_{j^*} \mid \mathbf{y}]$$

Capturing nonlinear interactions

A warm up example

$$
\begin{aligned}
K(\mathbf{u},\mathbf{v}) &= (\mathbf{u}^T\mathbf{v}+c)^2 \\
&= \sum_{k=1}^{p}\left(\mathbf{u}_k^2\right)\left(\mathbf{v}_k^2\right) + \sum_{k=2}^{p}\sum_{j=1}^{k-1}\left(\sqrt{2}\mathbf{u}_k\mathbf{u}_j\right)\left(\sqrt{2}\mathbf{v}_k\mathbf{v}_j\right) \\
&\quad + \sum_{k=1}^{p}\left(\sqrt{2c}\mathbf{u}_k\right)\left(\sqrt{2c}\mathbf{v}_k\right) + c^2
\end{aligned}
$$

Capturing nonlinear interactions

The Gaussian kernel

$$k(\mathbf{u}, \mathbf{v}) = \exp\{-h\|\mathbf{u}\|^2\} \exp\{-h\|\mathbf{v}\|^2\} \exp\{-h\langle \mathbf{u}, \mathbf{v}\rangle\}.$$

Capturing nonlinear interactions

The Gaussian kernel

$$k(\mathbf{u}, \mathbf{v}) = \exp\{-h\|\mathbf{u}\|^2\} \exp\{-h\|\mathbf{v}\|^2\} \exp\{-h\langle \mathbf{u}, \mathbf{v}\rangle\}.$$

Taylor expand the third term

$$\begin{aligned} \exp\{-h\langle \mathbf{u}, \mathbf{v}\rangle\} &= \sum_{m=0}^{\infty} \frac{h^m}{m!} (\mathbf{u}^\top \mathbf{v})^m \\ (\mathbf{u}^\top \mathbf{v})^m &= \sum_{j \in [p]^m} \left(\prod_{i=1}^{m} \mathbf{u}_{j_i} \right) \left(\prod_{i=1}^{m} \mathbf{v}_{j_i} \right). \end{aligned}$$

$[p]^m$ is the set of m coordinate subsets of the p coordinates.

Marginal Epistasis

Marginal epistatic effect — the combined pairwise interaction effects between a given variant and all other variants.

Marginal Epistasis

Marginal epistatic effect — the combined pairwise interaction effects between a given variant and all other variants.

$$\mathbf{y} = \mu + \mathbf{x}_k\beta_k + \sum_{l \neq k} \mathbf{x}_l\beta_l + \sum_{l \neq k} (\mathbf{x}_k \circ \mathbf{x}_l)\alpha_l + \boldsymbol{\varepsilon},$$

$$\mathbf{y} = \mu + \mathbf{x}_k\beta_k + \mathbf{m}_k + \mathbf{g}_k + \boldsymbol{\varepsilon}.$$

$\mathbf{x}_k\beta_k$ effect of k-th variant,

Marginal Epistasis

Marginal epistatic effect — the combined pairwise interaction effects between a given variant and all other variants.

$$\mathbf{y} = \mu + \mathbf{x}_k\beta_k + \sum_{l \neq k} \mathbf{x}_l\beta_l + \sum_{l \neq k} (\mathbf{x}_k \circ \mathbf{x}_l)\alpha_l + \boldsymbol{\varepsilon},$$

$$\mathbf{y} = \mu + \mathbf{x}_k\beta_k + \mathbf{m}_k + \mathbf{g}_k + \boldsymbol{\varepsilon}.$$

$\mathbf{x}_k\beta_k$ effect of k-th variant, $\mathbf{m}_k$ is the combined additive effects of all other variants with the k-th variant and $\mathbf{m}_k \sim \text{MVN}(0, \omega^2\mathbf{K}_k)$ with $\mathbf{K}_k = \mathbf{X}_{-k}\mathbf{X}_{-k}/(p-1)$,

Marginal Epistasis

Marginal epistatic effect — the combined pairwise interaction effects between a given variant and all other variants.

$$\mathbf{y} = \mu + \mathbf{x}_k\beta_k + \sum_{l \neq k} \mathbf{x}_l\beta_l + \sum_{l \neq k} (\mathbf{x}_k \circ \mathbf{x}_l)\alpha_l + \boldsymbol{\varepsilon},$$

$$\mathbf{y} = \mu + \mathbf{x}_k\beta_k + \mathbf{m}_k + \mathbf{g}_k + \boldsymbol{\varepsilon}.$$

$\mathbf{x}_k\beta_k$ effect of k-th variant, $\mathbf{m}_k$ is the combined additive effects of all other variants with the k-th variant and $\mathbf{m}_k \sim \text{MVN}(0, \omega^2\mathbf{K}_k)$ with $\mathbf{K}_k = \mathbf{X}_{-k}\mathbf{X}_{-k}/(p-1)$, $\mathbf{g}_k$ is the marginal epistatic effect between the k-th variant and all other variants and $\mathbf{g}_k \sim \text{MVN}(0, \sigma^2\mathbf{G}_k)$ with $\mathbf{G}_k = \mathbf{D}_k\mathbf{K}_k\mathbf{D}_k$.

MArginal ePIstasis Test (MAPIT)

Allows for testing of marginal epistatic effects—the combined pairwise interaction effects between a given variant and all other variants.

This results in p tests rather than $\binom{p}{2}$ tests, resulting in savings in computation and statistical power.

Outperforms: in simulations and on GEUVADIS eQTL data
PLINK – all pairs exhaustive search
Two-Step testing procedures

Collinearity and interpretation

From Gelman and Hill 2007

(1) Centering the Data: We center the approximate kernel matrix;

Collinearity and interpretation

From Gelman and Hill 2007

(1) Centering the Data: We center the approximate kernel matrix;

(2) Orthogonalized the Data: The spectral decomposition performs this step on the data;

Collinearity and interpretation

From Gelman and Hill 2007

(1) Centering the Data: We center the approximate kernel matrix;

(2) Orthogonalized the Data: The spectral decomposition performs this step on the data;

(3) Variable Selection: The g-prior specification on the kernel factor coefficients induces variable section on the original covariate effect sizes.

Genomic selection in mice

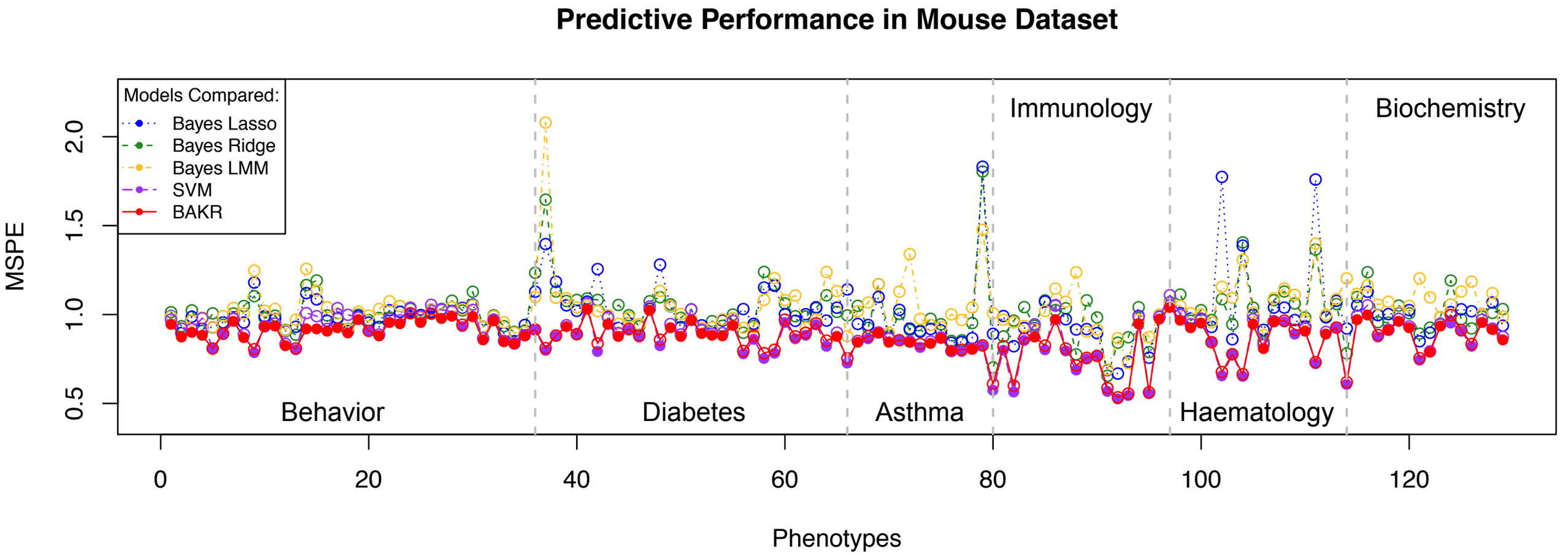
- Wellcome Trust Centre for Human Genetics
- Approximately 2,000 *related* mice and 10,000 SNPs
- Classify 129 traits into 6 broad groups: **behavior, diabetes, asthma, immunology, haematology, and biochemistry**
- **Models Compared:** Bayesian Lasso, Bayesian Ridge, Bayesian Linear Mixed Model (LMM), Support Vector Machine (SVM)
- Perform 80-20 (in/out of sample) inter-family splits; 100 times
- **Accuracy Measure:** Mean Squared Prediction Error (MSPE)

Genomic selection in mice

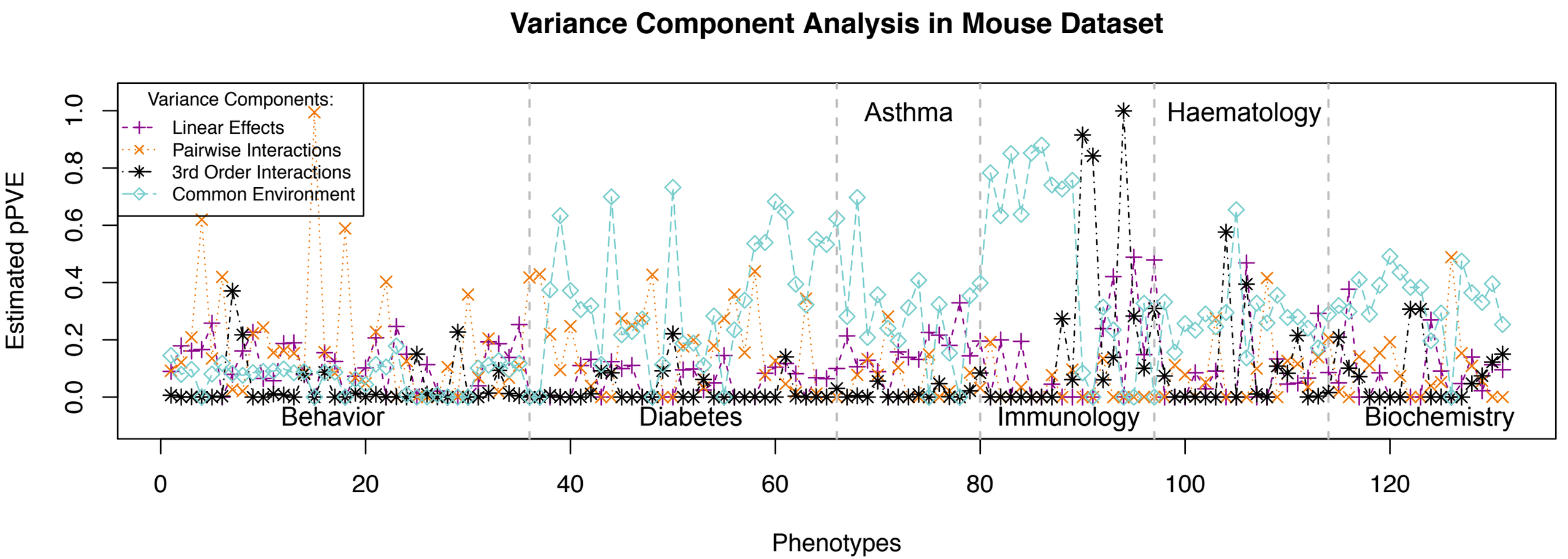
	Bayes Lasso	Bayes Ridge	Bayes LMM	SVM	BAKR
MSPE	1.00 (0.16)	1.01 (0.14)	1.04 (0.15)	0.88 (0.12)	**0.87 (0.10)**
Pr[Optimal]	0.016	0.00	0.00	0.434	**0.550**

Average MSPE across all 129 phenotypes. The number in parenthesis is the standard error due to random sampling

Genomic selection in mice



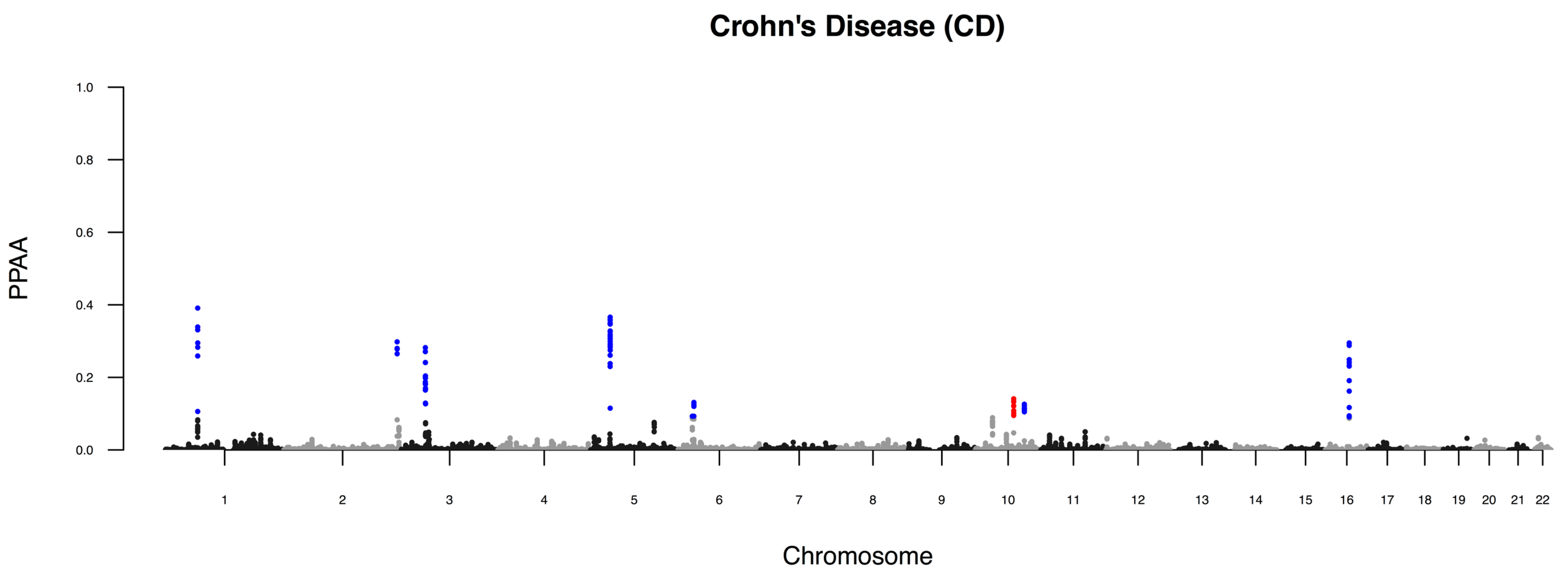
Genomic selection in mice



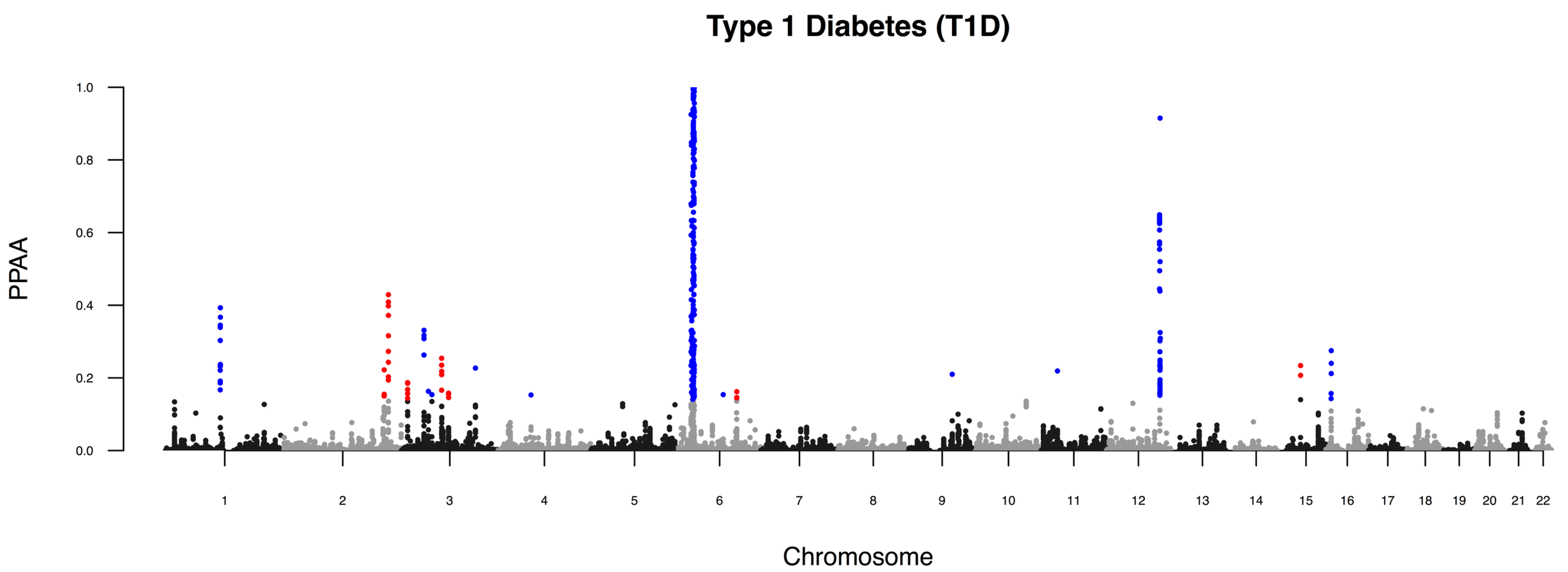
GWAS in WTCCC

- Wellcome Trust Case Control Consortium (WTCCC)
- Includes about 14,000 cases from seven common diseases and about 3,000 shared controls
- Approximately 450,000 SNPs measured for each individual
- **Diseases: Crohn's Disease (CD) and Type 1 Diabetes (T1D)**
- **Significance Measure:** Posterior Probability of Association Analogue (PPAA)

GWAS in WTCCC



GWAS in WTCCC



Future work

1. Using summaries in mapping.

Future work

1. Using summaries in mapping.
2. Relatedness and epistasis.

Future work

1. Using summaries in mapping.
2. Relatedness and epistasis.
3. Local estimates of heritability.

Future work

1. Using summaries in mapping.
2. Relatedness and epistasis.
3. Local estimates of heritability.
4. Bigger, faster, better.

Code

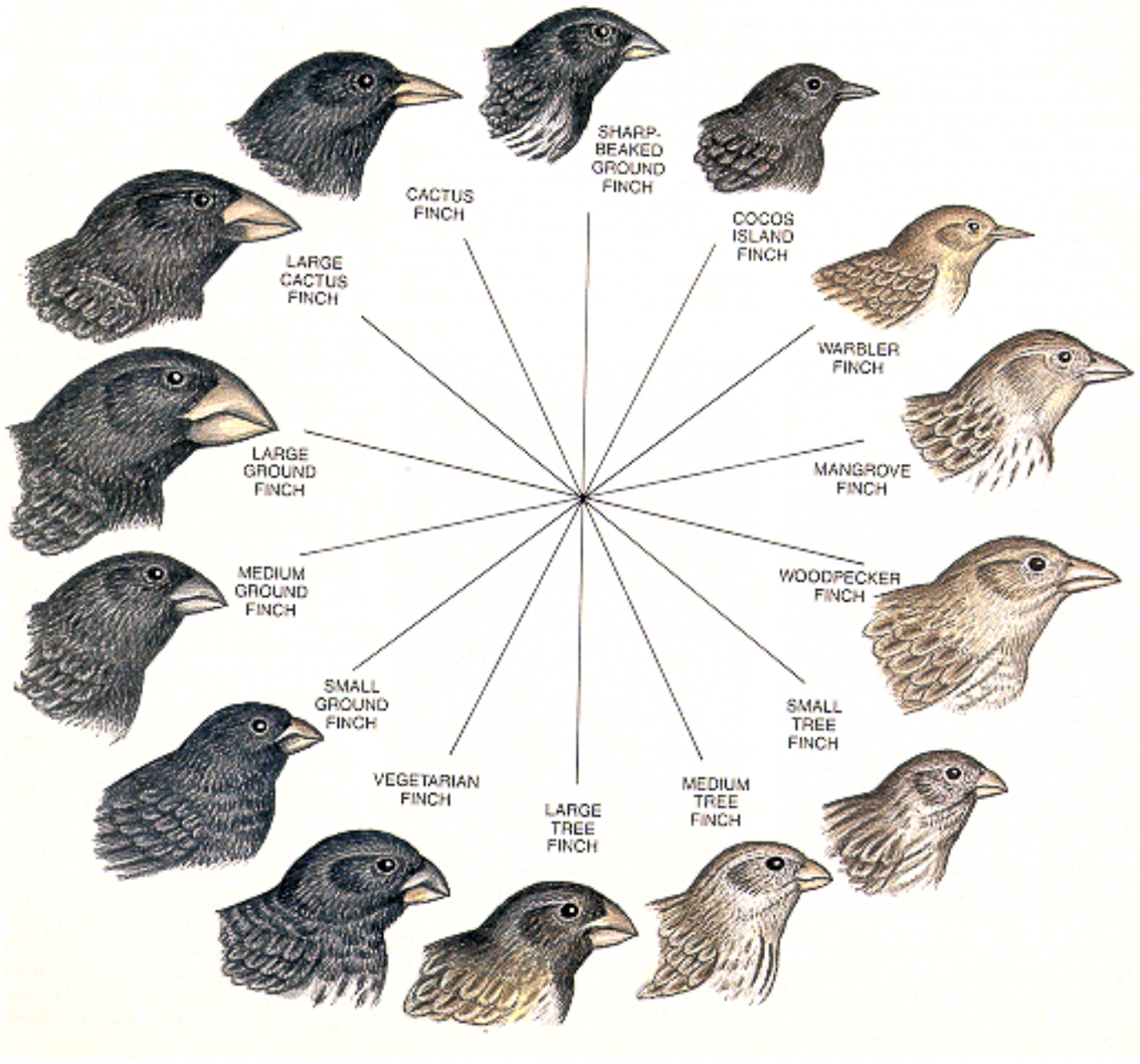
Bayesian Approximate Kernel Regression (BAKR)
https://github.com/lorinanthony/BAKR

MArginal ePIstasis Test (MAPIT)
https://github.com/lorinanthony/MEPIT

Other association mapping software
http://www.xzlab.org/software.html

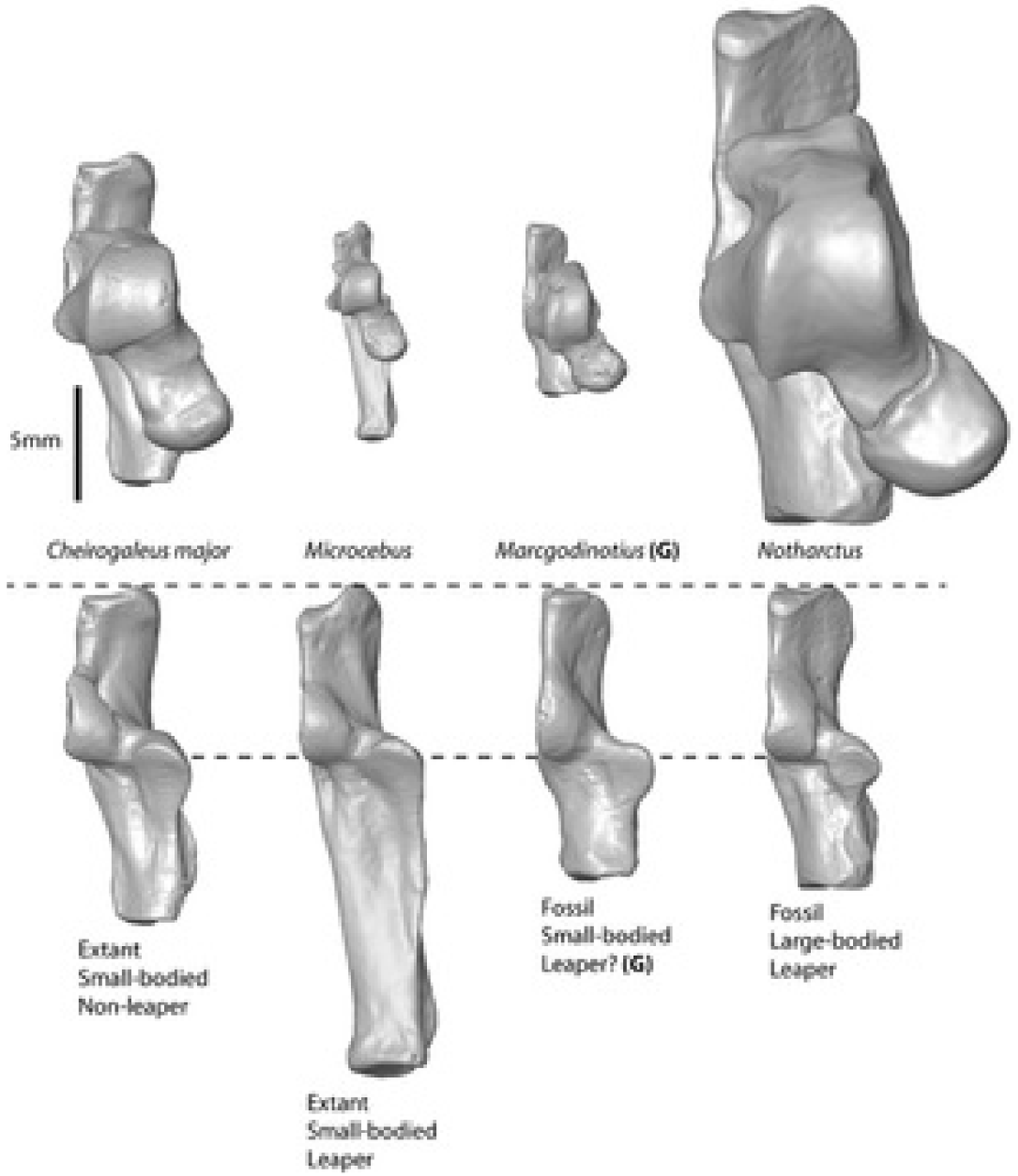
Digitizing shapes

Modeling variation in shapes



S. J. Gould

Variation in calcanei



D. Boyer.

Models of surfaces

(1) Shape spaces: Based on landmarks on shapes, defines shape manifolds, procrustean metrics, and complex projective spaces.

Models of surfaces

(1) Shape spaces: Based on landmarks on shapes, defines shape manifolds, procrustean metrics, and complex projective spaces.

(2) Diffeomorphism based: Variational problems on flows of diffeomorphisms for image matching. Magnitude of a continuous transformation between shapes.

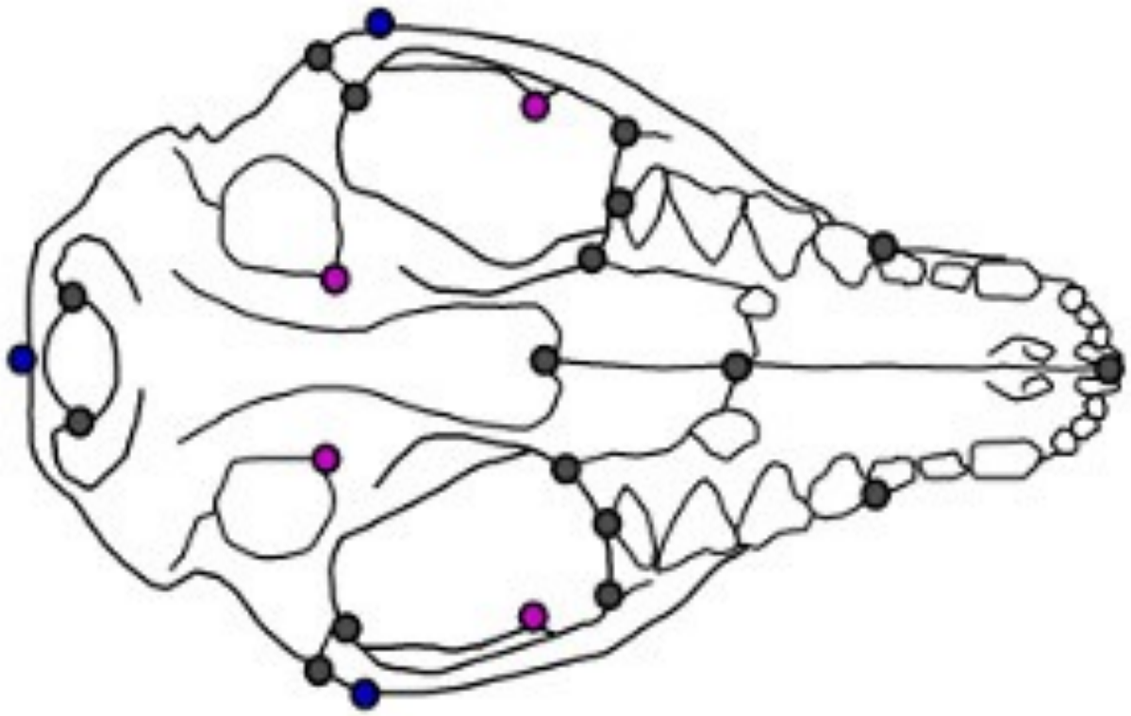
Models of surfaces

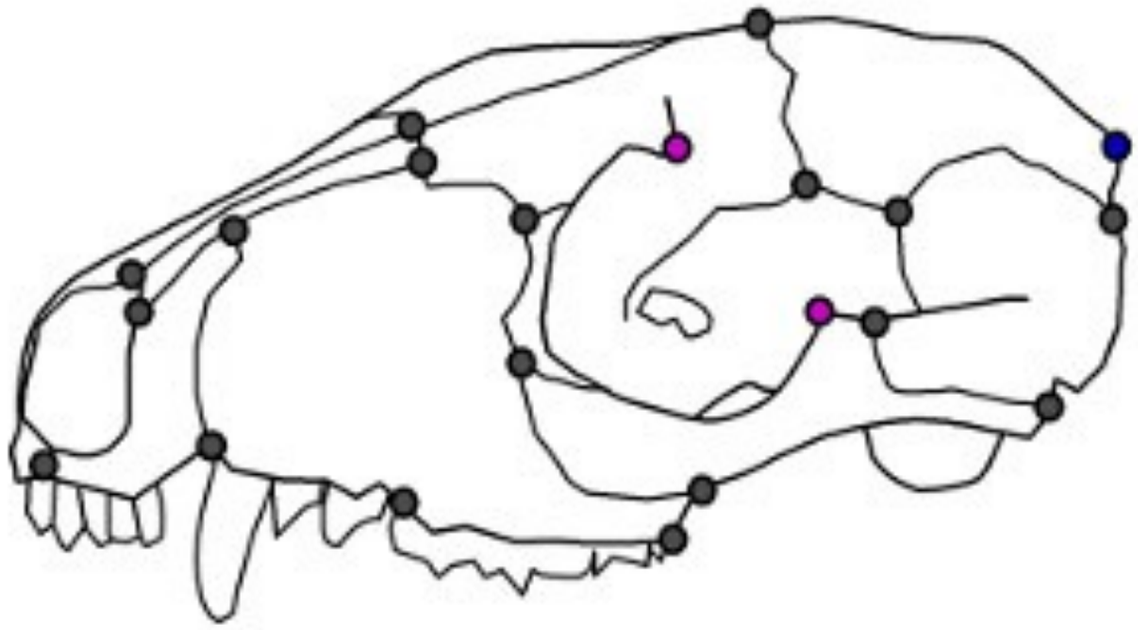
(1) Shape spaces: Based on landmarks on shapes, defines shape manifolds, procrustean metrics, and complex projective spaces.

(2) Diffeomorphism based: Variational problems on flows of diffeomorphisms for image matching. Magnitude of a continuous transformation between shapes.

(3) Integral geometry: Using topological invariants of extrema of random sets or random fields.

Landmarks and shape spaces





3D image repositories

MORPHO SOURCE

ABOUT | BROWSE | DASHBOARD

LOGIN/REGISTER

Getting Started

Find & Download Datasets

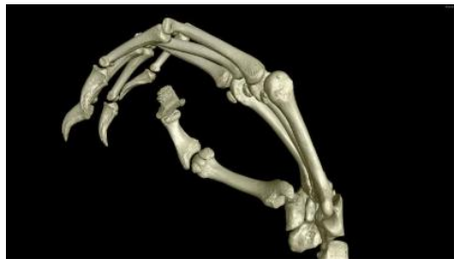
BROWSE or

enter search terms

LOGIN OR REGISTER

Useful Info

- Information for Users
- Information for Contributors
- Terms
- User Guide



foot of Daubentonia madagscariensis *scanned at 38micron resolution at Duke Evolutionary Anthropology department's new high resolution microCt facility. Click here if you are interested in details on the facility*

Recently Published

Four bones of a new species of Homo from South Africa.

→ See all the bones of the newly described Homo naledi

→ Read the published article

PREVIOUS NEXT

Welcome

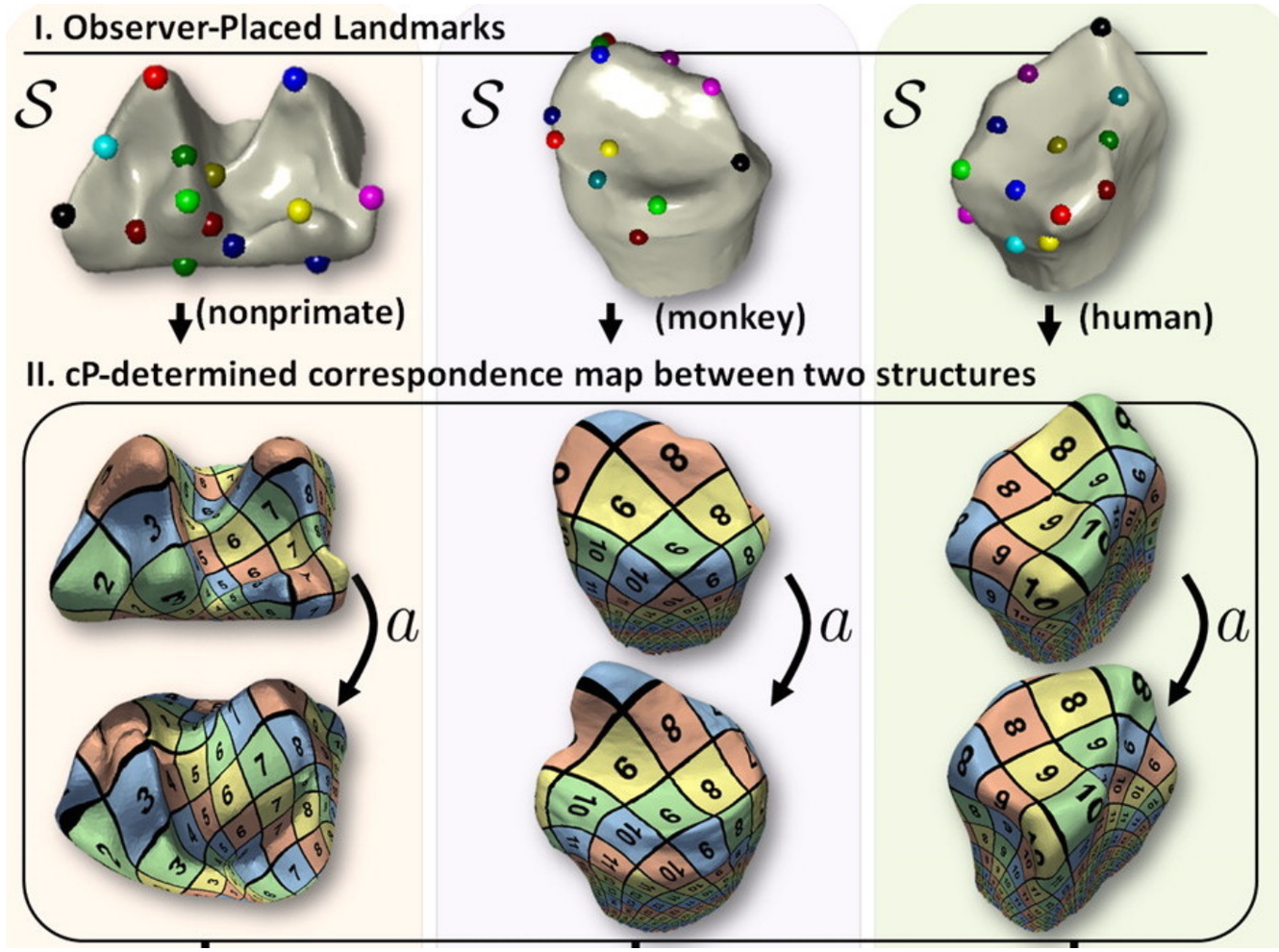
MorphoSource is a project-based data archive that allows researchers to store and organize, share, and distribute their own 3d data. Furthermore any registered user can immediately search for and download 3d morphological data sets that have been made accessible through the consent of data authors.

The goal of **MorphoSource** is to provide rapid access to as many researchers as possible, large numbers of raw microCt data and surface meshes representing vouchered specimens.

File formats include tiff, dicom, stanford ply, and stl. The website is designed to be self explanatory and to assist you through the process of uploading media and associating it with meta data. If you are interested in using the site for your own data but have questions about security or anything else contact the site administrator. Otherwise please download whatever data you need and check back frequently to see what's new.

Diffeomorphism based approach

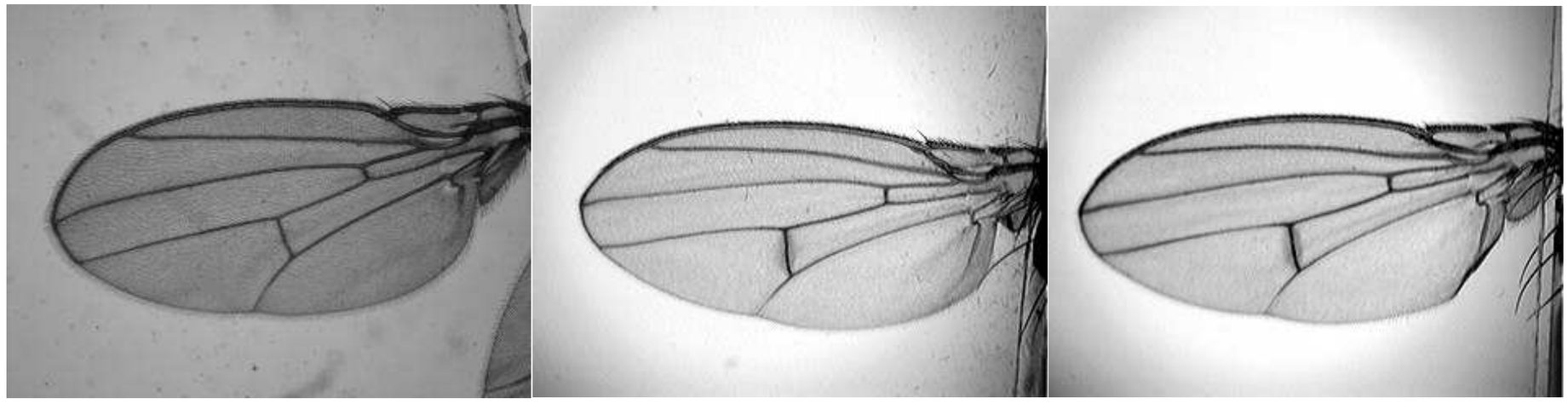
Similarity between teeth: Algorithms to automatically quantify the geometric similarity of anatomical surfaces, Boyer et. al. PNAS 2011.



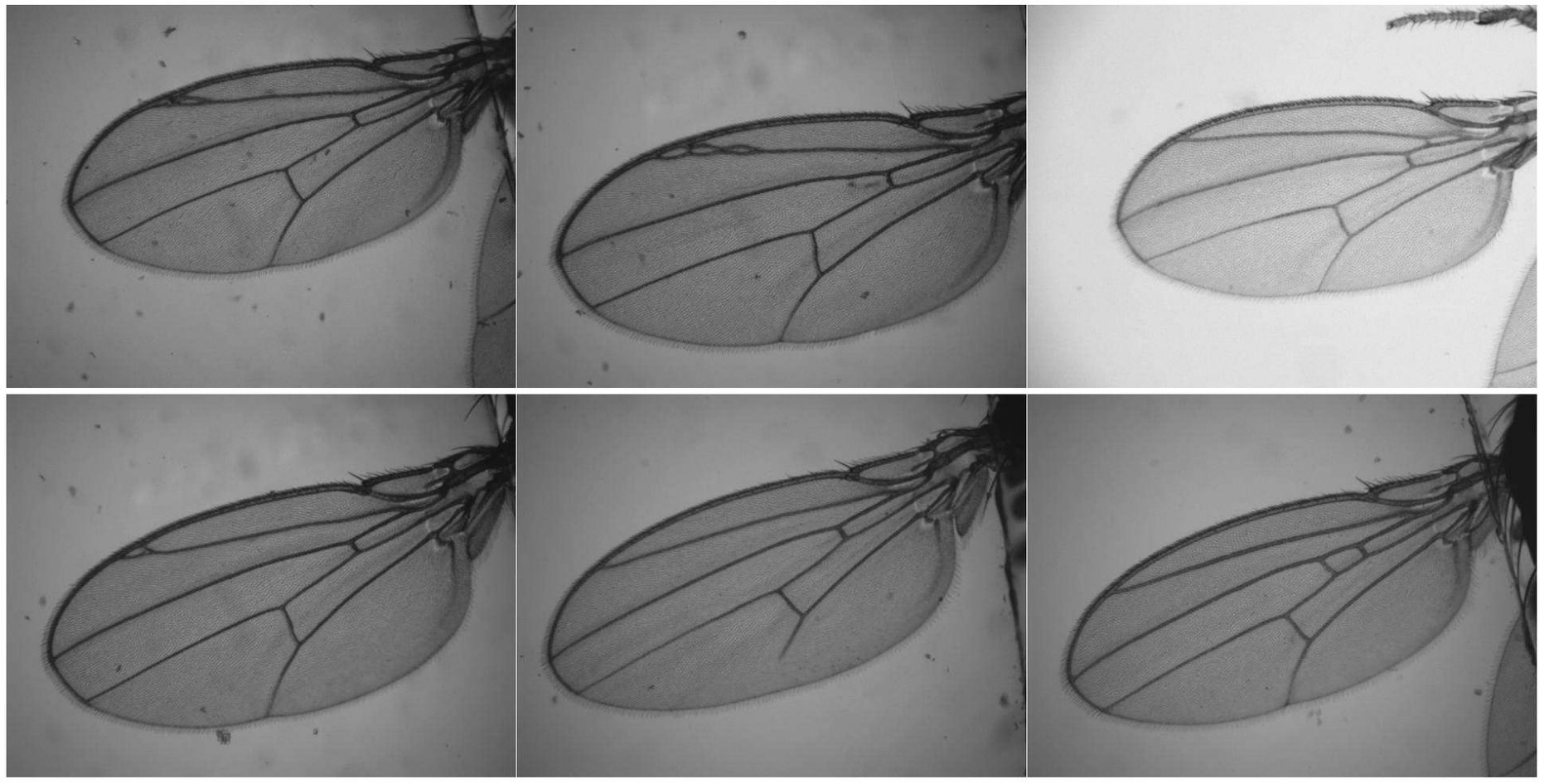
Homeomorphism between shapes

Fly wings are not homeomorphic

Normal fly wings [photos from David Houle's lab]:



Topologically abnormal veins:



Our objective

Model shapes without requiring landmarks or diffeomorphisms.

Our objective

Model shapes without requiring landmarks or diffeomorphisms.

Transform the data/object into a representation that can be modeled using standard methods.

Our objective

Model shapes without requiring landmarks or diffeomorphisms.

Transform the data/object into a representation that can be modeled using standard methods.

Desired properties

(1) The transformation is injective, ideally the summary statistic is sufficient.

Our objective

Model shapes without requiring landmarks or diffeomorphisms.

Transform the data/object into a representation that can be modeled using standard methods.

Desired properties

(1) The transformation is injective, ideally the summary statistic is sufficient.

(2) The transformed space is nice (may be a matrix). Can compute distances or place probability models in the transformed space.

Our objective

Model shapes without requiring landmarks or diffeomorphisms.

Transform the data/object into a representation that can be modeled using standard methods.

Desired properties

(1) The transformation is injective, ideally the summary statistic is sufficient.

(2) The transformed space is nice (may be a matrix). Can compute distances or place probability models in the transformed space.

(3) Pull back from the transformed space to positions on shapes.

Two topological summaries

(1) Euler characteristics.

Two topological summaries

(1) Euler characteristics.

(2) Persistent homology.

Euler characteristic

For a mesh M in 3 dimensions the Euler characteristic is

$$\chi(M) = \#\text{vertices} - \#\text{edges} + \#\text{faces}.$$

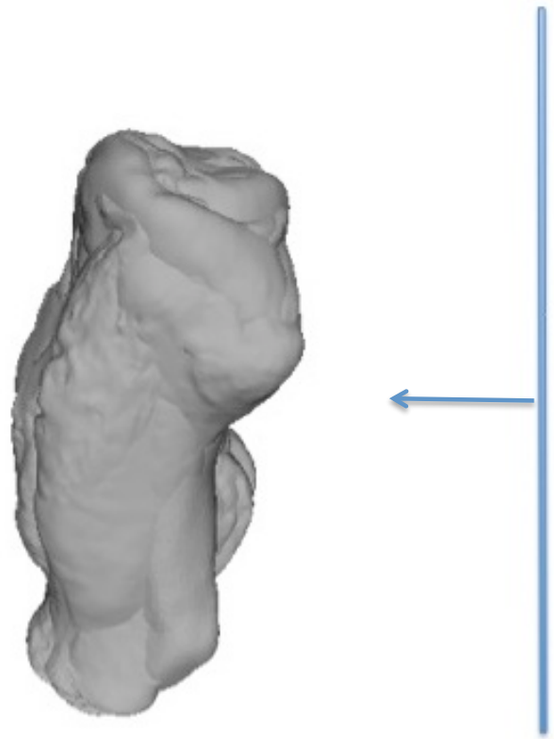
Euler characteristic

For a mesh M in 3 dimensions the Euler characteristic is

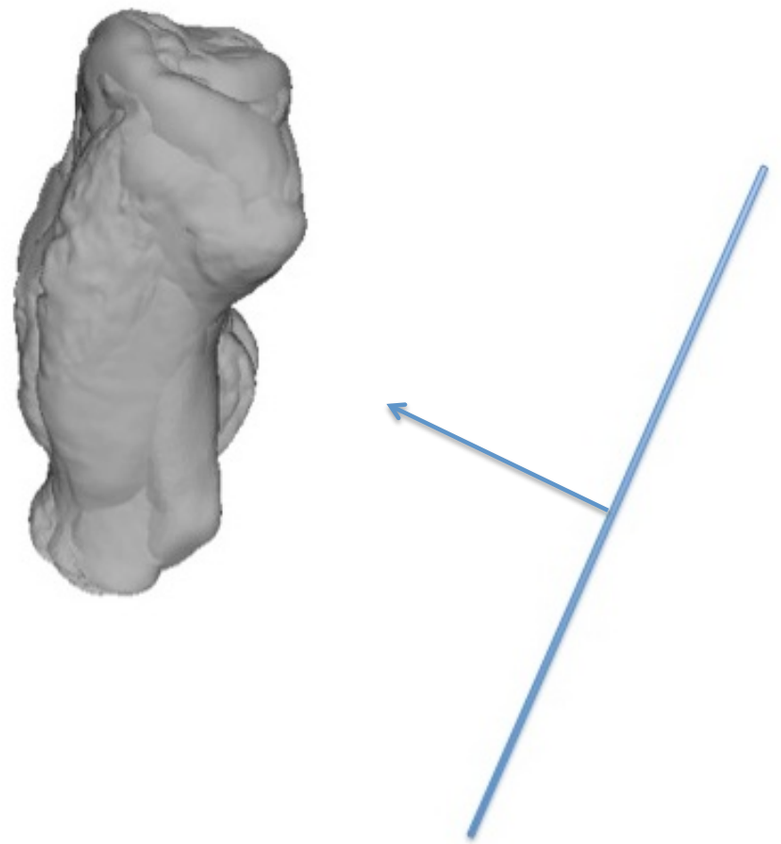
$$\chi(M) = \#\text{vertices} - \#\text{edges} + \#\text{faces}.$$

χ=2 χ=0 χ=−34

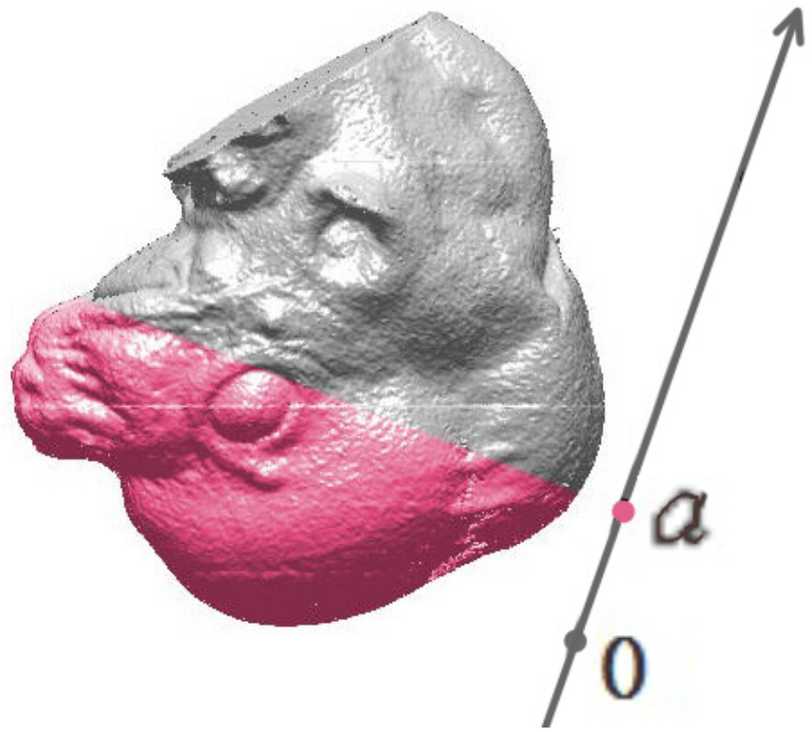
Height function: v_1



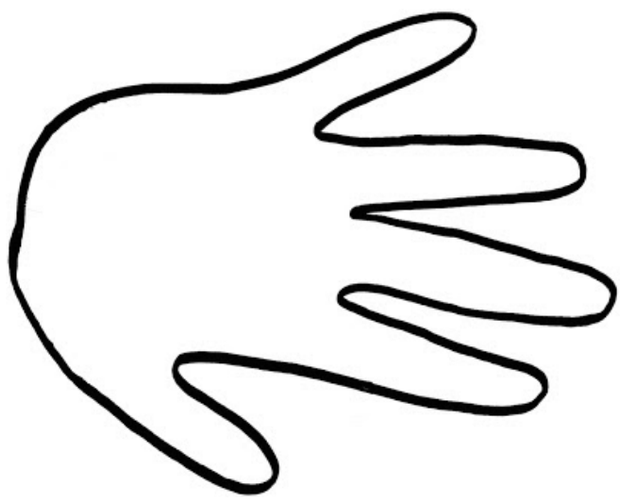
Height function: v_2

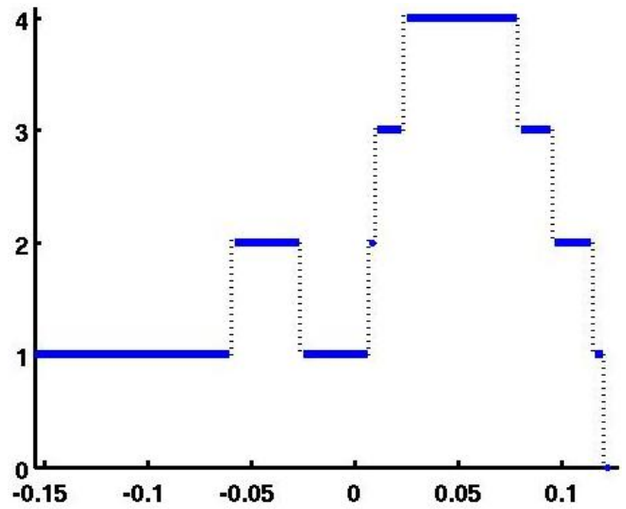


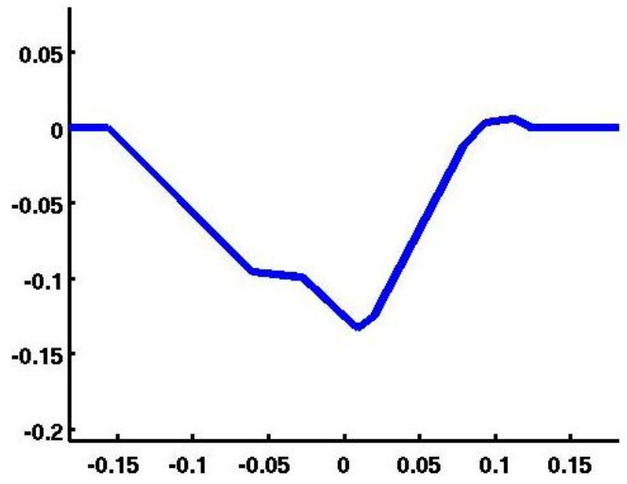
Sublevel set of mouse embryo head



Euler characteristic curves

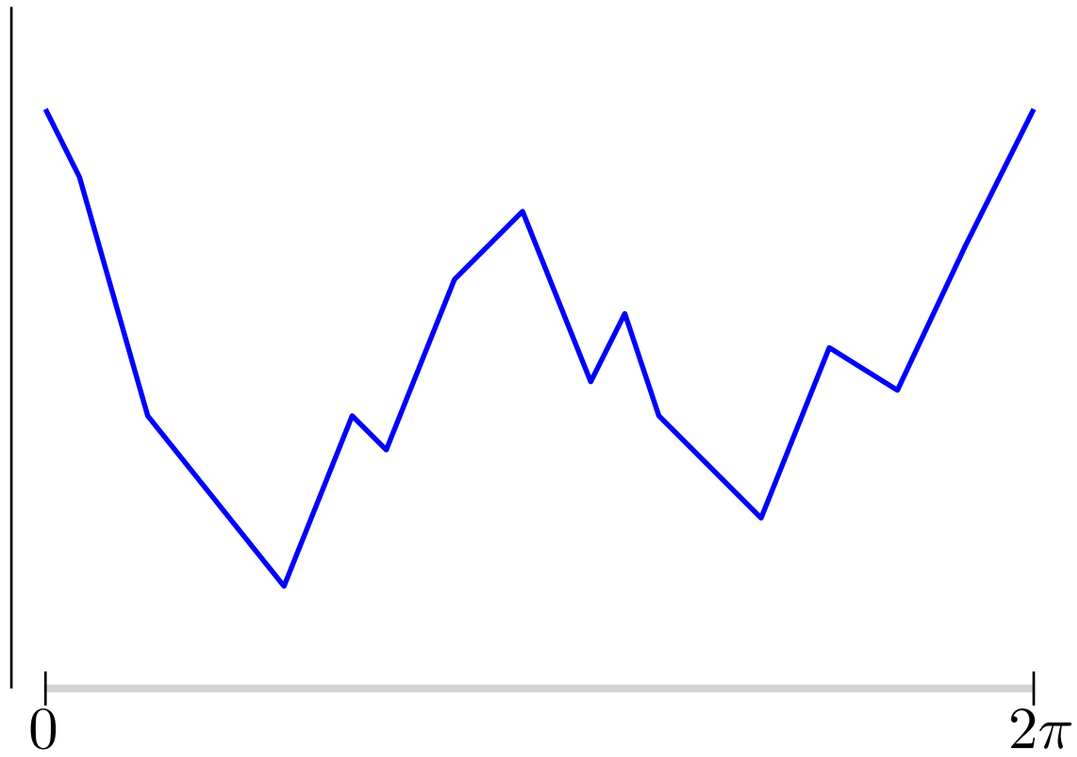


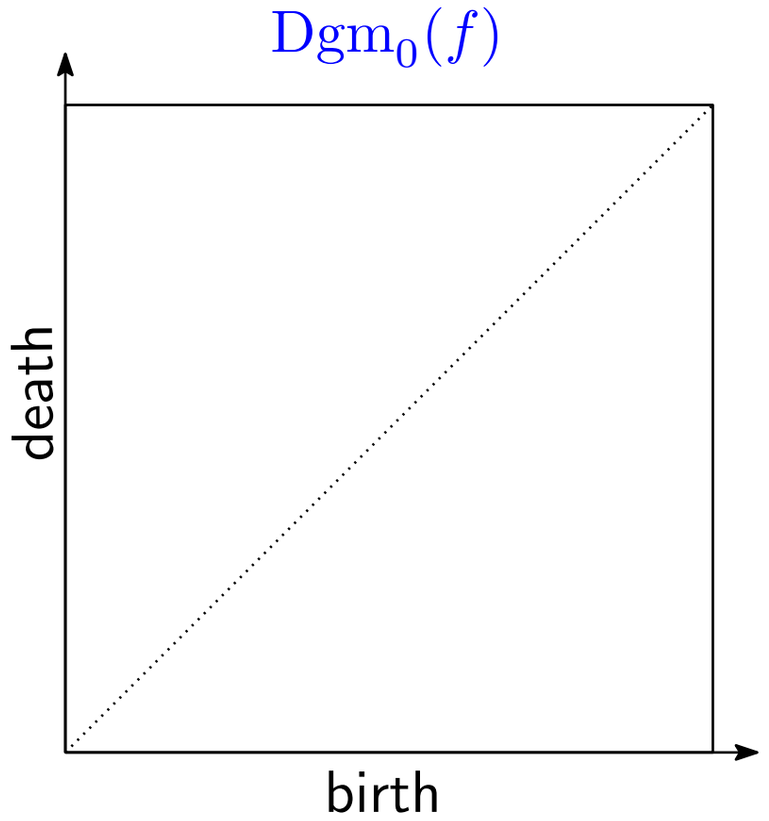




Persistent homology: Morse theory

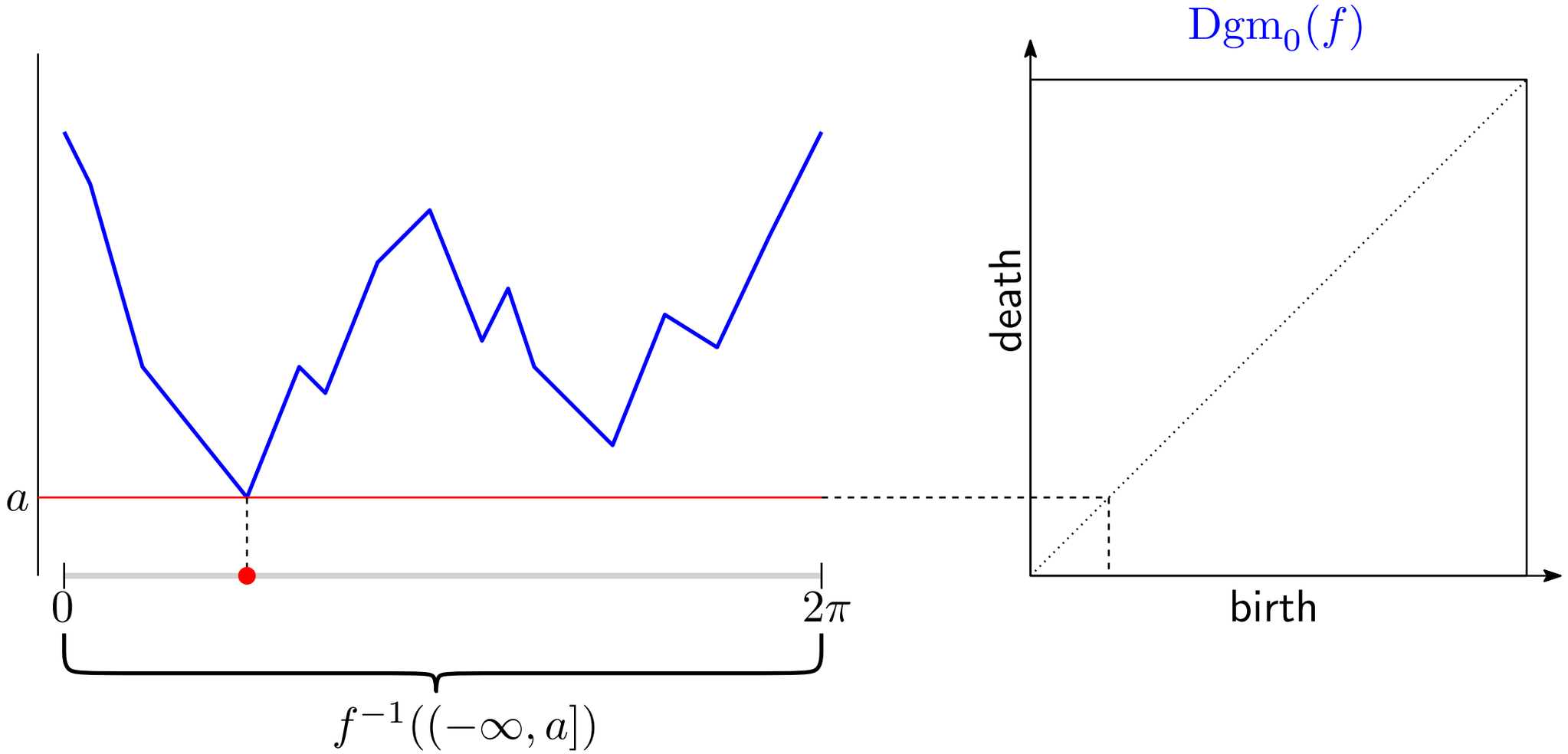
Evolution of homology as birth-death pair.





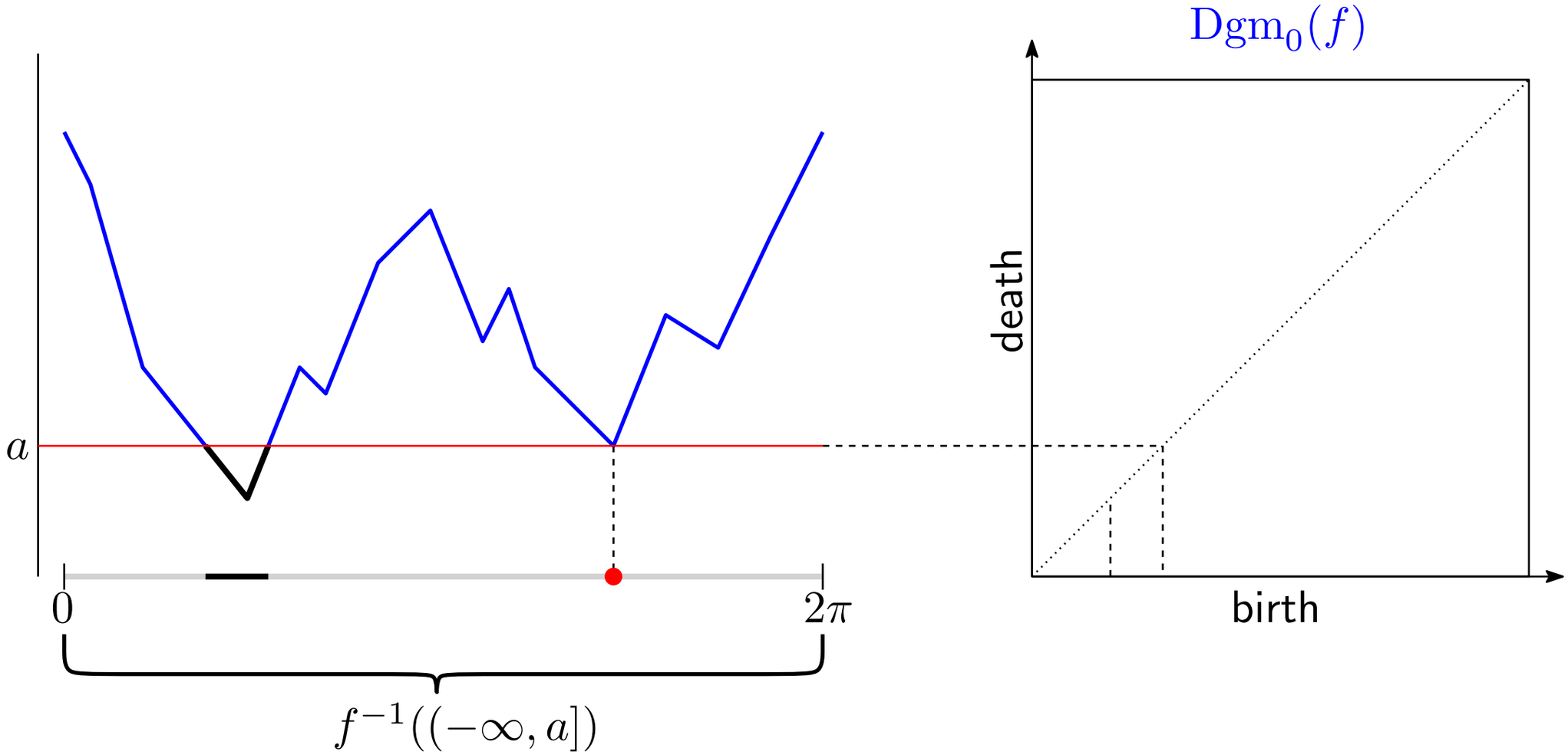
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



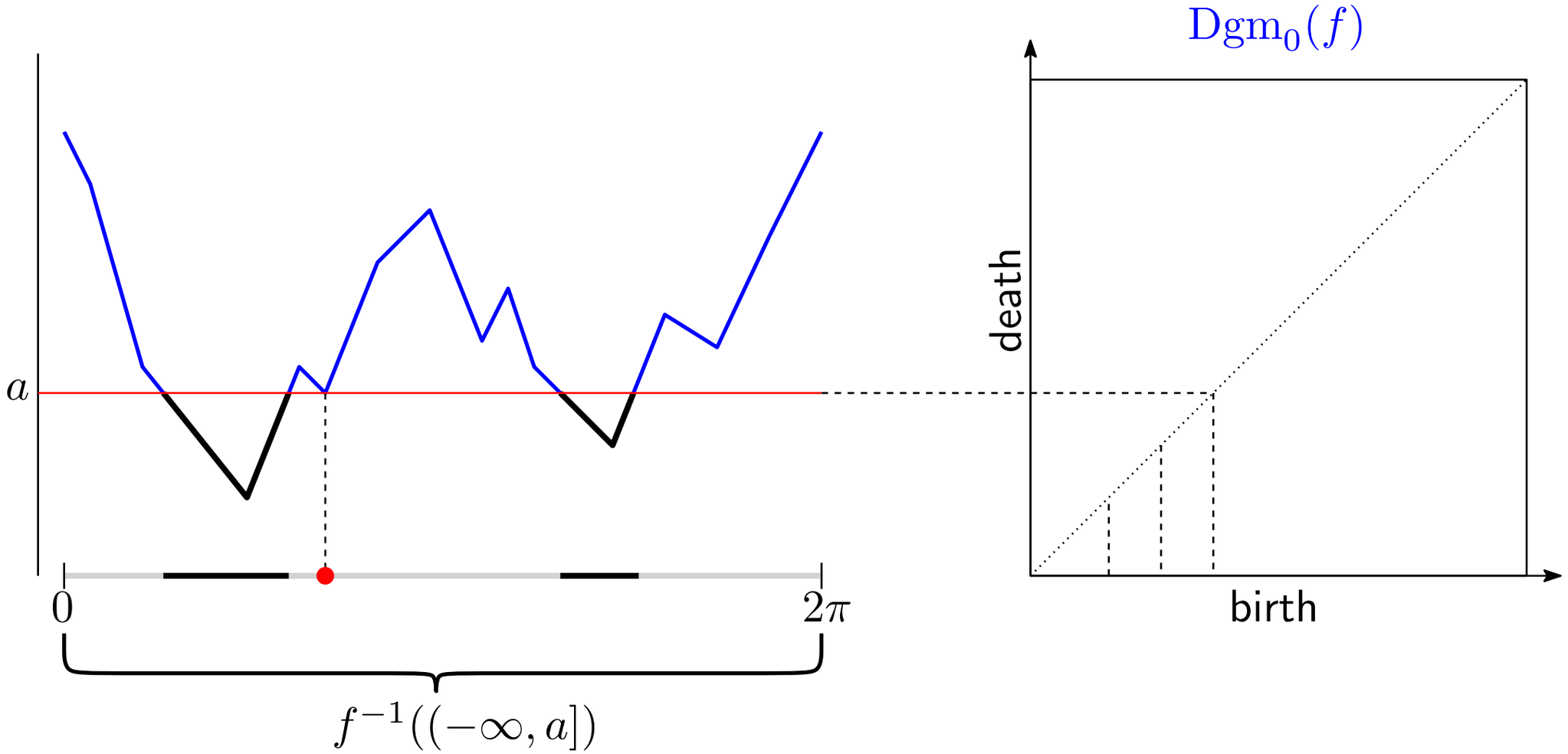
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



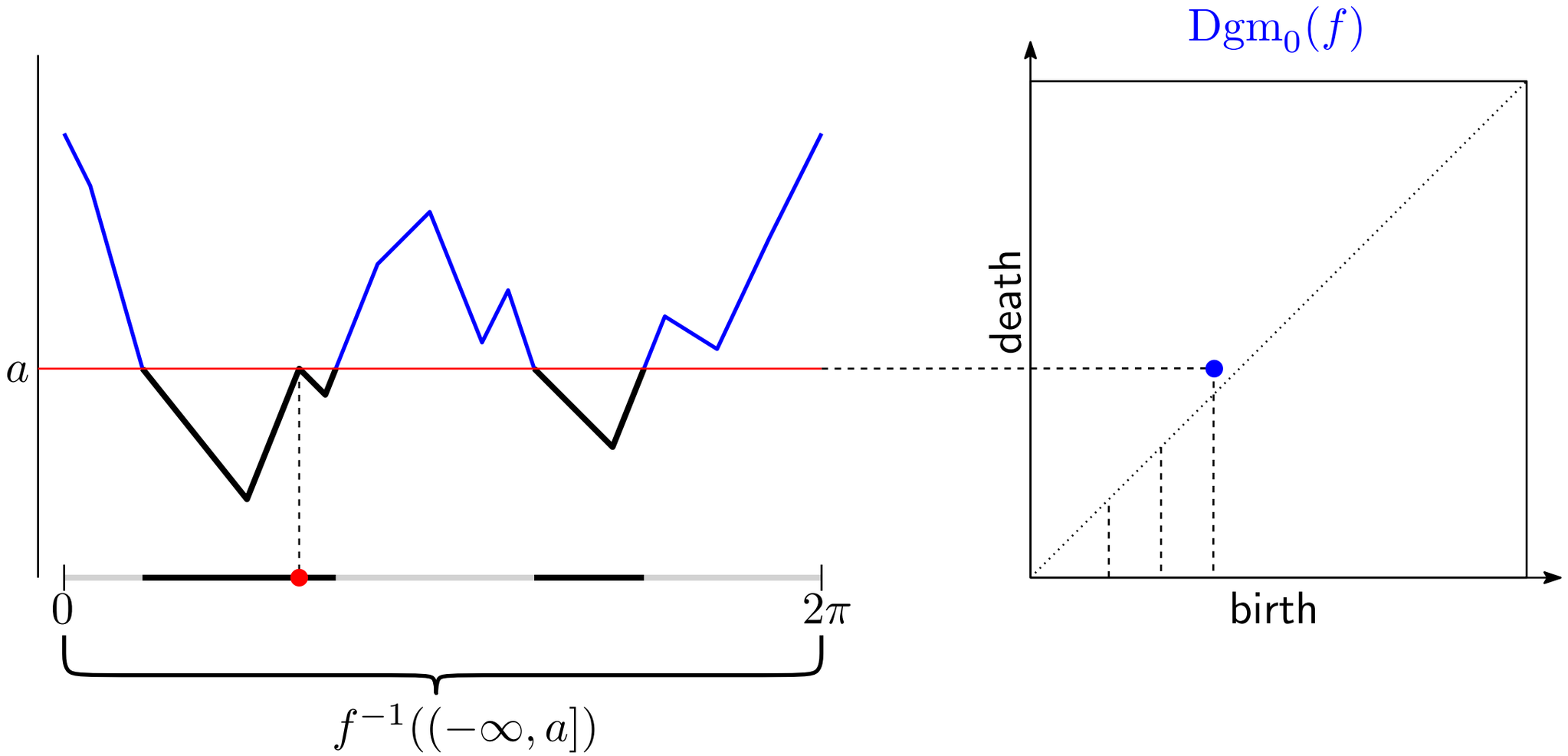
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



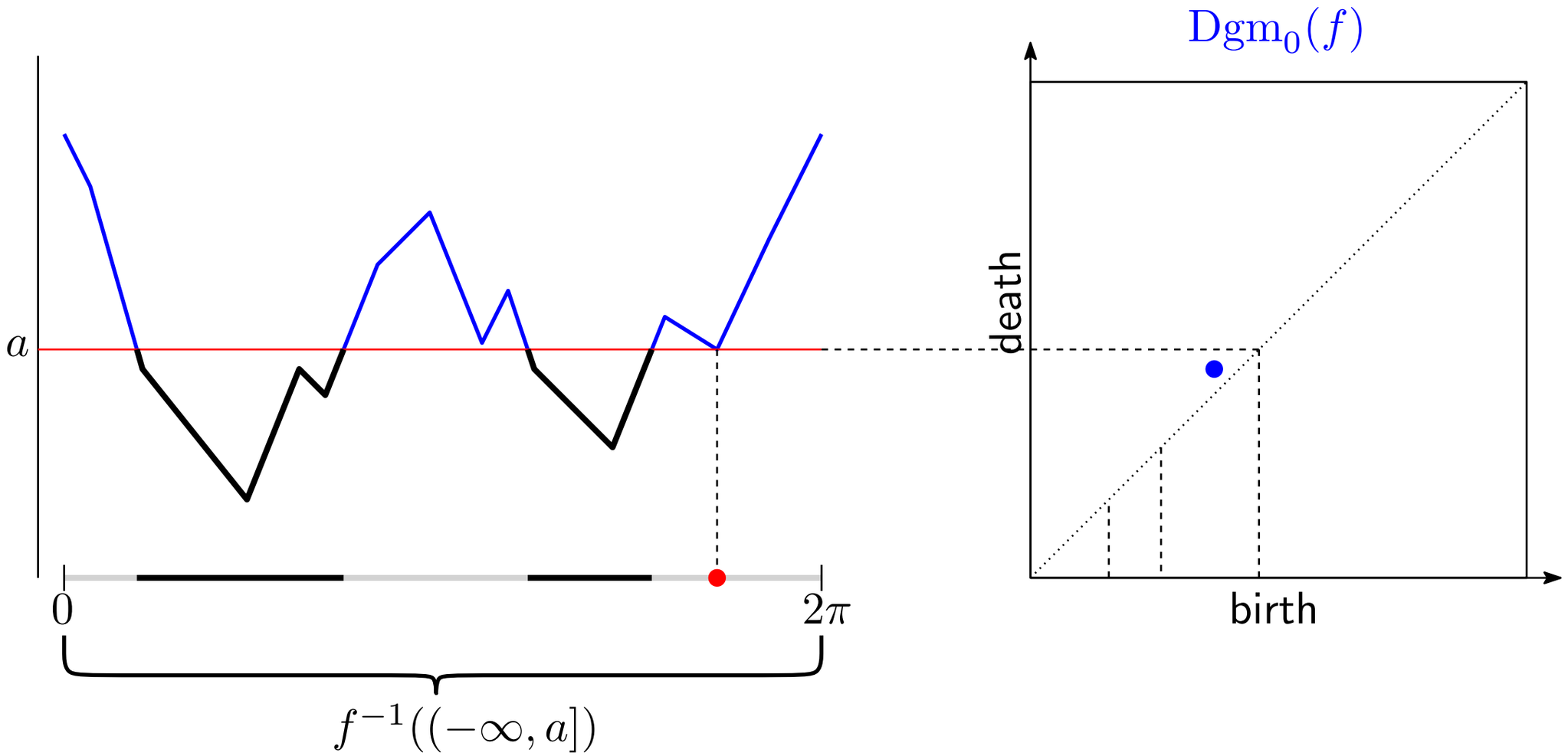
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



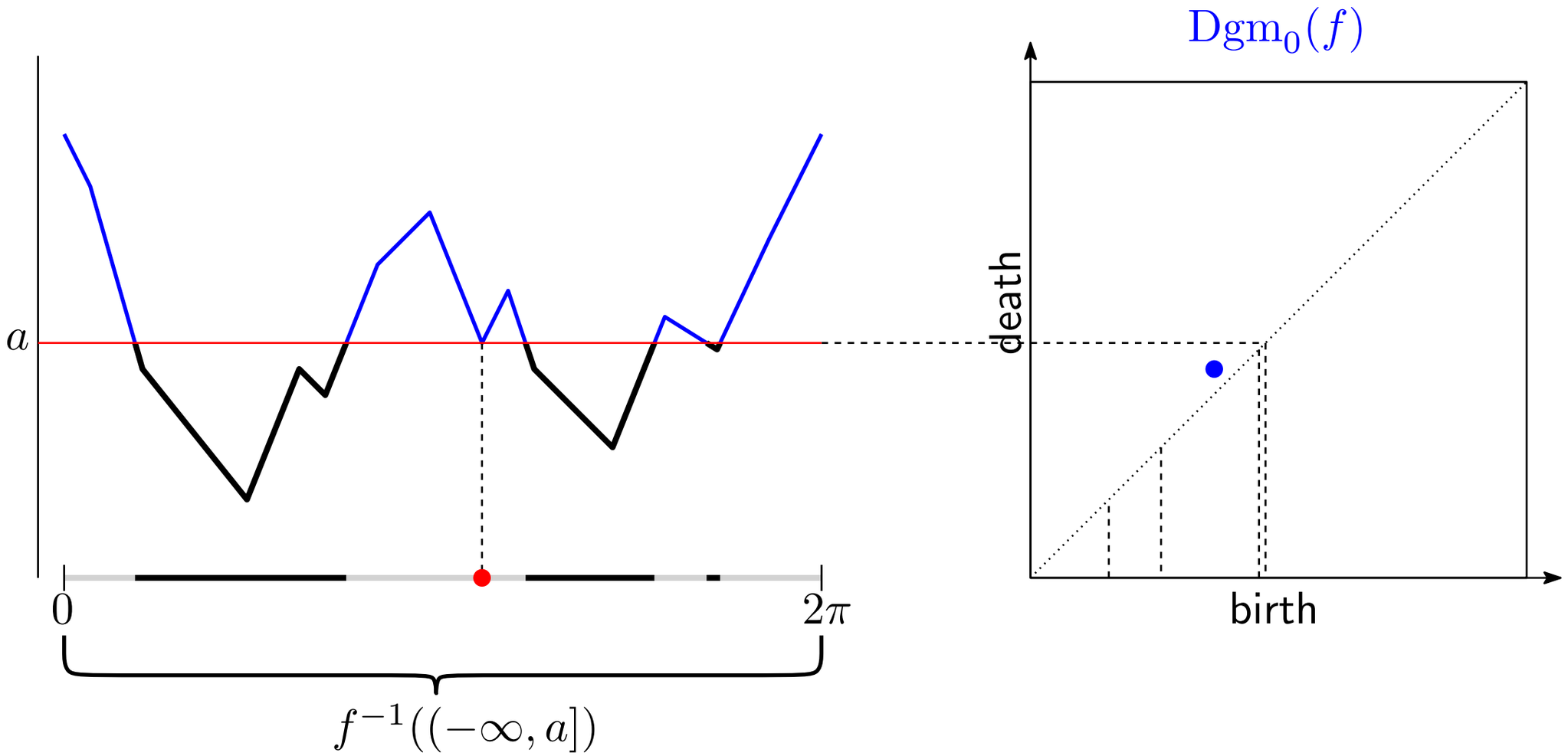
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



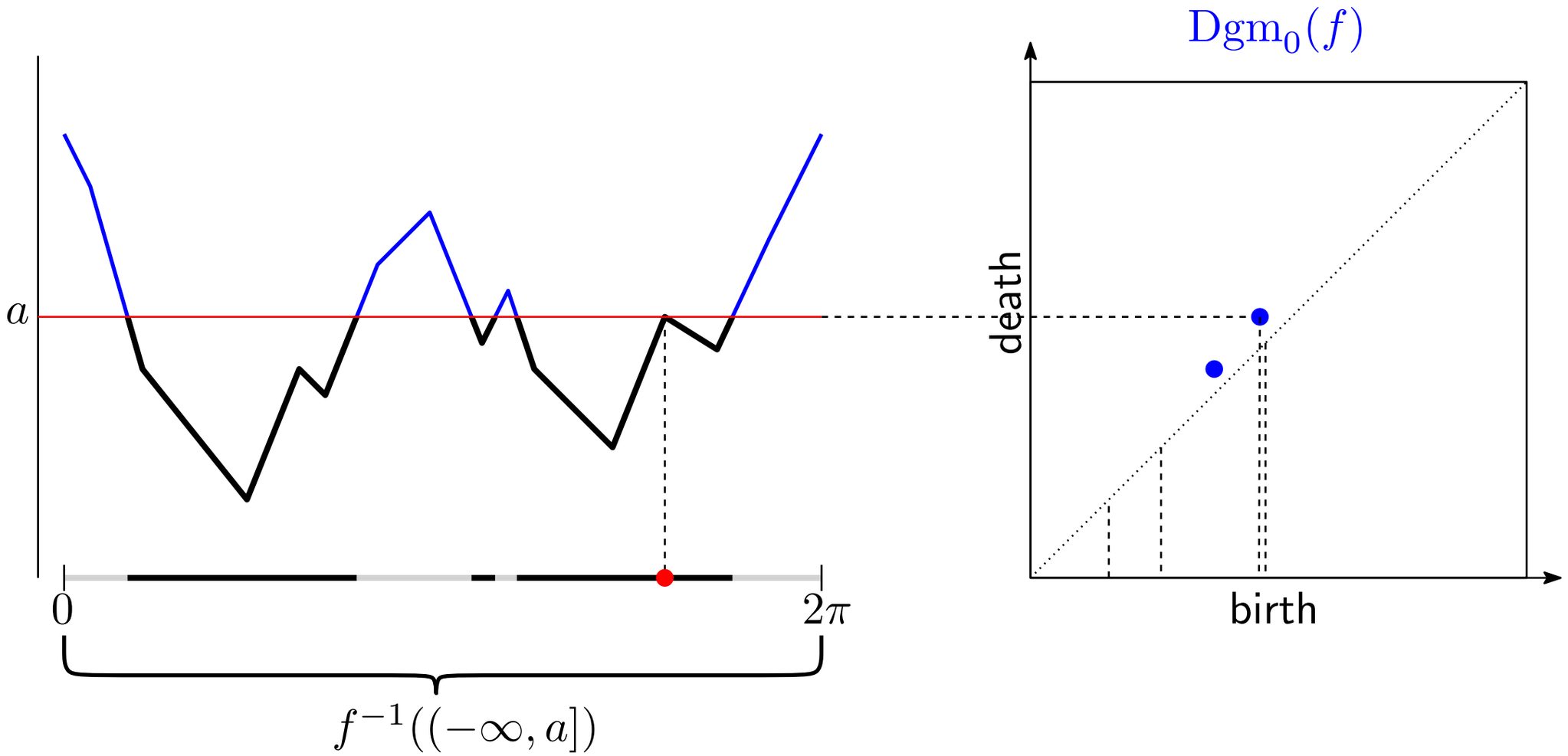
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



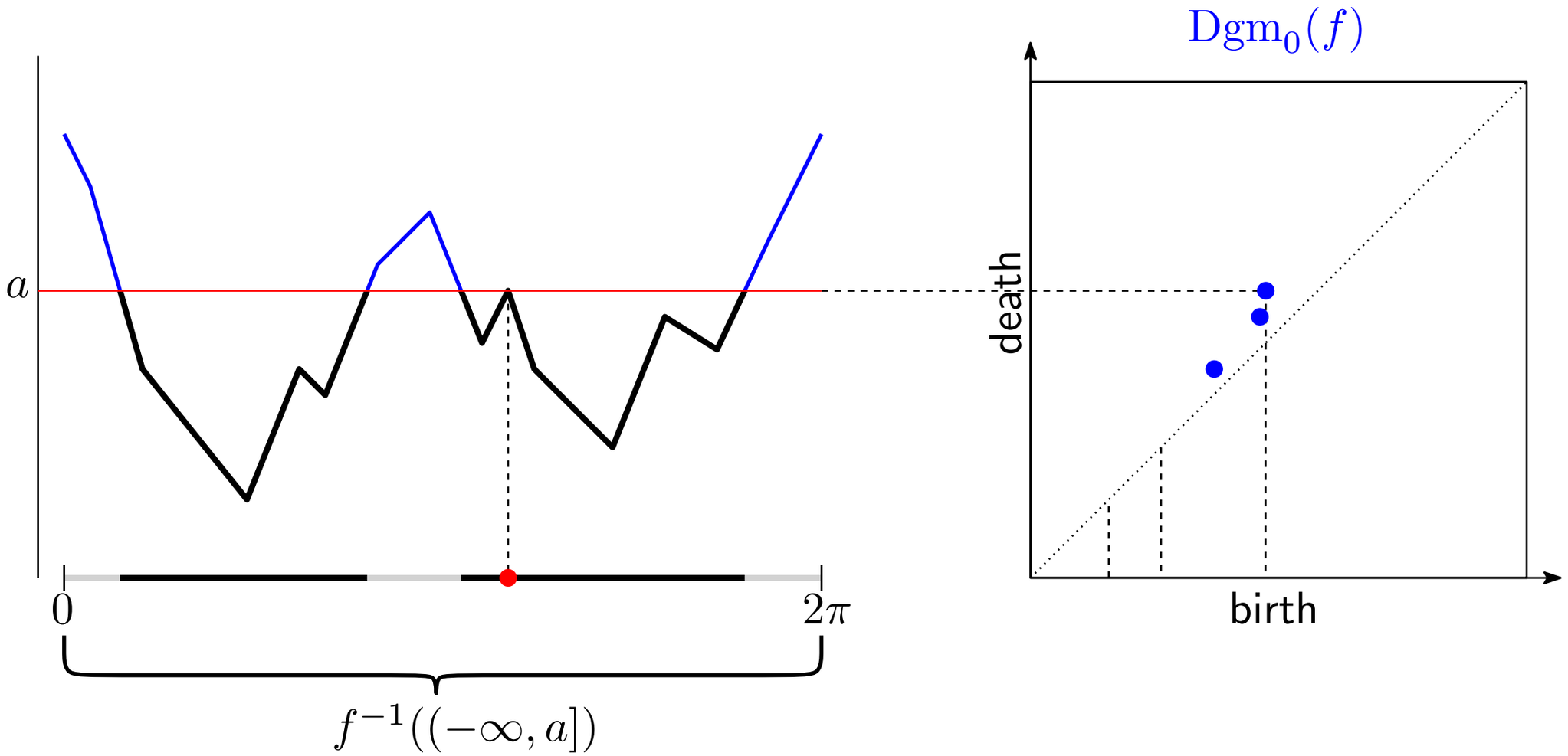
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



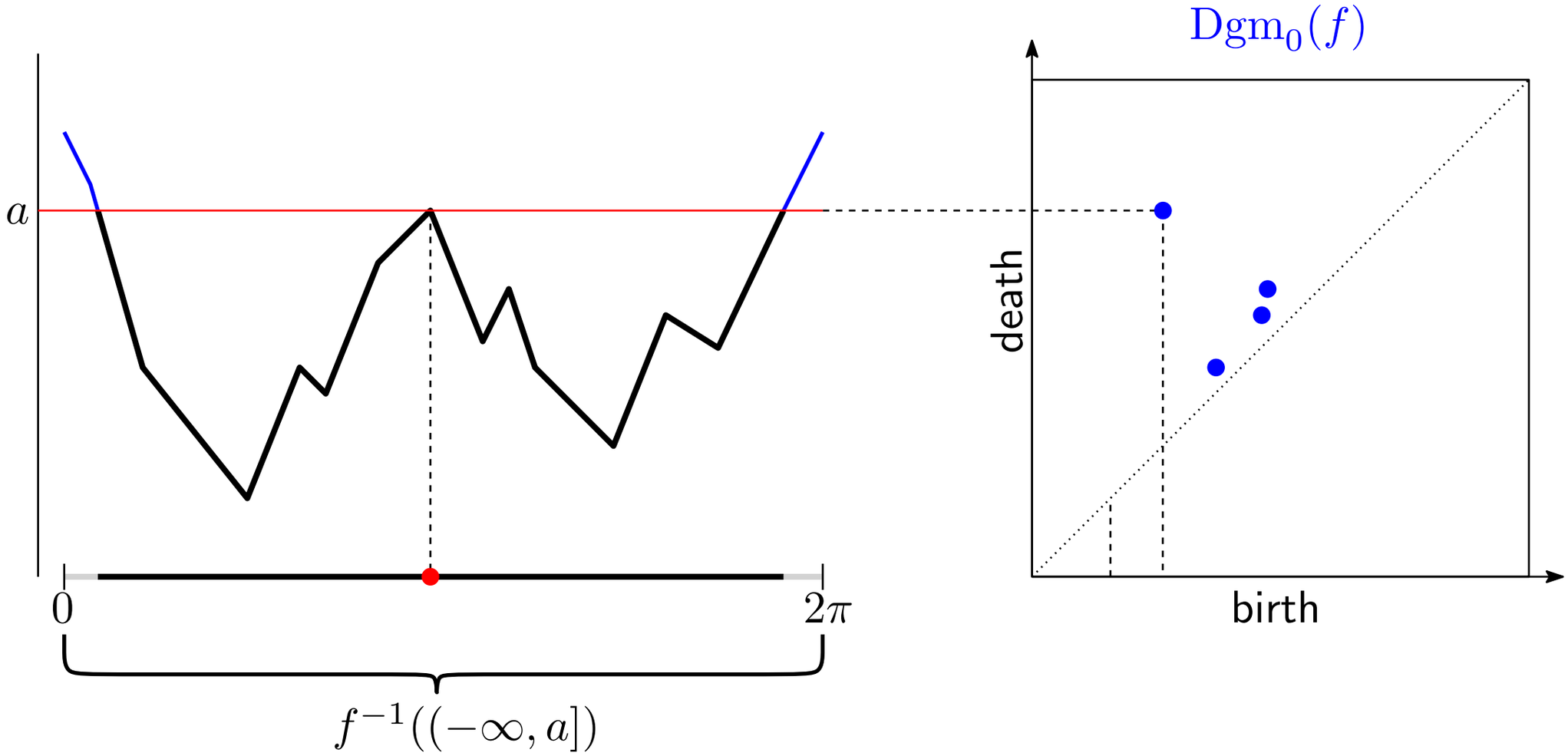
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



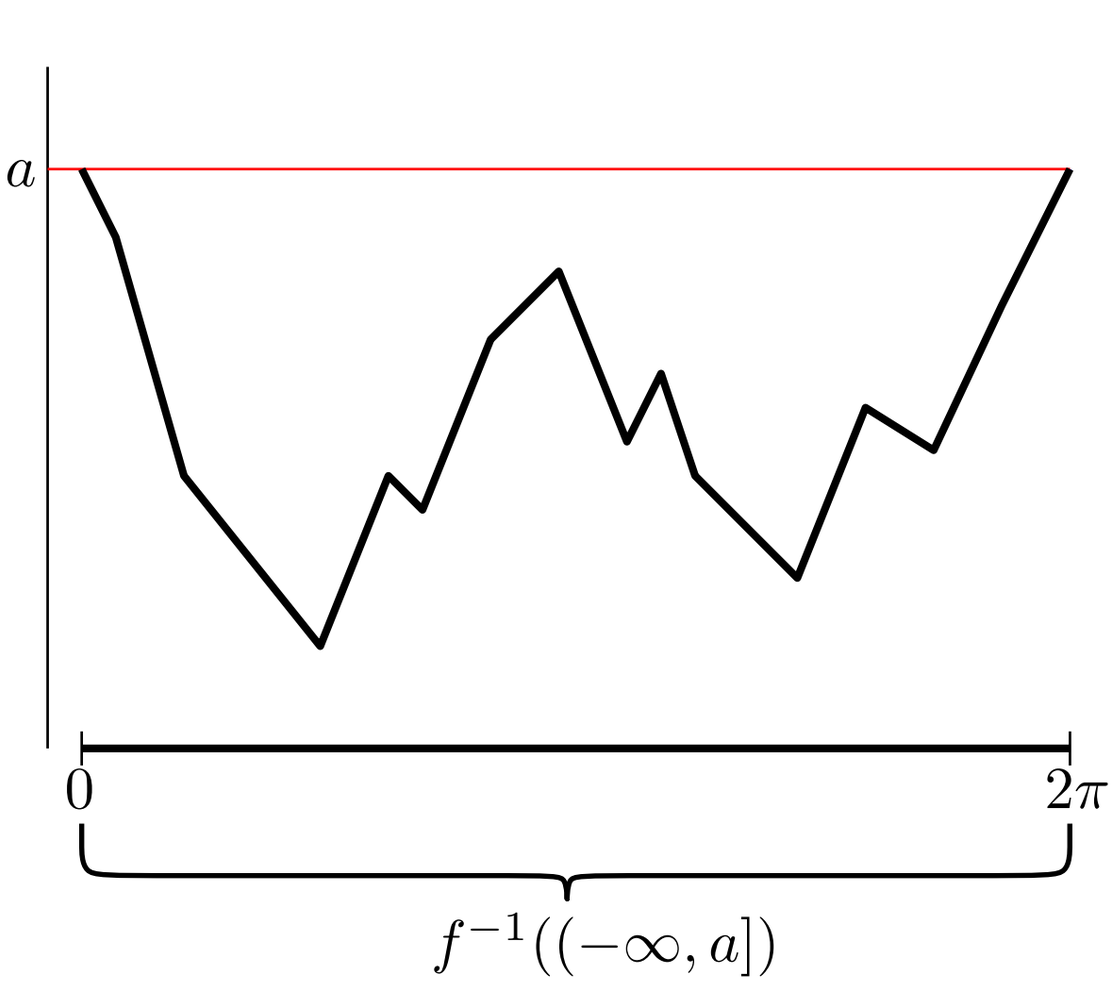
Persistent homology: Morse theory

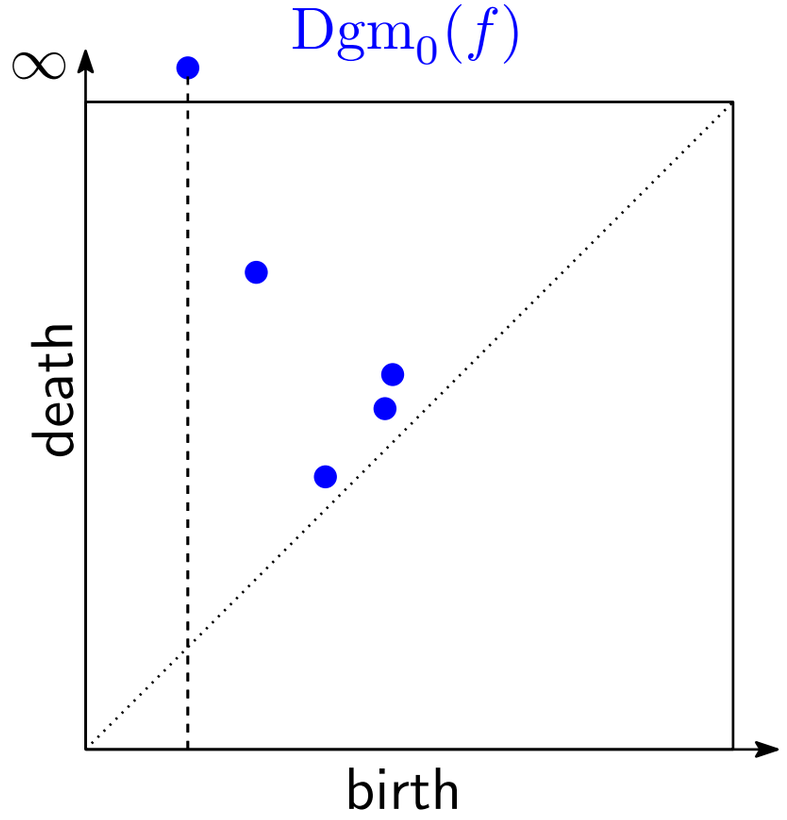
Evolution of homology as birth-death pair.



Persistent homology: Morse theory

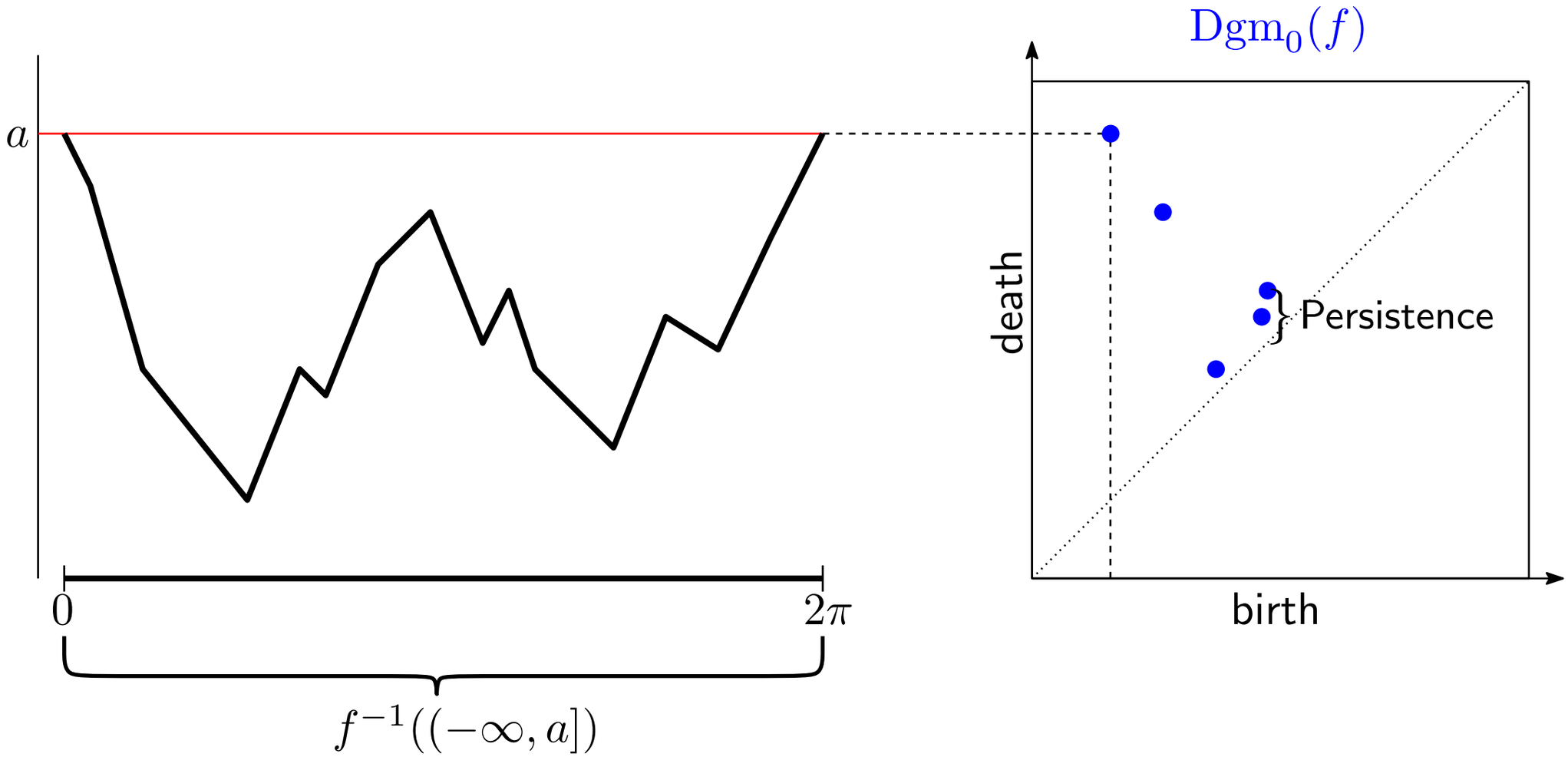
Evolution of homology as birth-death pair.



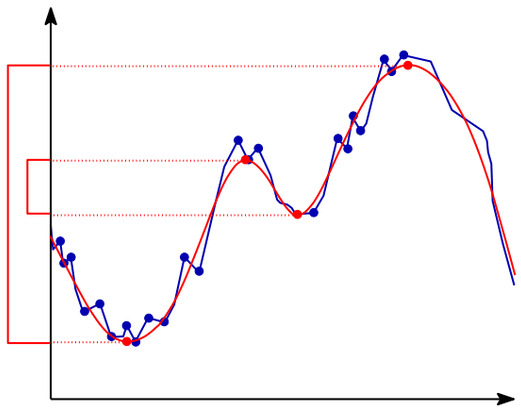


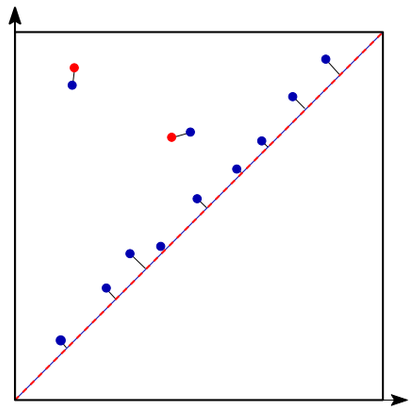
Persistent homology: Morse theory

Evolution of homology as birth-death pair.



Metrics on diagrams





L^2-Wasserstein distance

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

ϕ is the set of bijections between the points in X plus copies of the diagonal and points in Y with copies of he diagonals.

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

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

Smooth Euler characteristic transform (SECT)

The smooth Euler curve for each direction is

$$f(y) = \chi(M, v), \quad F(x) = \int_0^x f(y)dy - \overline{\chi(M, v)}.$$

Smooth Euler characteristic transform (SECT)

The smooth Euler curve for each direction is

$$f(y) = \chi(M, v), \quad F(x) = \int_0^x f(y)dy - \overline{\chi(M, v)}.$$

Definition

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

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

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

$$d_{\mathcal{M}_d}(M_1, M_2) := \sqrt{\int_{S^{d-1}} \left\| F(M_1, v) - F(M_2, v) \right\|^2 d\nu(v)}.$$

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

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

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

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

Injectivity and sufficiency of the ECT

Theorem (Turner-M-Boyer)

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

Injectivity and 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.*

Injectivty 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

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.

A sampling theory for shapes

How many directions to sample ?

A sampling theory for shapes

How many directions to sample ?

(1) For 2-D: 162 directions

(2) For 3-D: Over 700 directions.

A sampling theory for shapes

How many directions to sample ?

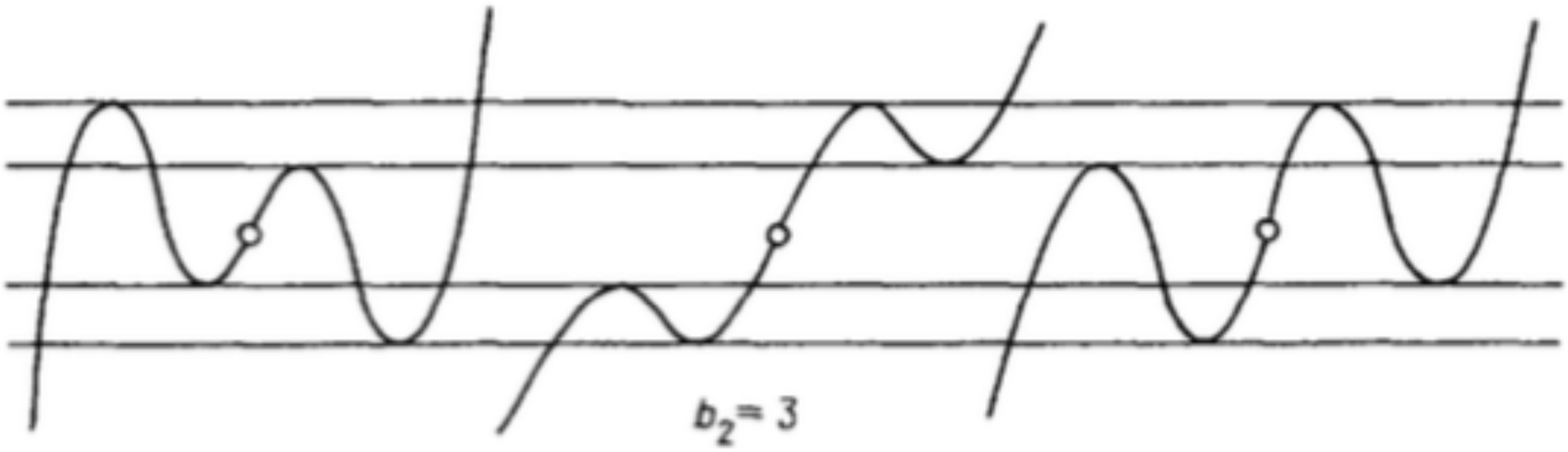
(1) For 2-D: 162 directions

(2) For 3-D: Over 700 directions.

A sampling theory for shapes — complexity metric for families shapes in terms of directions required.

Calculus of snakes

V.I. Arnol'd, The calculus of snakes and the combinatorics of Bernoulli, Euler and Springer numbers of Coxeter groups.



Exponential family and SECT

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 SECT

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 SECT

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

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

Distances without alignment

Theorem (Turner-M)

Let $f : S^2 \to \mathbb{Z}^{\mathbb{R}}$ and $g : S^2 \to \mathbb{Z}^{\mathbb{R}}$ 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

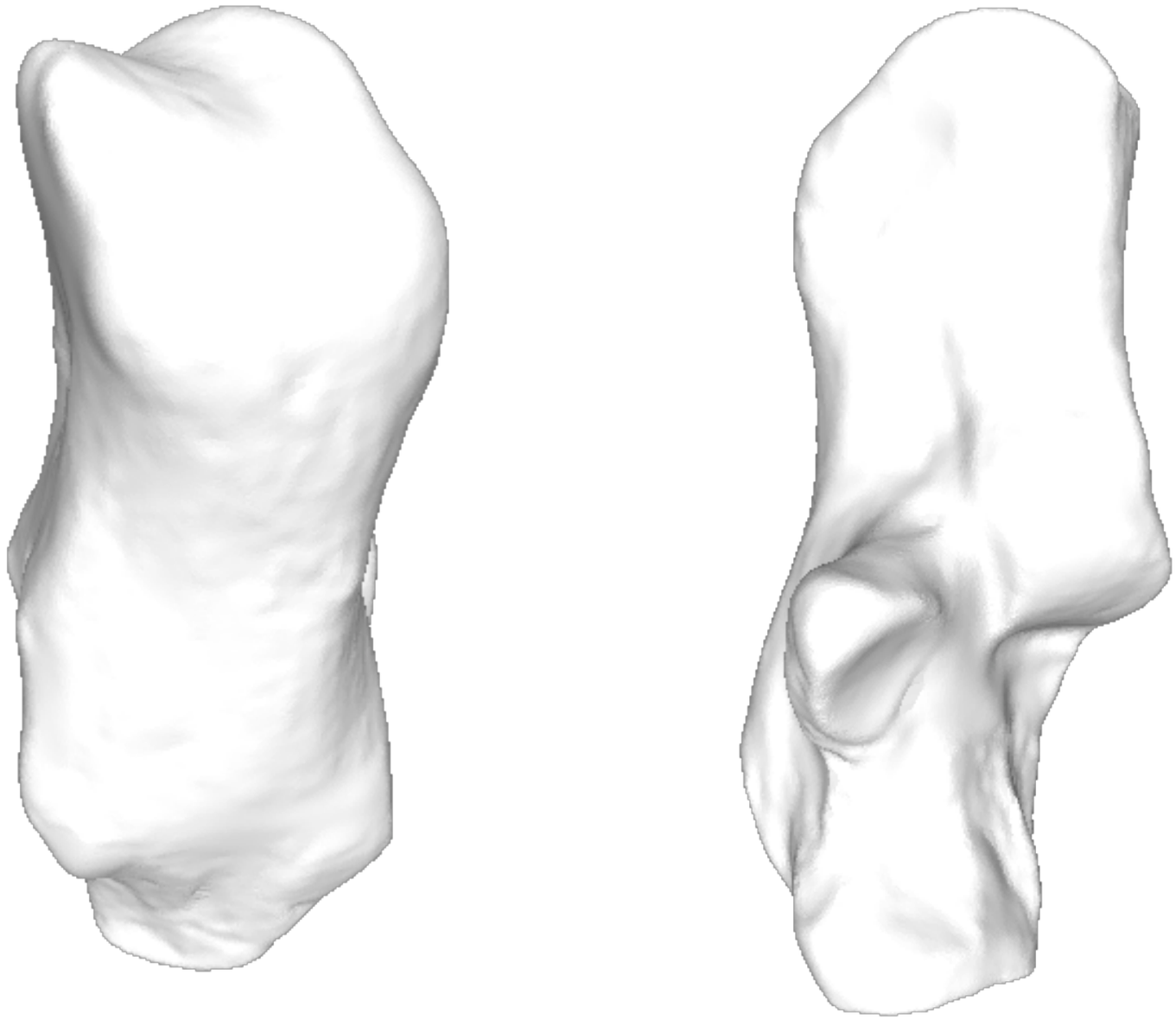
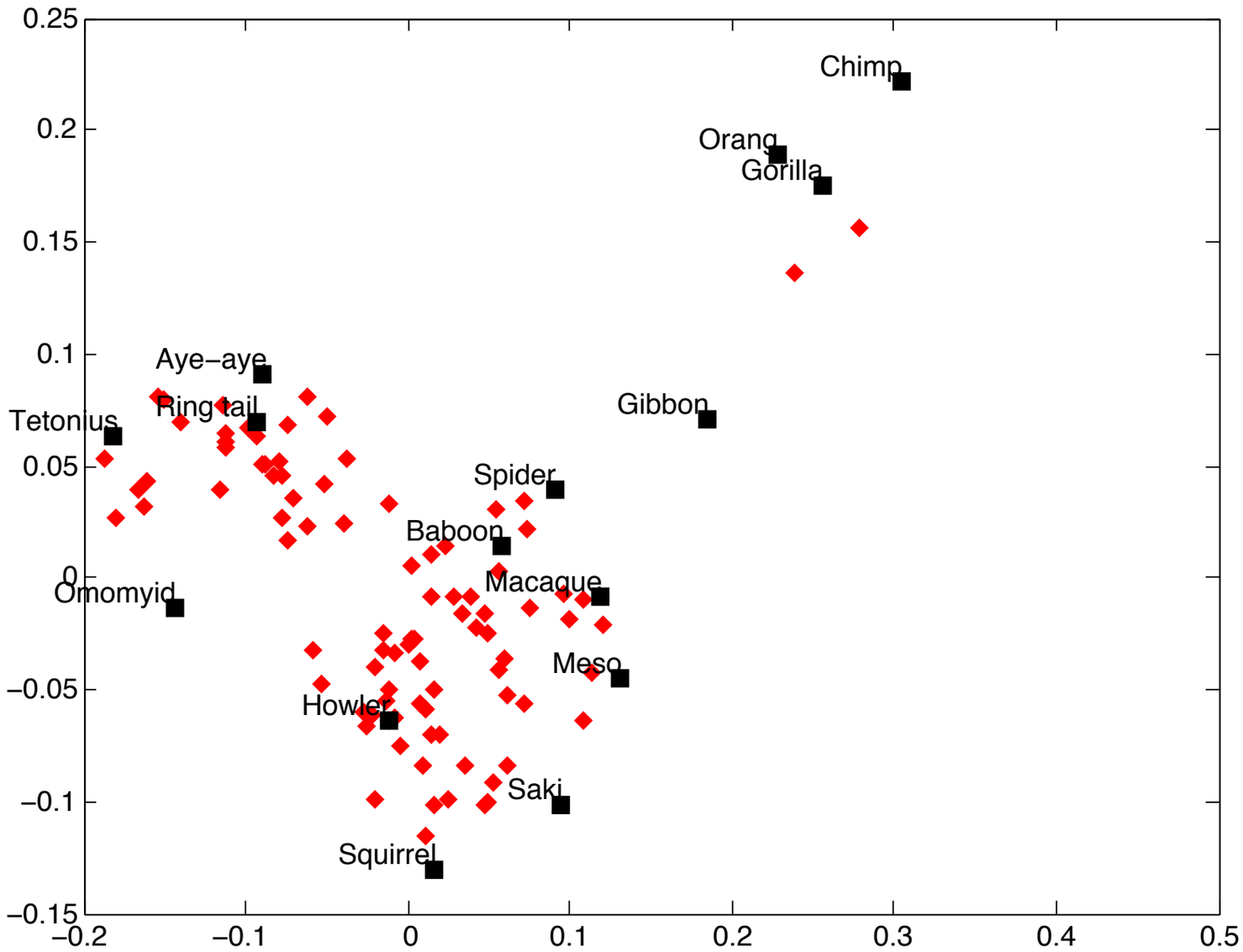
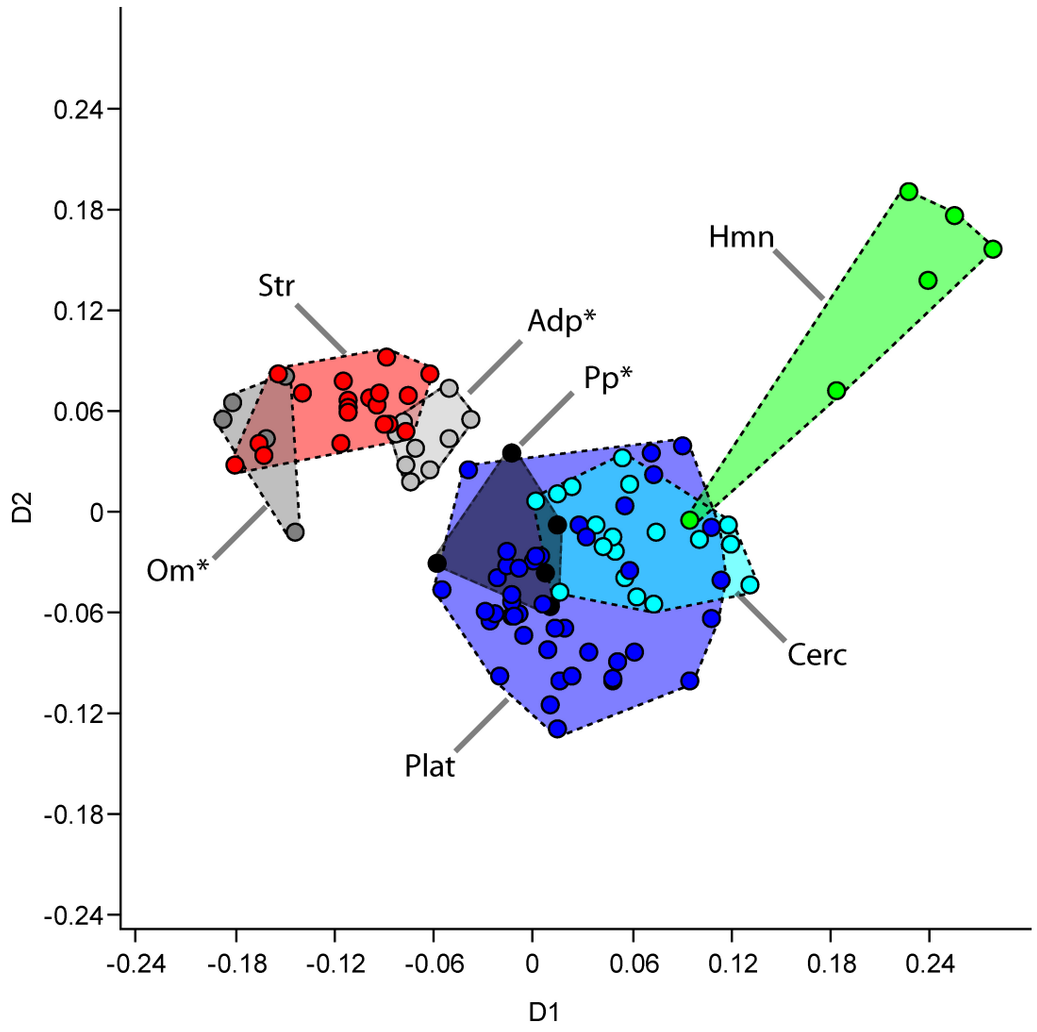


Figure: Images of a calcaneus from two different angles.

106 primates



Primate calcanei



Phylogenetic groups of primate calcanei with 67 genera. Asterisks indicate groups of extinct taxa. Abbreviations: Str, Strepsirrhines; Plat, platyrrhines; Cerc, Cercopithecoids; Om, Omomyiforms; Adp, Adapiforms; Pp, parapithecids; Hmn, Hominoids.

Comment from Doug

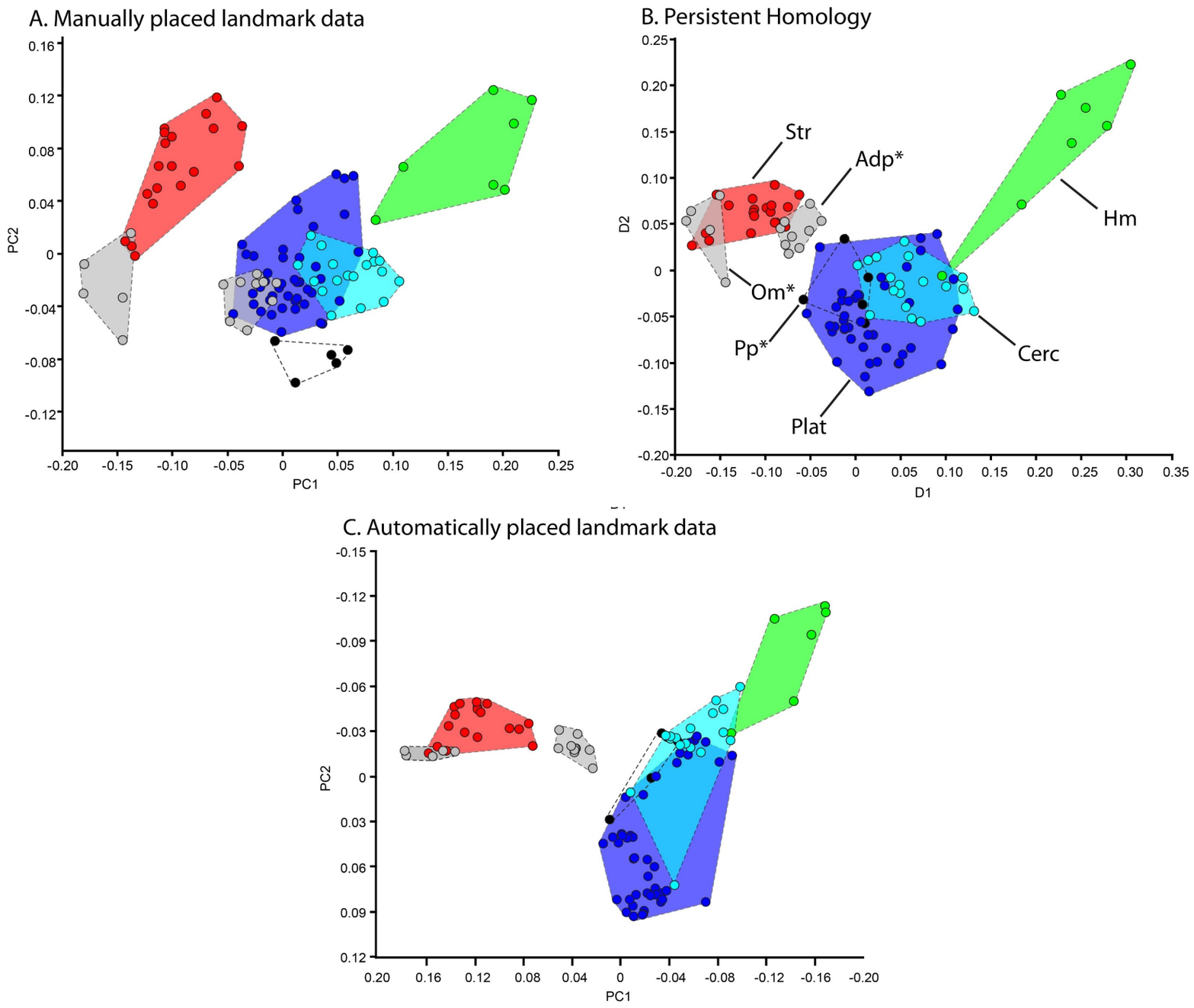
”In at least one way the method matched shapes with family groups better than any of the other previous methods... it linked a Hylobates specimen with the the other ape specimens (pan, gorilla, pongo, and oreopithecus). Previous both hylobatids (which ARE apes) always ended up closest to some Alouatta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

Comment from Doug

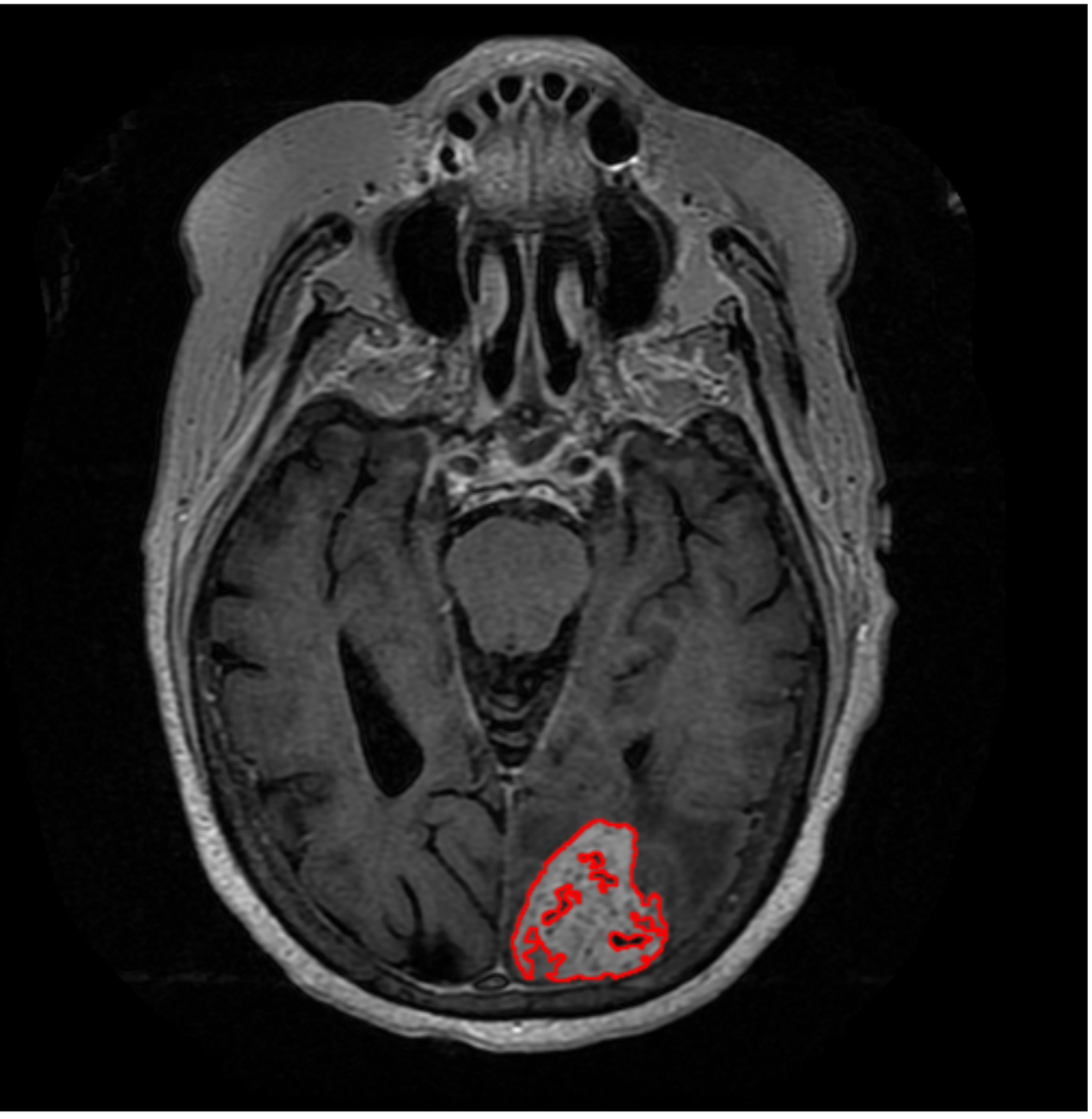
”In at least one way the method matched shapes with family groups better than any of the other previous methods... it linked a Hylobates specimen with the the other ape specimens (pan, gorilla, pongo, and oreopithecus). Previous both hylobatids (which ARE apes) always ended up closest to some Alouatta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

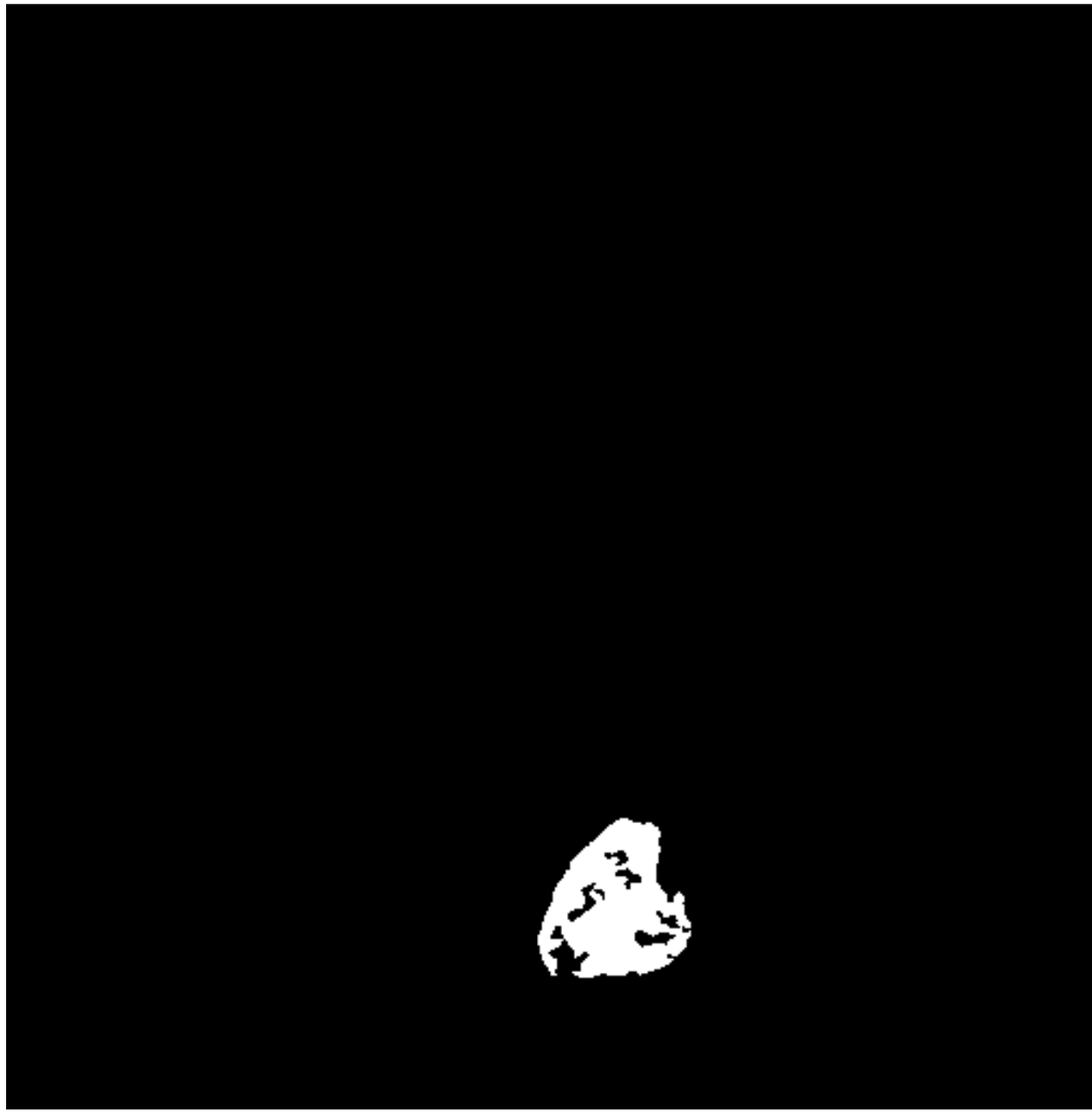
hylobatids = Gibbons
Alouatta = Howler monkey

Comparing methods



Glioblastoma and radiogenomics





The data

92 patients with matched gene expression and MRI data from the TCGA.

Gene expression: $p_g = 9215$
Morphometric features: $p_m = 212$
Volumetric features: $p_v = 5$
Topological features: $p_s = 7200$

The data

92 patients with matched gene expression and MRI data from the TCGA.

Gene expression: $p_g = 9215$
Morphometric features: $p_m = 212$
Volumetric features: $p_v = 5$
Topological features: $p_s = 7200$

Response:
Disease Free Survival (DFS): The period after a successful treatment during which there are no signs or symptoms of the cancer that was treated.

The data

92 patients with matched gene expression and MRI data from the TCGA.

Gene expression: $p_g = 9215$
Morphometric features: $p_m = 212$
Volumetric features: $p_v = 5$
Topological features: $p_s = 7200$

Response:
Disease Free Survival (DFS): The period after a successful treatment during which there are no signs or symptoms of the cancer that was treated.
Overall Survival (OS): The entire period after the start of treatment during which the cancer patient is still alive.

The model

Consider the following kernel/GP model for the SECT

$$f(\mathbf{F}) = \sum_{i=1}^{n} \alpha_i \, k(\mathbf{F}, \mathbf{F}_i), \quad k(\mathbf{F}_i, \mathbf{F}_j) = \exp(-\kappa \|\mathbf{F}_i - \mathbf{F}_j\|^2).$$

The model

Consider the following kernel/GP model for the SECT

$$f(\mathbf{F}) = \sum_{i=1}^{n} \alpha_i \, k(\mathbf{F}, \mathbf{F}_i), \quad k(\mathbf{F}_i, \mathbf{F}_j) = \exp(-\kappa \|\mathbf{F}_i - \mathbf{F}_j\|^2).$$

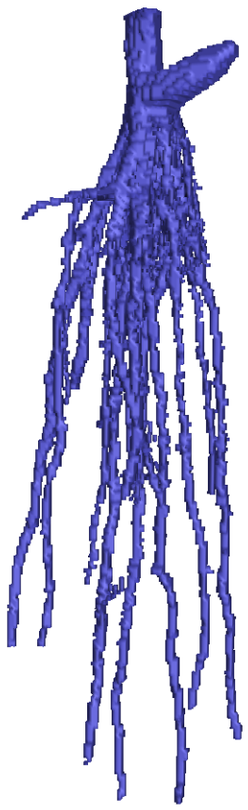
Can use standard functional data analysis. The same kernel model in part I of the talk for regression.

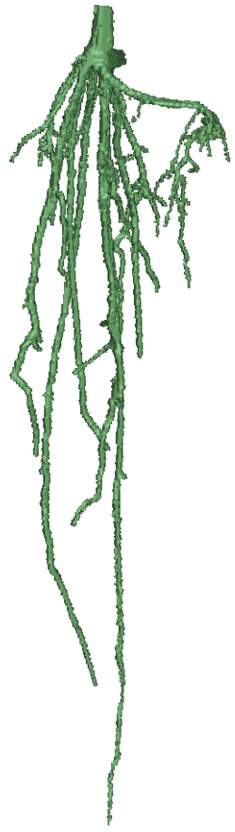
Results

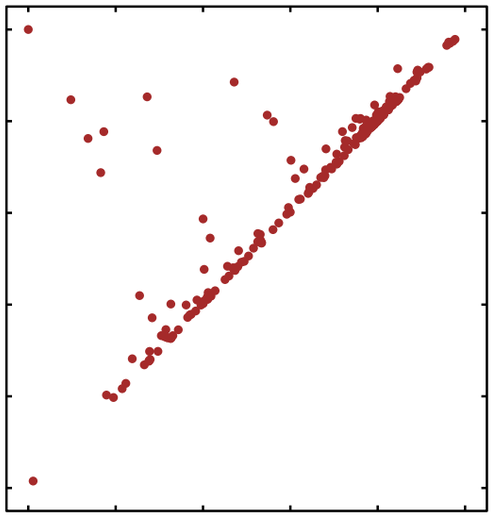
	Disease Free Survival		Overall Survival	
Data Type	RMSEP	Optimal%	RMSEP	Optimal%
Gene Expression	0.944 (0.035)	0.20	0.981 (0.030)	0.27
Morphometrics	0.942 (0.035)	0.07	0.965 (0.029)	0.15
Volume	0.939 (0.035)	0.04	0.964 (0.029)	0.16
SECT	**0.803 (0.035)**	**0.69**	**0.958 (0.028)**	**0.42**

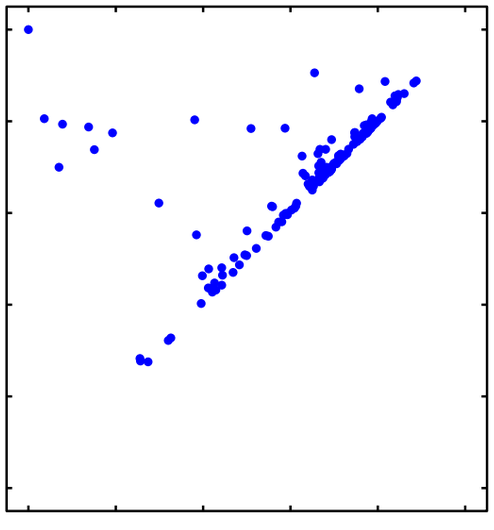
Table 1: **Results for predicting disease free survival (DFS) and overall survival (OS) using the Gaussian kernel function.** The first and third panels show comparisons of root mean squared errors of prediction (RMSEP) for the four considered data types. The second and fourth panels detail the percentage of the time that a model exhibits the lowest RMSEP. This is denoted as Optimal%. All values in bold represent the method with the lowest RMSEP or the method that most frequently performs best, respectively. These values are based on 100 random different 80-20 splits for each clinical outcome. Standard errors for each model are given the parentheses.

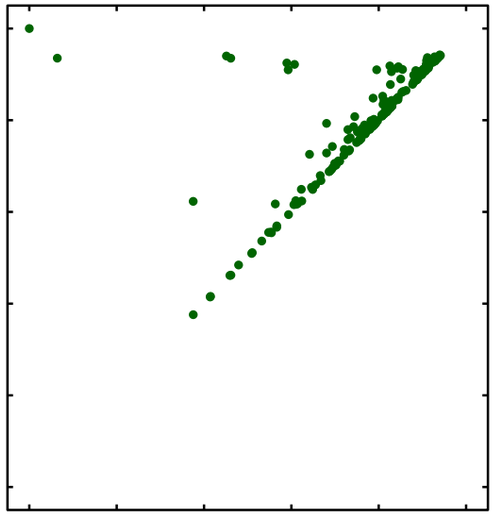
Association studies of shape phenotypes



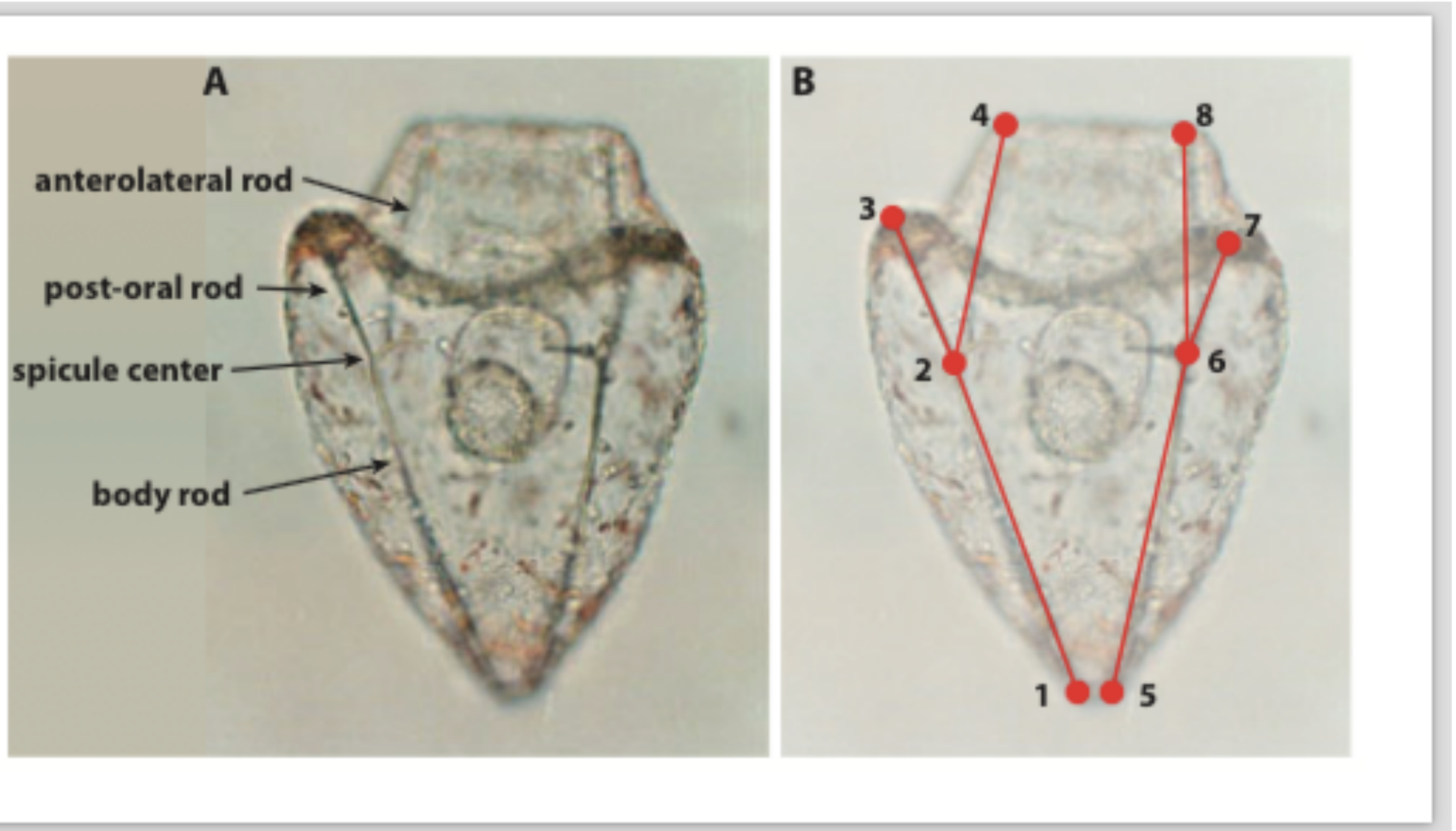






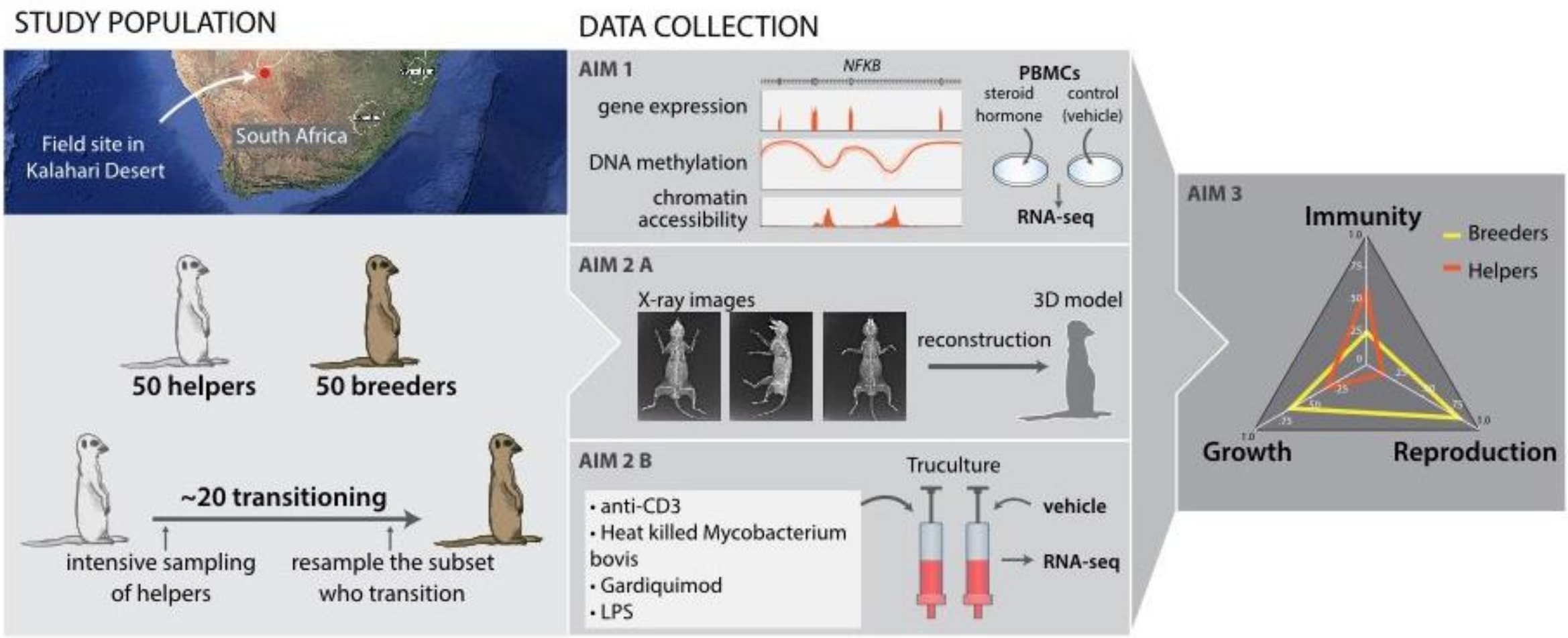


Quantitative genetics

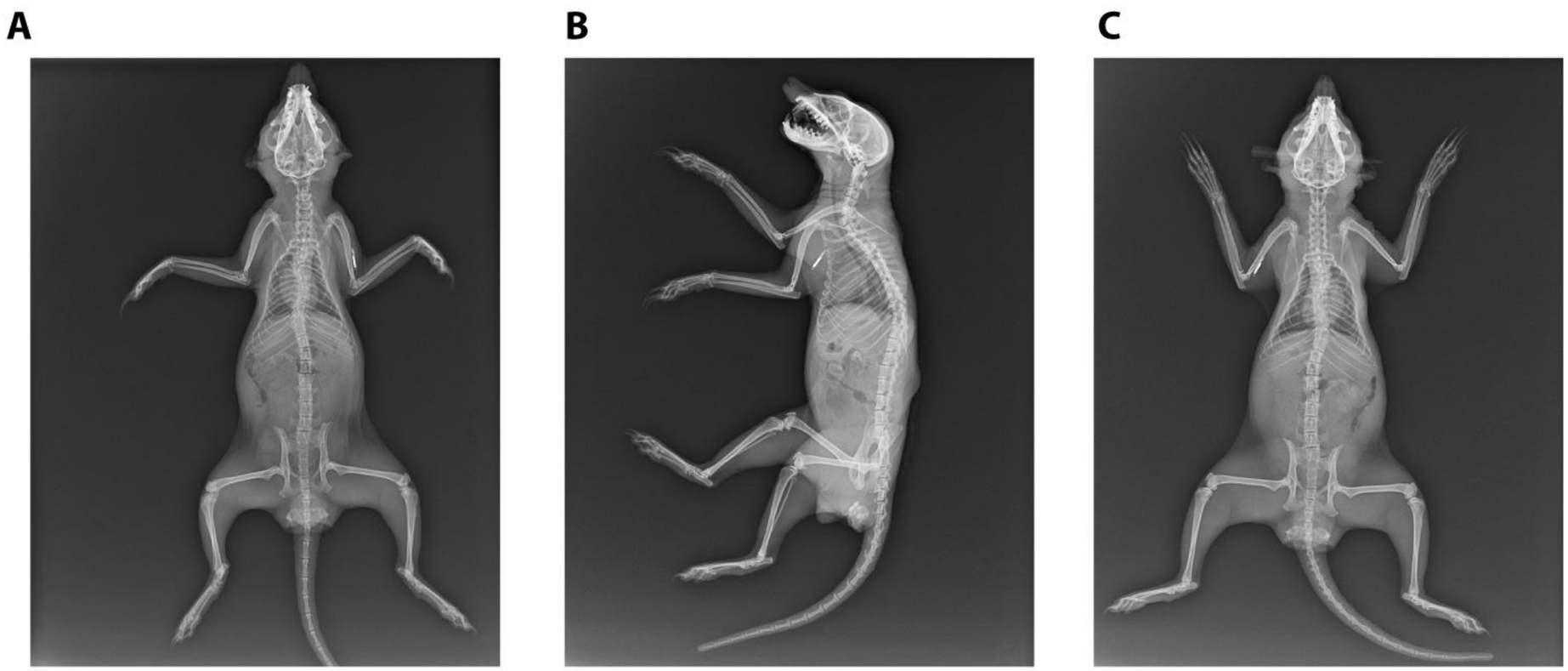


Evolution of cooperation in mammals

Evolution of cooperation in mammals



Evolution of cooperation in mammals



Open questions and problems

(1) Localized transforms: The PHT and ECT can be generalized as Euler integration

$$\int_X h \, d\chi, \quad h \text{ is a (localized) basis function.}$$

Open questions and problems

(1) Localized transforms: The PHT and ECT can be generalized as Euler integration

$$\int_X h \, d\chi, \quad h \text{ is a (localized) basis function.}$$

(2) Dictionary learning: representing a family of shapes as a sparse linear combination of prototype shapes.

Open questions and problems

(1) Localized transforms: The PHT and ECT can be generalized as Euler integration

$$\int_X h \, d\chi, \quad h \text{ is a (localized) basis function.}$$

(2) Dictionary learning: representing a family of shapes as a sparse linear combination of prototype shapes.

(3) Reconstructing shapes from summaries.

Open questions and problems

(1) Localized transforms: The PHT and ECT can be generalized as Euler integration

$$\int_X h \, d\chi, \quad h \text{ is a (localized) basis function.}$$

(2) Dictionary learning: representing a family of shapes as a sparse linear combination of prototype shapes.

(3) Reconstructing shapes from summaries.

(4) Phylogenies of surfaces and traits: we have a likelihood model, how do we get a substitution model.

Open questions and problems

(1) Localized transforms: The PHT and ECT can be generalized as Euler integration

$$\int_X h \, d\chi, \quad h \text{ is a (localized) basis function.}$$

(2) Dictionary learning: representing a family of shapes as a sparse linear combination of prototype shapes.

(3) Reconstructing shapes from summaries.

(4) Phylogenies of surfaces and traits: we have a likelihood model, how do we get a substitution model.

(5) Extending the diffeomorphism based approach to address cases where only subsets of the objects are diffeomorphic, learning transformations from data.

Open questions and problems

(1) Localized transforms: The PHT and ECT can be generalized as Euler integration

$$\int_X h \, d\chi, \quad h \text{ is a (localized) basis function.}$$

(2) Dictionary learning: representing a family of shapes as a sparse linear combination of prototype shapes.

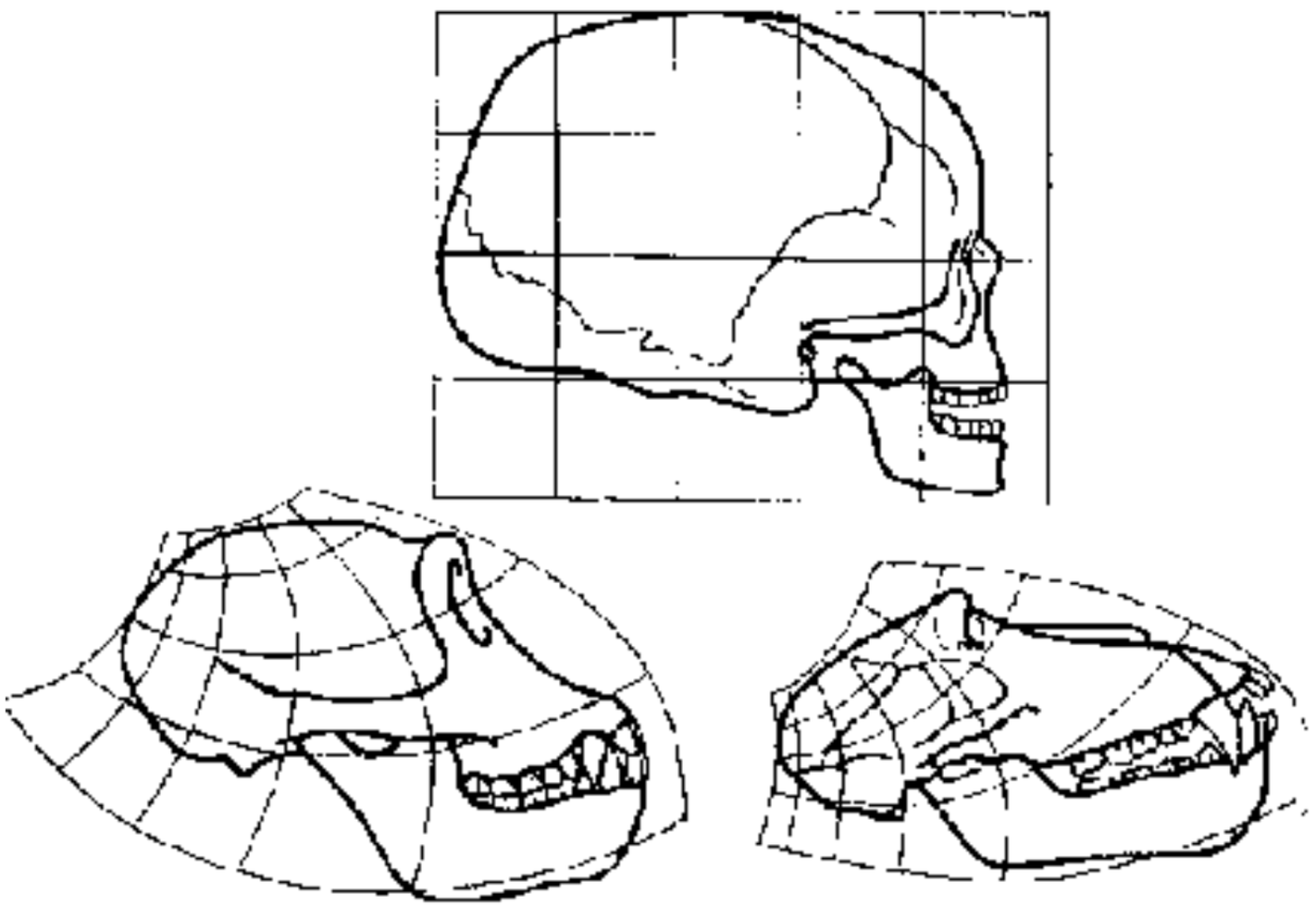
(3) Reconstructing shapes from summaries.

(4) Phylogenies of surfaces and traits: we have a likelihood model, how do we get a substitution model.

(5) Extending the diffeomorphism based approach to address cases where only subsets of the objects are diffeomorphic, learning transformations from data.

(6) Generalization to graphs and networks.

Modeling variation in shapes



Skulls of a human, a chimpanzee and a baboon
and transformations between them

D'Arcy Thompson, On Growth and Form

Thank
You
köszönöm
Kiitos
Vinaka
Maake
Asante
Shukria
Dhanyavadagalu
شكرا
Manana
Dankon
감사합니다
Dankscheen
спасибо
Kam Sah Hammida
Matondo
Terima Kasih
Diolch i Chi
Biyan
Mauruuru
Dank Je
Taiku
Tack
Blagodaram
Ngiyabonga
Dziekuje
Chokrane
Grazie
Arigato
Juspaxar
Mochchakkeram
Gracias
Tingki
நன்றி
Bedankt
Ďakujem
धन्यवाद
cảm ơn bạn
Khap
Paldies
Gratias Tibi
Ua Tsaug Rau Koj
Grazas
Di Ou Mèsi
Kia Ora
Obrigado
Děkuji
Nuwun
Nirringrazzjak
Kop Khun
ありがとう
Suksama
Rahmat
Matur
Hvala
谢谢
хвала
Welalin
Djiere Dieuf
Misaotra
Danke
Merci
Go Raibh Maith Agat
Tuke
Eskerrik Asko
Salamat
ขอขอบคุณคุณ
Najis

Acknowledgements

Part I. Dan Runcie, Julian Ayroles, Jenny Tung, Beth Hauser
Part II. Ingrid Debauchies, Yuliy Barishnikov, Jenny Tung, Washington Mio, Mao Li

Funding:

- NSF DMS, CCF, IIS
- DARPA
- AFOSR
- NIH