Modeling Multiple Correlated Functional Outcomes with Spatially Heterogeneous Shape Characteristics

David Ruppert, Kunlaya Soiaporn, Raymond Carroll

Cornell University, Cornell University, and Texas A&M University

Aug 6, 2013

- Functional Data Analysis (FDA) is a well-recognized and active research area
- On each subject one observes one of more functions such as
  - spectral data: power versus wavelength
  - weather data: temperature versus day of the year
- Much of the research on FDA assumes that the functions are generated by a Gaussian process
- Typically one function is observed on each subject

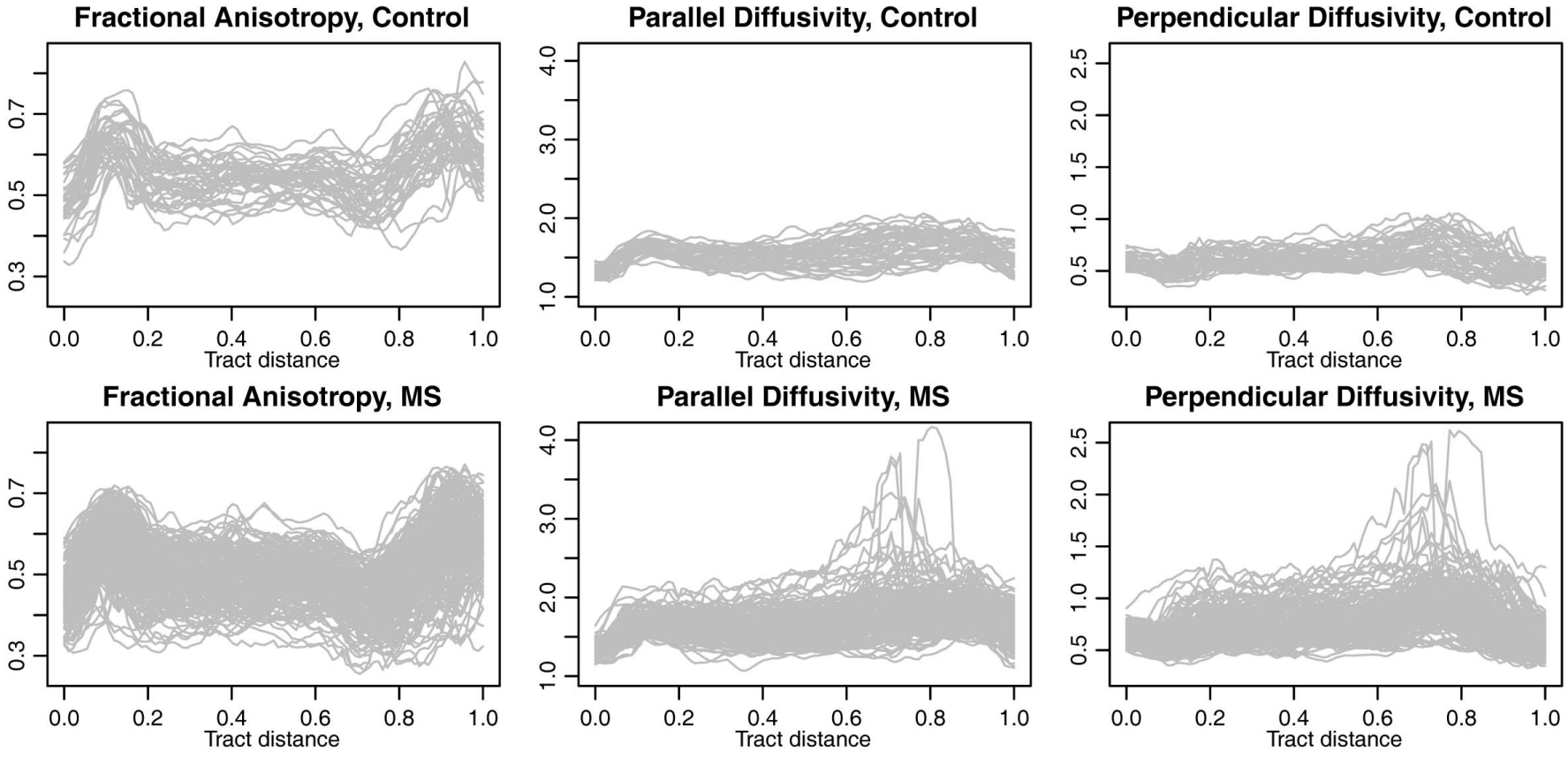
Diffusion Tensor Imaging

- We have been working with Diffusion Tensor Imaging (DTI) data
- White matter tracts connect different parts of the brain
- DTI measures the diffusion of water along the tract
- We have data on 162 multiple sclerosis (MS) patients and 42 controls

DTI Data

- The data are skewed, with the skewness spatially varying
- The amount of skewness is different for Multiple Sclerosis (MS) patients and controls
- There are several DTI signals on each tract
  - Fractional Anisotropy
  - Parallel Diffusivity
  - Perpendicular Diffusivity
- There are also multiple tracts, although we will only consider the corpus callosum

Plot of DTI Data



Top row: controls

Bottom row: MS patients – more variability and more skewness

Patients versus controls

- Patients and controls differ as much by variability and skewness as in their mean functions
- Might their correlation functions also differ?

Model For a Single Outcome

$Y_{ip}(\cdot) = p$th functional outcome observed on ith subject
Staicu, Crainineanu, Reich, and Ruppert (2011, *Biometrics*) proposed a copula model:

$$Y_{ip}(t) = \mu_p(t) + \sigma_p(t)\, G^{-1}\Big\{ W_{ip}(t); \alpha_p(t) \Big\}$$

- $\mu_p(\cdot)$ is the mean function
- $\sigma_p(t)$ is the standard deviation function
- $G(\cdot; \alpha)$ is a parametric family of CDFs with mean 0, standard deviation 1, and shape parameter α.
  - E.g., skewed Gaussian (Azzalini, 1985)
- $W_{ip}(t)$ is a process (in t) with unif(0,1) marginal distributions

From previous page:

$$Y_{ip}(t) = \mu_p(t) + \sigma_p(t)\Big[G^{-1}\big\{\underbrace{W_{ip}(t)}_{\text{unif}(0,1)}; \alpha_p(t)\big\}\Big]$$

It follows that:

- $E\{Y_{ip}(t)\} = \mu_p(t)$
- $\text{sd}\{Y_{ip}(t)\} = \sigma_p(t)$
- $Y_{ip}(t)$ has shape parameter $\alpha_p(t)$

Estimation of Marginal Distributions

Staicu et al. develop penalized spline estimators of

- $\mu_p(t)$
- $\sigma_p(t)$
- $\alpha_p(t)$
- copula model for within-function dependencies

Their methodology applies to the functional outcomes one-at-a-time.
Cross-dependencies cannot be studied without new methodology.

A copula is a multivariate CDF with uniform marginal distributions.

Copulas allow one to decompose the modeling of a multivariate distribution into two independent steps:

1. modeling the dependencies via a copula
2. modeling the univariate marginal distributions

We have already done Step 2.

Gaussian Copula Model for Dependence Structure

Copula model for W_{ip}:

- $R_{ip}(t) := \Phi^{-1}\{W_{ip}(t)\} \sim N(0,1)$
- For each p, assume $R_{ip},\ i = 1, \ldots, n$, are iid Gaussian processes with mean 0 and variance 1
- The correlation function of R_{ip} determines the dependence structure of W_{ip} and therefore of Y_{ip}
- Since R_{ip} is a Gaussian process, for any $t_1, \ldots, t_M$, $\{W_{ip}(t_1), \ldots, W_{ip}(t_M)\}$ has a Gaussian copula induced by the Gaussian distribution of $\{R_{ip}(t_1), \ldots, R_{ip}(t_M)\}$
- This is also the copula of $\{Y_{ip}(t_1), \ldots, Y_{ip}(t_M)\}$

Modeling Dependencies Across Outcomes

We model R_{ip} as the sum of a finite Karhuenen-Loève expansion and white noise:

$$R_{ip}(t) = \sum_{k=1}^{K_p} Z_{ipk} f_{kp}(t) + \epsilon_{ip}(t)$$

Here

- $Z_{ikp},\ k = 1, \ldots, K_p$, are independent $N\{0, \text{var}(Z_{ipk})\}$
- $f_{1p}, \ldots, f_{K_p p}$ are eigenfunctions of the covariance function of R_{ip}
- $\epsilon_{ip}(t)$ is white noise with a constant variance $\sigma^2_{\epsilon p}$

It follows from the above that

$$\sum_{k=1}^{K_p} f^2_{kp}(t) \text{var}(Z_{ipk}) + \sigma^2_{\epsilon p} \equiv 1.$$

Spline Model for the Eigenfunctions

Let $b(t) = \{b_1(t), \ldots, b_q(t)\}$ be an orthogonal spline basis.

Assume $f_{kp}(t) = h(t)^{\mathsf{T}}\theta_{kp}$ for some coefficient vector θ_{kp}.

Let Θ_p be the matrix with kth row equal to θ_{kp}.

We will use a penalty on Θ to prevent overfitting.

Pseudo-Likelihood

We use pseudo-likelihood estimation which has two stages:

1. Estimate the parameters in the marginal distributions
2. Estimate the copula parameters by
   - plugging the marginal parameter estimates into the penalized log-likelihood,
   - acting as if they were the true parameters, and
   - maximizing this pseudo penalized log-likelihood over the copula parameters.

Let

$$\begin{aligned} R_{ip} &= \{R_{ip}(t_1), \ldots, R_{ip}(t_m)\}^{\mathsf{T}} \text{ (data)} \\ B &= \{b(t_1), \ldots, b(t_m)\}^{\mathsf{T}} \text{ (known)} \\ \epsilon_{ip} &= \{\epsilon_{ip}(t_1), \ldots, \epsilon_{ip}(t_m)\}^{\mathsf{T}}. \end{aligned}$$

Then

$$\begin{aligned} R_{ip} &= B\Theta_p Z_{ip} + \epsilon_{ip} \\ \epsilon_{ip} &\sim \mathsf{N}(0, \sigma^2_{\epsilon p} I_m), \\ Z_{ip} &\sim \mathsf{N}(0, D_p), \text{ (unobserved latent variables)} \\ \mathsf{cov}(Z_{ip}, Z_{ip'}) &= C_{pp'}, \quad \text{for} \quad p \neq p'. \end{aligned}$$

We must estimate the parameters Θ_p, D_p, and $C_{pp'}$, $p, p' = 1, \ldots, P$

We also need to deal with the latent $Z_i = (Z_{i1}, \ldots, Z_{iP})$, $i = 1, \ldots, n$.

Pseudo-Data:

$$R_i = \begin{pmatrix} R_{i1} \\ \vdots \\ R_{iP} \end{pmatrix}, \quad i = 1, \ldots, n.$$

The E-step: Compute the conditional distribution of Z_i given R_i

The M-Step: Update the parameter estimates by minimizing

$$-2E\left\{\sum_{i=1}^{N}\log L(R_i,Z_i)\middle|R_i\right\}+\sum_{p=1}^{P}\lambda_p\sum_{k=1}^{K_p}\Theta_{pk}^T\int b''(t)b''(t)^T dt\Theta_{pk}$$

White Matter Tracts, DTI, and MS

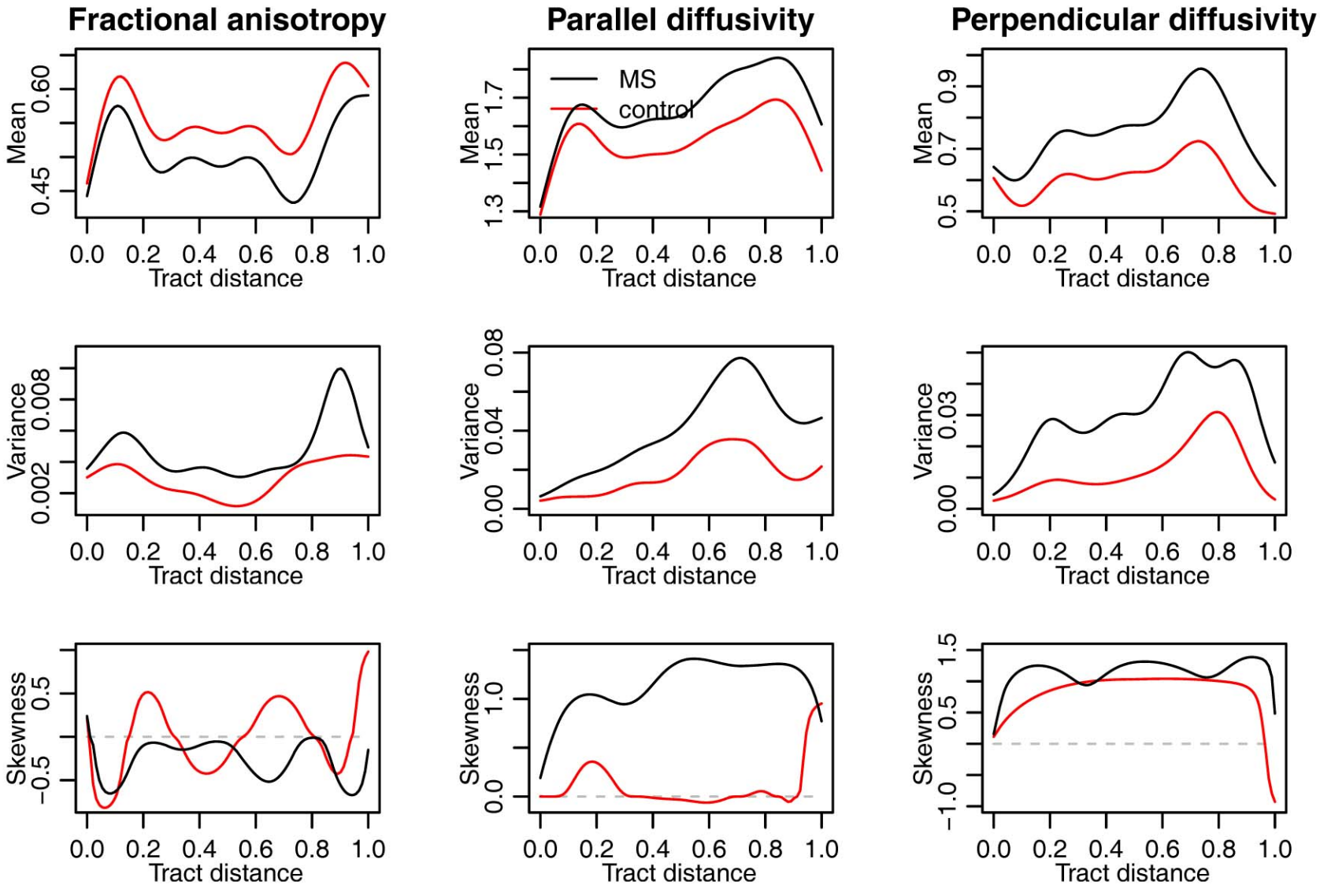
- White matter tracts are made up of axons that transmit signals between different regions of the brain.
  - These axons are insulated by a fatty substance called myelin.
- Multiple sclerosis is an autoimmune disease associated with damage to myelin.
  - Can lead to significant disabilities.
- DTI is a magnetic resonance imaging technique that measures the diffusion of water in tissue.
  - Anisotropy of water diffusion allows images of the white matter in the brain to be generated.
- A subset of our DTI data set is available in the `refund` package of `R`

Three DTI Outcomes

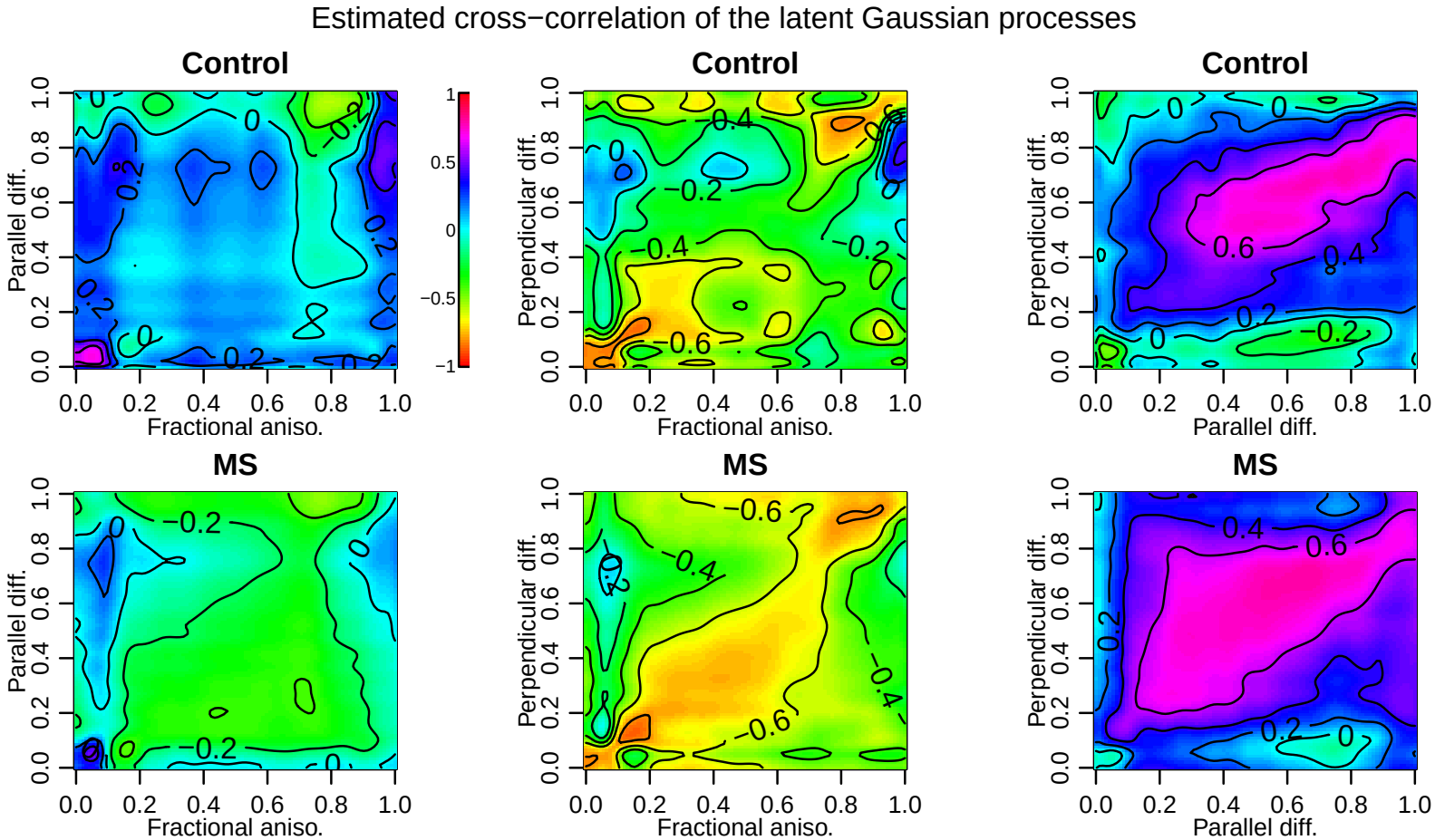
- At each tract location one obtains a 3×3 positive-definitive, symmetric matrix
  - the matrix describes the diffusion at that location
- Let $\lambda_1 > \lambda_2 > \lambda_3$ be the eigenvalues
- Parallel diffusivity $= \lambda_1$
- Perpendicular diffusivity $= (\lambda_1 + \lambda_2)/2$
- Fractional anisotropy $=$

$$\left[\frac{3\left\{(\lambda_1 - \bar{\lambda})^2 + (\lambda_2 - \bar{\lambda})^2 + (\lambda_3 - \bar{\lambda})^2\right\}}{2(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)} \right]^{1/2}$$

Marginal Distributions: Cases versus Controls

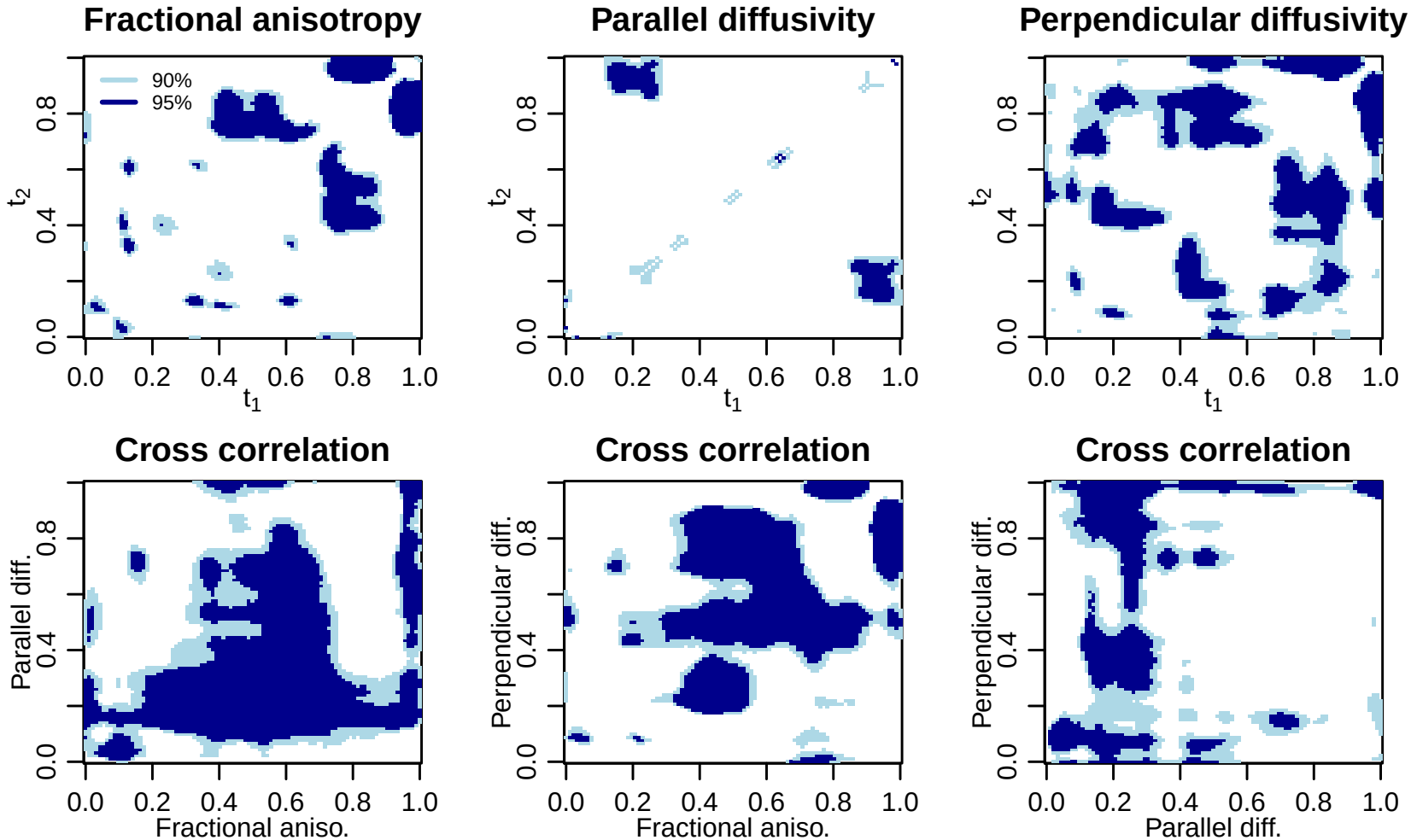


Cross-Correlations



Differences Between Correlation Functions

Differences of the correlations between the two groups

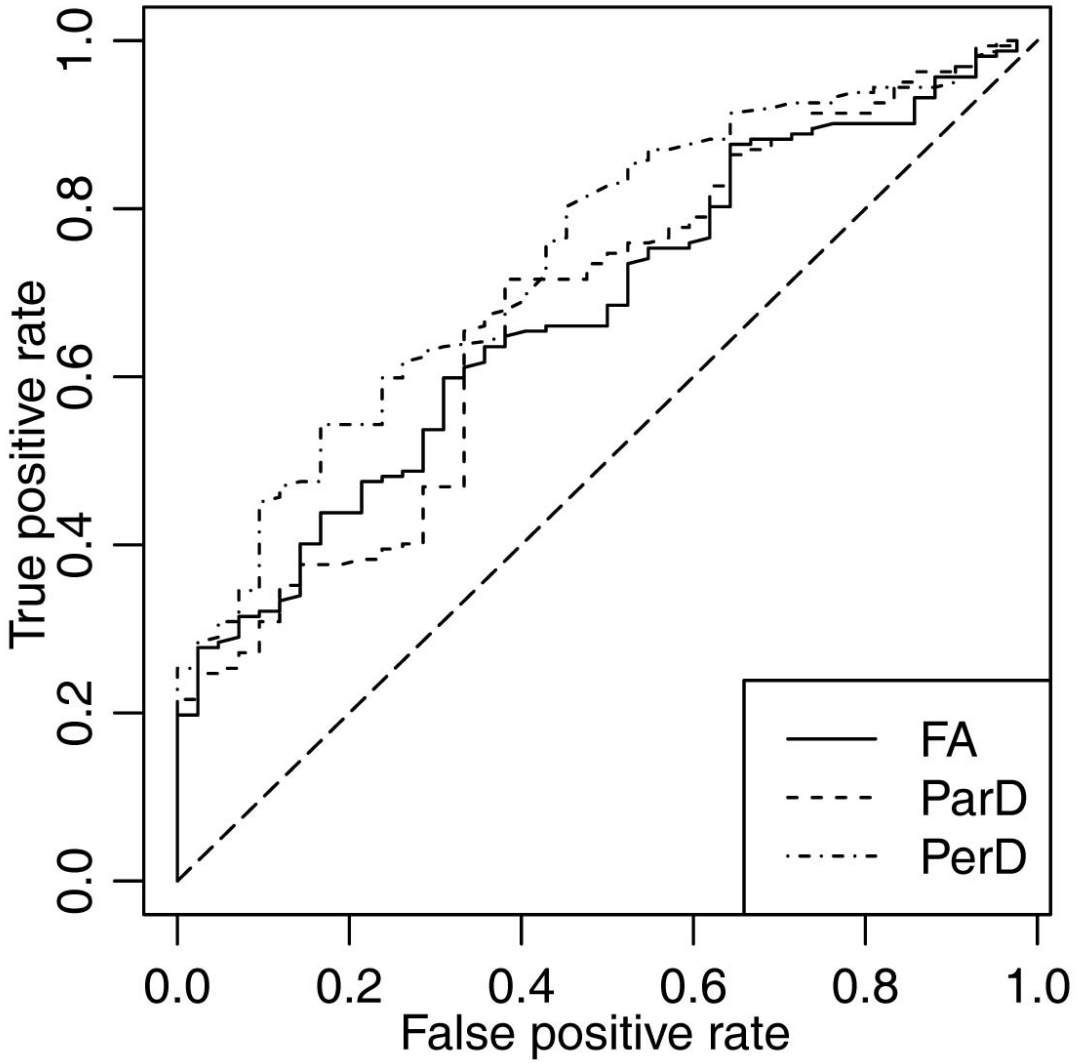


Predicting Case Status from Correlations

We predicted each outcome from the other two outcomes

- The prediction was done twice using the parameters estimated for
  1. cases
  2. controls
- The area between the predicted and each of the observed curves was computed
  - The subject was classified as case/control according to which area was smallest
- Cross-validation to prevent over-optimism
  - parameters were estimated without data from the subject being classified

CV-ROC Curve for Predicting Case Status



Thanks for coming!