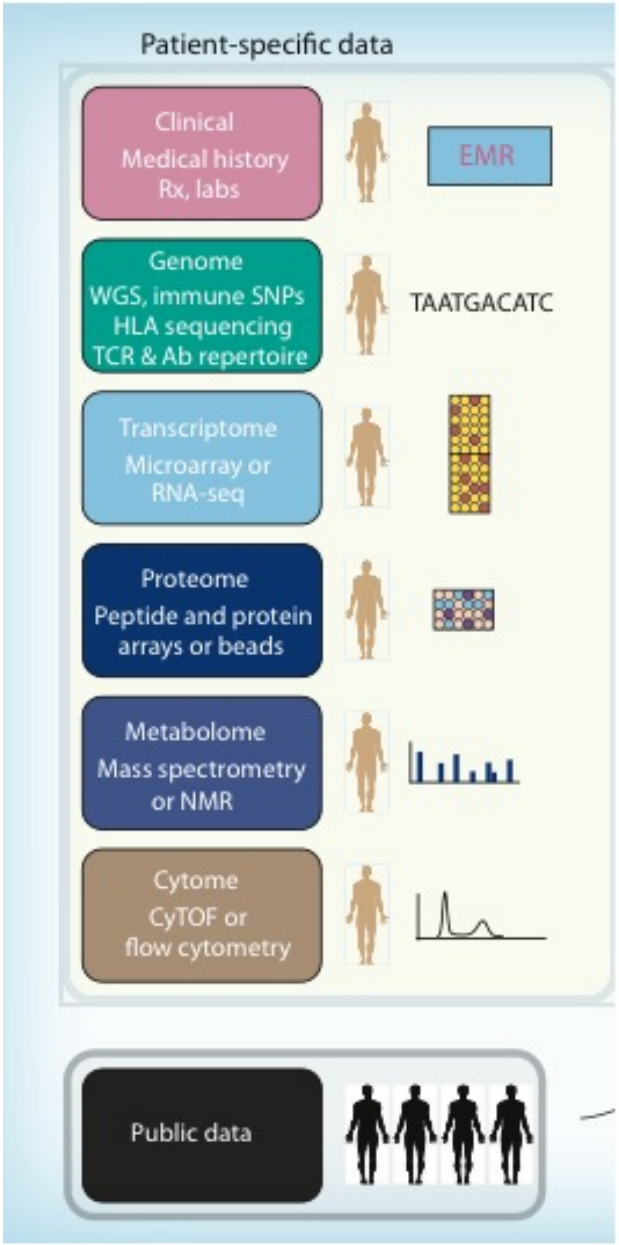
Bayesian Models for Richly Structured Data in Biomedicine

Veera Baladandayuthapani
The University of Texas M.D. Anderson Cancer Center
veera@mdanderson.org

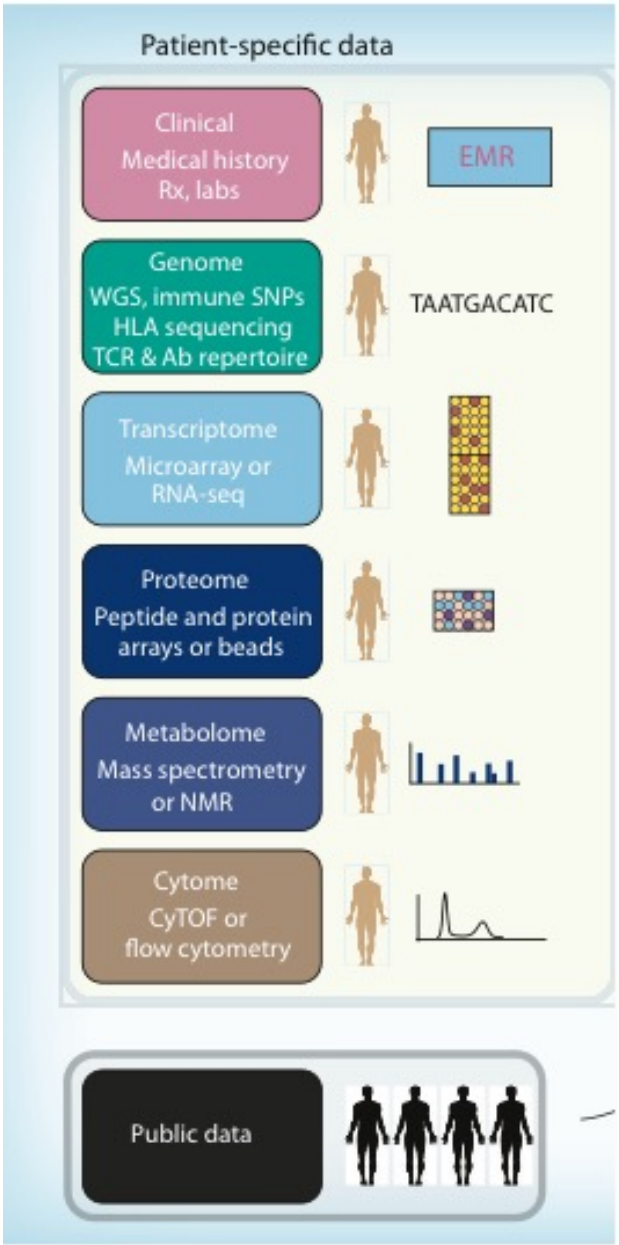
Joint work with K. Bharath, S. Kurtek, A. Rao, A. Saha, H. Yang, J. S. Morris

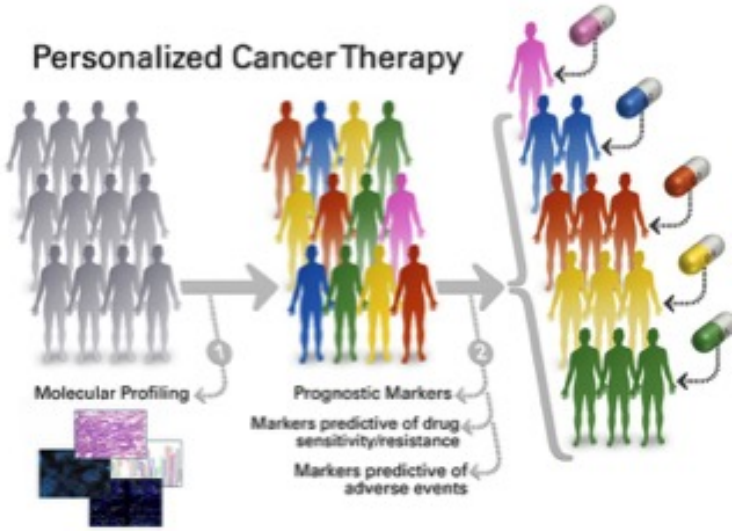
Precision (Personalized/Stratified) Medicine Continuum



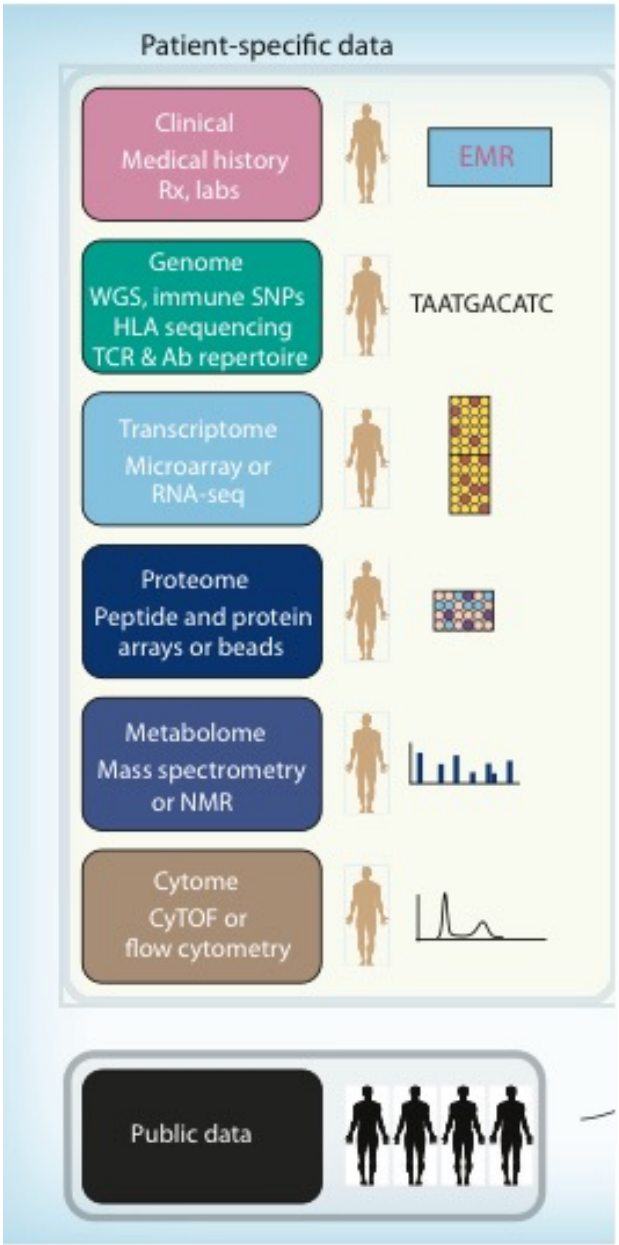
(Kidd et al)

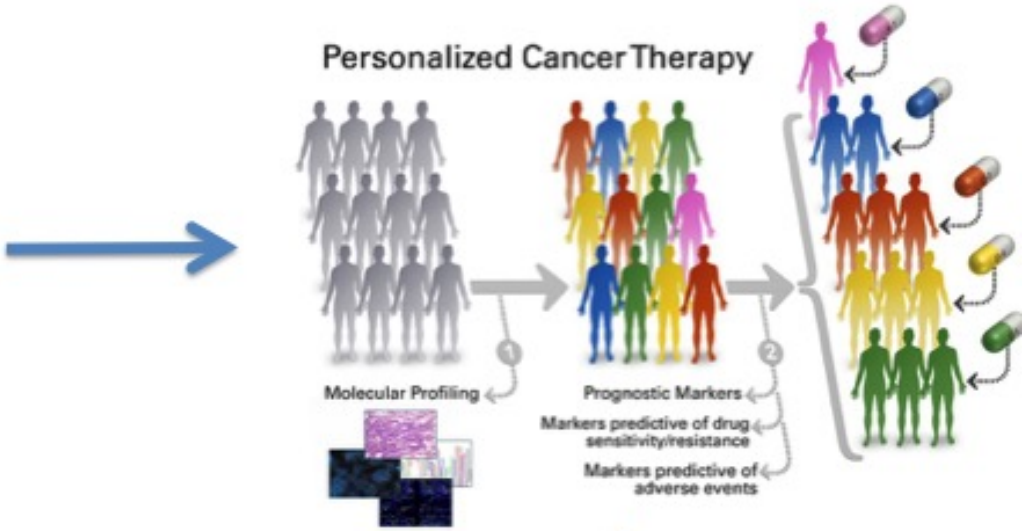
Precision (Personalized/Stratified) Medicine Continuum

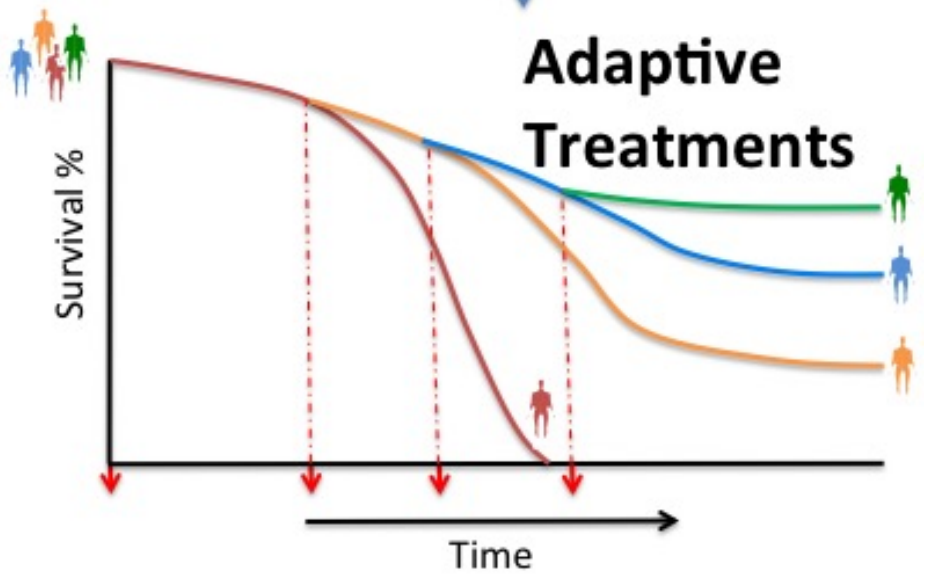




Precision (Personalized/Stratified) Medicine Continuum



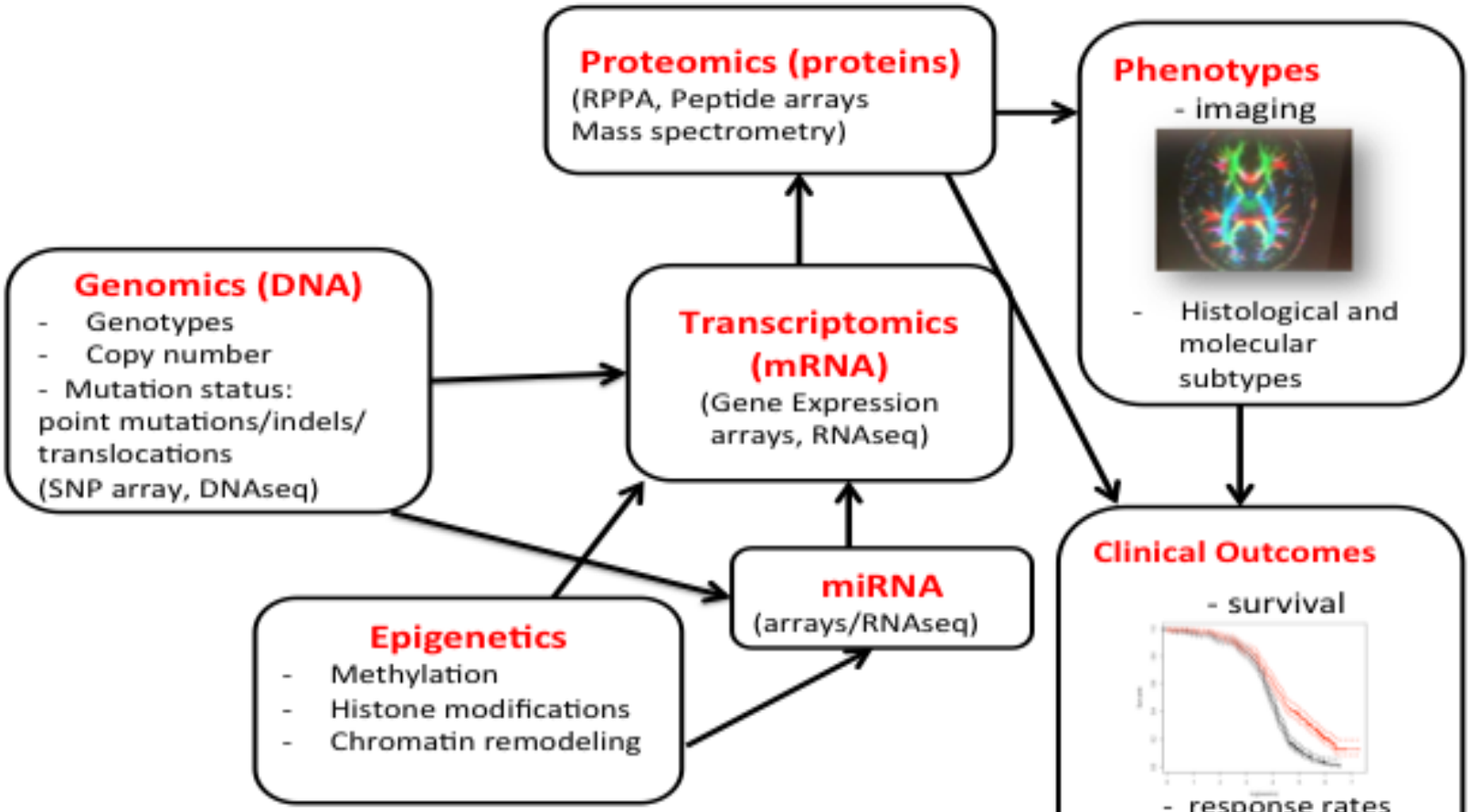




Motivation

- Cancer is one of the most well-characterized path-biological disease systems at different molecular levels
- Multiple types of high-throughput data now available on the multiple (matched) tumor samples
  - Genomics (multiple cancers): The Cancer Genome Atlas (TCGA, cancergenome.nih.gov); International Cancer Genome Consortium (ICGC, icgc.org); Genomic Data Commons (gdc.cancer.gov) Cancer Moonshots!
  - Imaging: The Cancer Imaging Archive (TCIA, cancerimagingarchive.net)
  - Key: same set of samples (not meta-analysis)

Multiple genomes at play



Morris and Baladandayuthapani, 2017

Beyond 'Omics

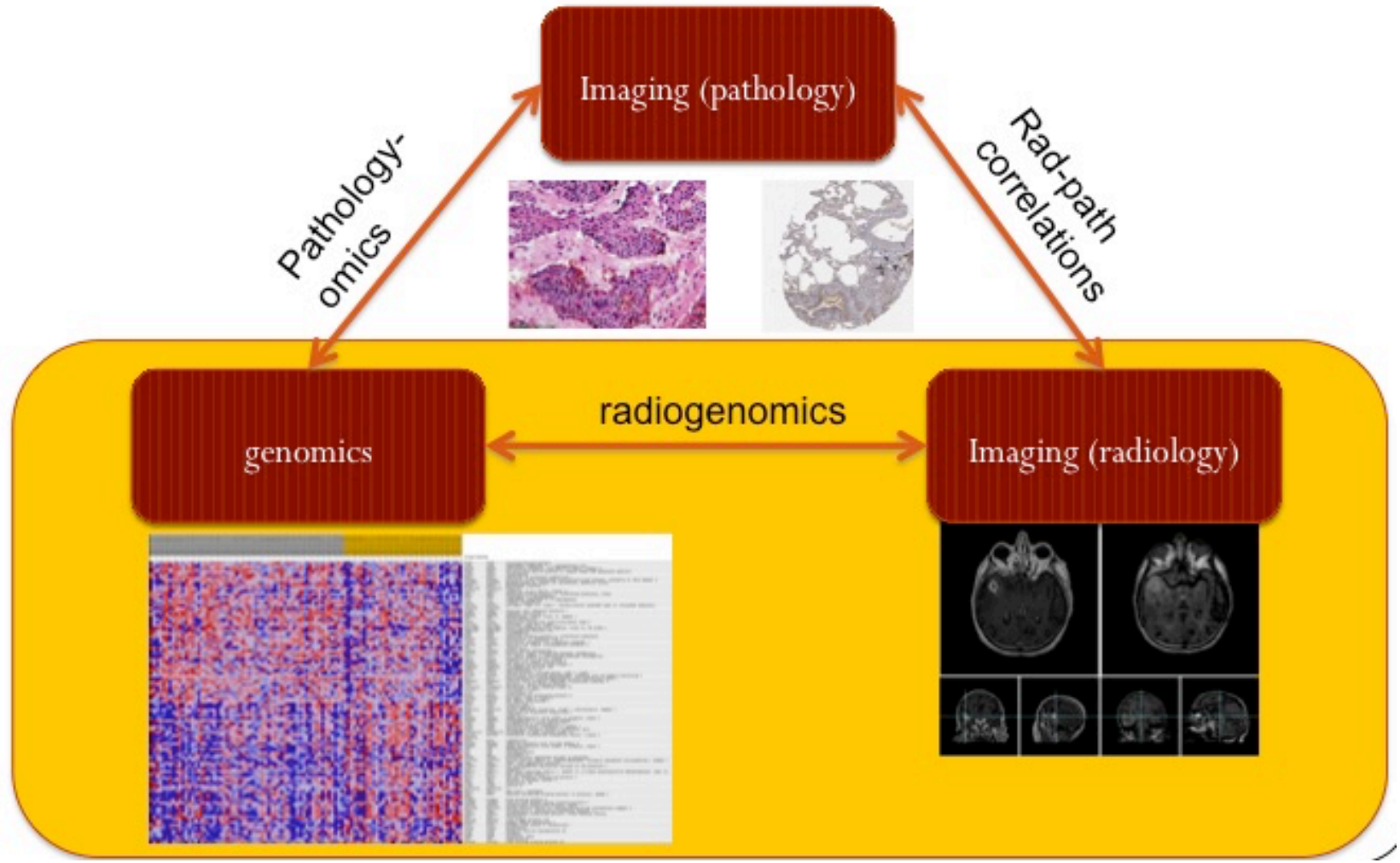


Image courtesy: Arvind Rao

Goals

Two competing continuums:

Stability/Ease of Measurement $\Leftrightarrow$ Biological/Phenotypic Relevance

Goals

Two competing continuums:

Stability/Ease of Measurement $\Leftrightarrow$ Biological/Phenotypic Relevance

- Main scientific goals:
  - Single type of alteration tells only part of the story
  - Systems-level: understand basic cancer biology (regulatory mechanisms)
  - Translational-level: correlation with clinical outcomes; biomarker discovery; personalized/precision medicine

Goals

Two competing continuums:

Stability/Ease of Measurement $\Leftrightarrow$ Biological/Phenotypic Relevance

- Main scientific goals:
  - Single type of alteration tells only part of the story
  - Systems-level: understand basic cancer biology (regulatory mechanisms)
  - Translational-level: correlation with clinical outcomes; biomarker discovery; personalized/precision medicine
- Statistically (or analytically): joint models for information rich, complex-structured, heterogenous, multi-modal high-dimensional data

Practical Challenges to Data Integration

- Missing data
  - Sample size shrinks when "matching" samples
- Experimental design/batch effects/preprocessing
  - Systematic biases/noise
  - Worse for complex, high-dimensional data
  - Each platform has own challenges/difficulties
- Data management
  - Management of large data sets
  - Ability to link genomic, imaging, clinical and electronic medical record data
- Choice of modeling unit
  - Different platforms have different observational units (probes, segments etc)
  - How to match up elements across platforms (genes/proteins?)
  - Data on different scales (continuous/ordinal/discrete/non-euclidean)

Statistical Contributions to Bioinformatics: Design, Modeling, Structure Learning, and Integration
Morris and Baladandayuthapani (2017+, Stat Modeling Discussion paper & Rejoinder)

Practical Challenges to Data Integration

- Missing data
  - Sample size shrinks when "matching" samples
- Experimental design/batch effects/preprocessing
  - Systematic biases/noise
  - Worse for complex, high-dimensional data
  - Each platform has own challenges/difficulties
- Data management
  - Management of large data sets
  - Ability to link genomic, imaging, clinical and electronic medical record data
- Choice of modeling unit
  - Different platforms have different observational units (probes, segments etc)
  - How to match up elements across platforms (genes/proteins?)
  - Data on different scales (continuous/ordinal/discrete/non-euclidean)
- Lurking structured dependencies between multi-modal data sources

Statistical Contributions to Bioinformatics: Design, Modeling, Structure Learning, and Integration
Morris and Baladandayuthapani (2017+, Stat Modeling Discussion paper & Rejoinder)

Structured Dependencies

- Both biological and induced by experimental design

Structured Dependencies

- Both biological and induced by experimental design
- Biological dependencies
  - Serial genomic-location correlation (copy number, methylation)
  - Pathway based correlations (mRNA, protein expression)
  - Mechnanistic (DNA $\longrightarrow$RNA, RNA $\longrightarrow$protein etc)
  - Non-linear interactions
  - Intra- and Inter- tumor heterogeneity
- Experimental/Design-based
  - Sampling based; treatment subgroups; biomarker-based randomized clinical trials (BATTLE, I-SPY trials)
  - Spatial characterizations of tumor development; imaging

Structured Dependencies

- Both biological and induced by experimental design
- Biological dependencies
  - Serial genomic-location correlation (copy number, methylation)
  - Pathway based correlations (mRNA, protein expression)
  - Mechnanistic (DNA $\longrightarrow$RNA, RNA $\longrightarrow$protein etc)
  - Non-linear interactions
  - Intra- and Inter- tumor heterogeneity
- Experimental/Design-based
  - Sampling based; treatment subgroups; biomarker-based randomized clinical trials (BATTLE, I-SPY trials)
  - Spatial characterizations of tumor development; imaging
- many more...
- profound implications in modeling and interpretations

Examples

- Pathway based correlations (mRNA, protein expression) (think graphical/network models!)
- Mechnanistic (DNA $\longrightarrow$RNA, RNA $\longrightarrow$protein etc) (think hierarchical models!)
- Non-linear interactions (think non-parametric models!)
- Serial correlation (copy number, methylation) (think functional data models!)
- Spatial characterizations; imaging (think spatial process models!)
- Combine diverse genomics data (integrative models...integromics!)

Radiogenomics

- "Radiomics": Extraction of large numbers of image features from radiological data (CT/MRI etc); radio-phenotypes
- "Radiogenomics": Radiomics + genomics

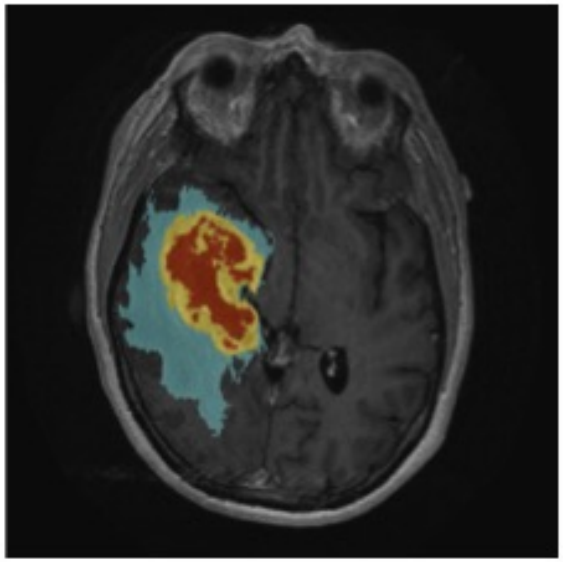
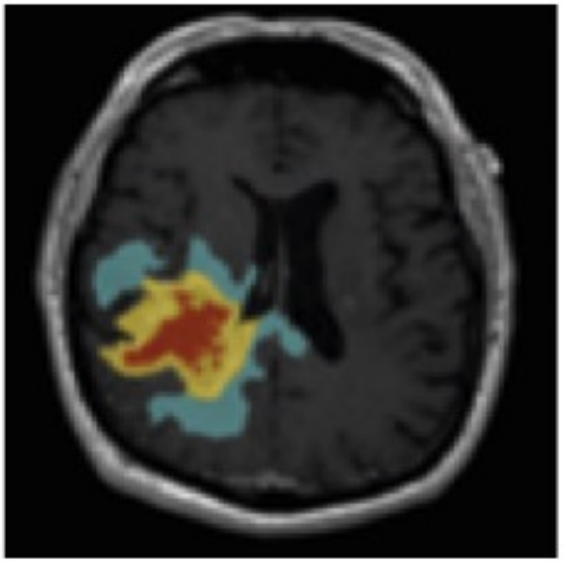
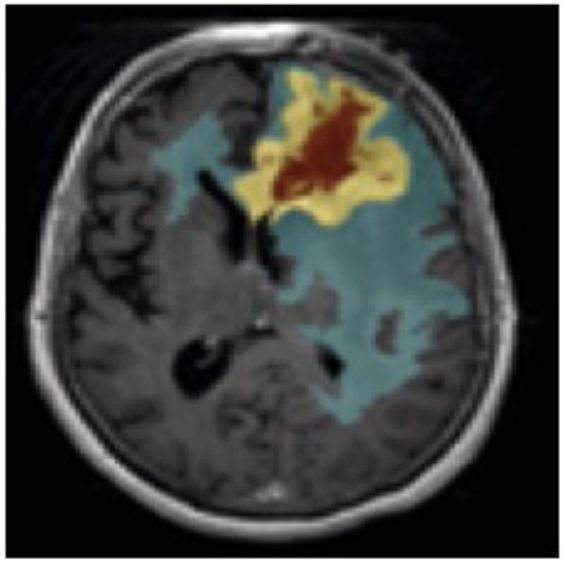
Radiogenomics

- "Radiomics": Extraction of large numbers of image features from radiological data (CT/MRI etc); radio-phenotypes
- "Radiogenomics": Radiomics + genomics
- Imaging is non-invasive; virtual biopsy
- Goal: find diagnostic/prognostic/predictive imaging biomarkers that are genomically driven and by what mechanism

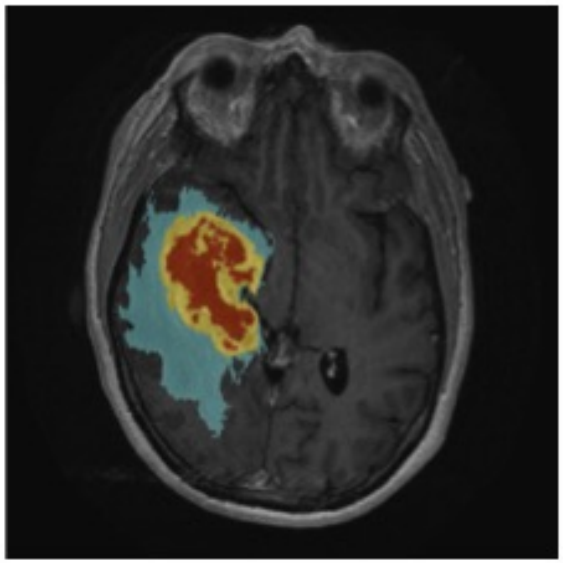
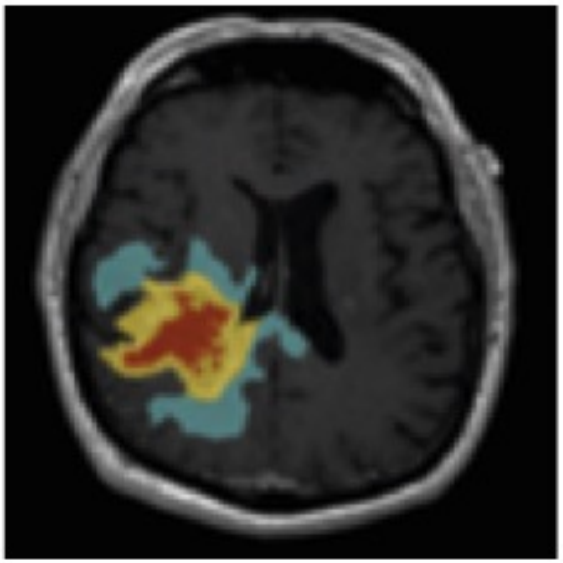
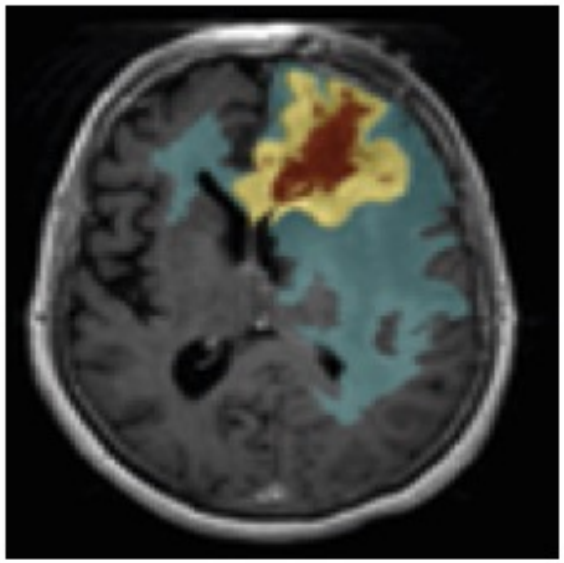
Radiogenomics

- "Radiomics": Extraction of large numbers of image features from radiological data (CT/MRI etc); radio-phenotypes
- "Radiogenomics": Radiomics + genomics
- Imaging is non-invasive; virtual biopsy
- Goal: find diagnostic/prognostic/predictive imaging biomarkers that are genomically driven and by what mechanism
- Lurking challenges
  - Tumors differ in shapes/size/areas/organs; non-conformable objects for population level analyses
  - Tumor heterogeneity at multiple levels (genomic+imaging)

Glioblastoma Multiforme

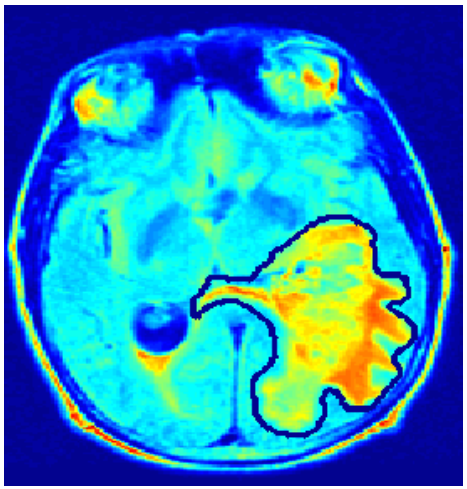




Glioblastoma Multiforme

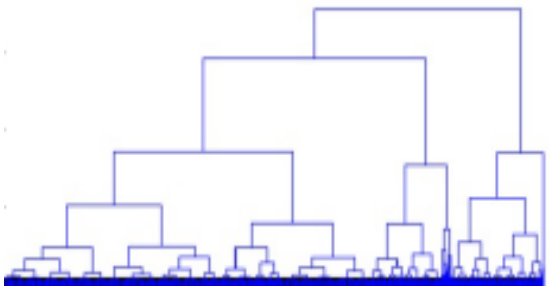




Need better metrics of tumor heterogeneity; capture different "architecture" of the tumor development

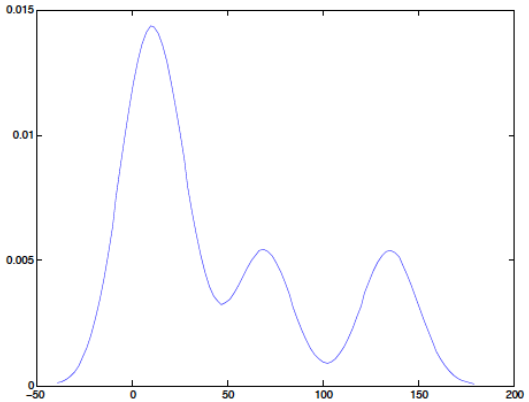
Imaging-based Metrics of (intra) Tumor Heterogeneity



Tumor heterogeneity tree

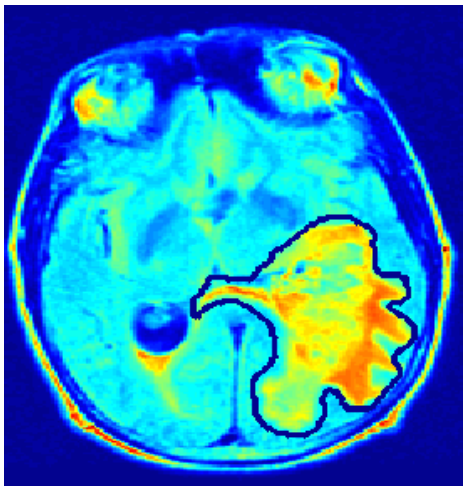


Voxel density estimate

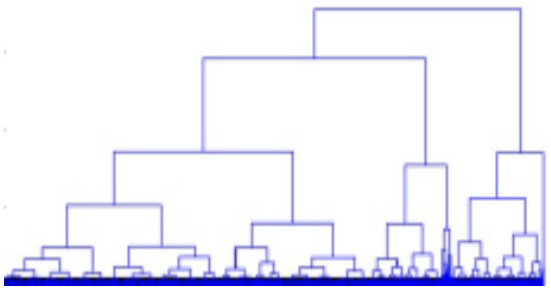


Tumor shape

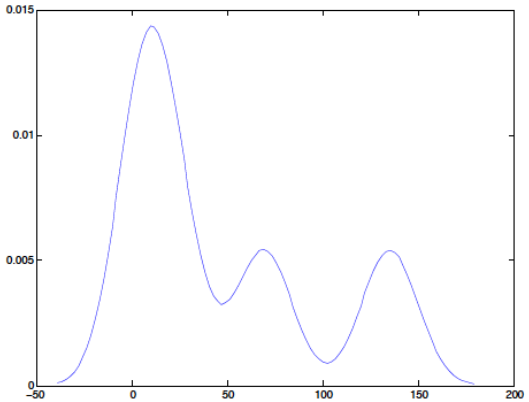
Imaging-based Metrics of (intra) Tumor Heterogeneity



Tumor heterogeneity tree



Voxel density estimate

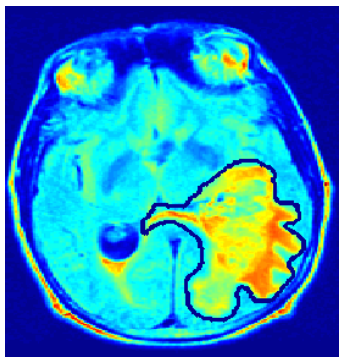


Tumor shape

"Geometric" Functional Data

(with K. Bharath, S. Kurtek, A. Rao)

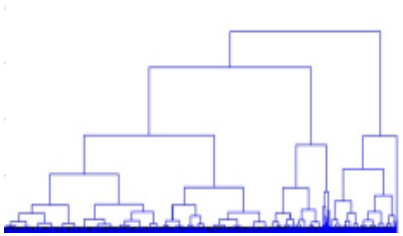
Today's talk

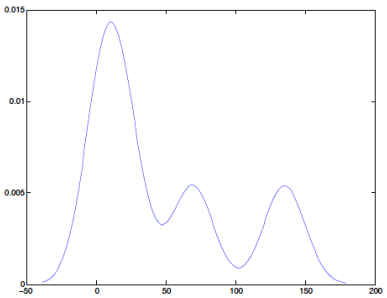


Tumor heterogeneity tree

Voxel density estimate

Tumor shape





Statistical Analyses of Tree Structured data (Bharath et al; JASA, 2017)

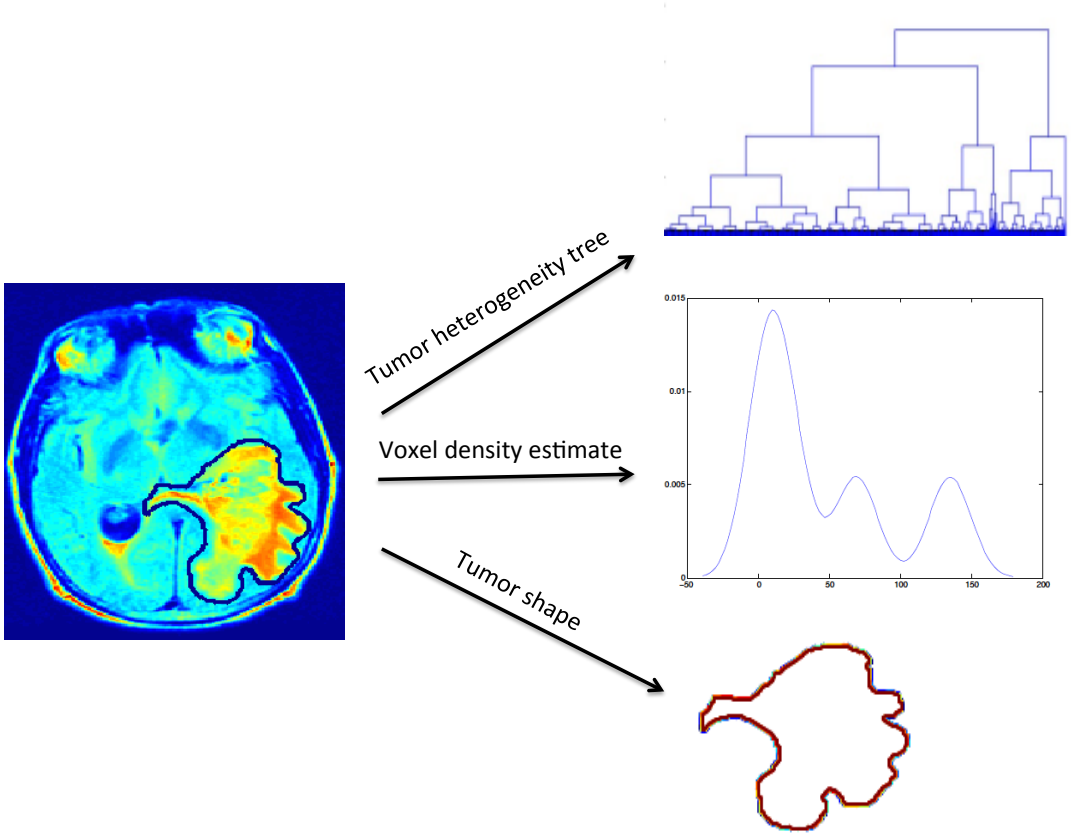
Quantile Functional Regression using Quantlets (Yang et al, JASA, under revision)

Non-parametric clustering of densities (Saha et al; Neuroimage, 2016)

Radiologic Image-based Statistical Shape Analysis of Brain Tumors (Bharath, Kurtek et al; JRSSC, 2018+)

"Geometric" Functional Data

Density-based Characterizations



Quantile Functional Regression using Quantlets (Yang et al, JASA, under revision)

Non-parametric clustering of densities (Saha et al; Neuroimage, 2016)

Glioblastoma Multiforme (GBM)

- Most common and aggressive form of brain cancer
- No current prevention approaches, and poor outcomes
  - Median survival 12mo, 3-5% 5yr survival
- Exhibits heterogeneous physiological and morphological features as it proliferates
- Investigating these heterogeneities and relating them to clinical/genetic outcomes can lead to the development of personalized treatment strategies.

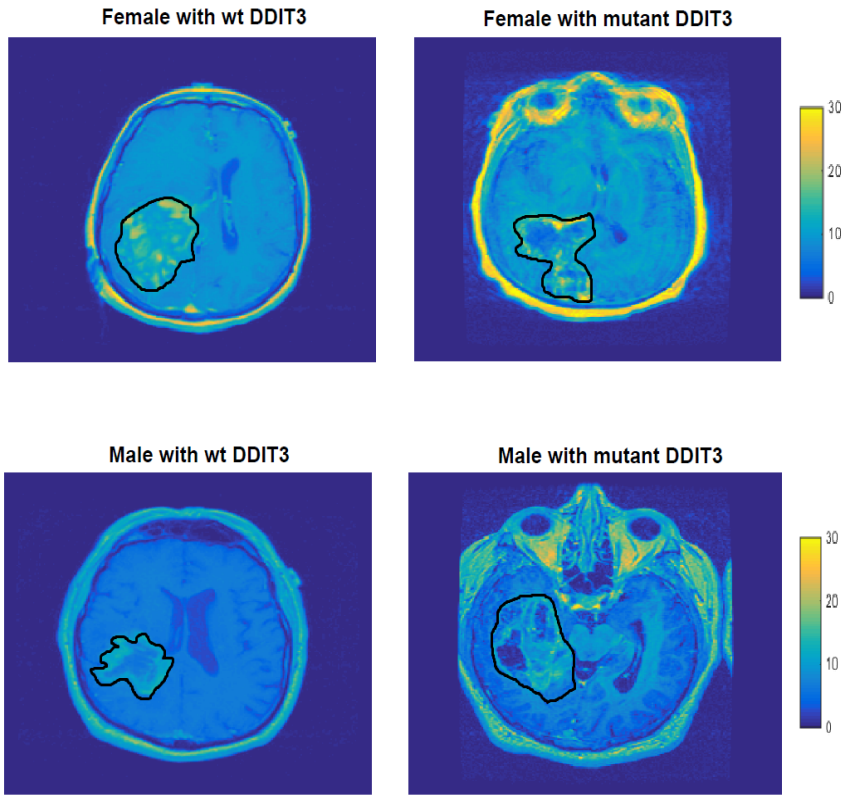
Glioblastoma Multiforme (GBM)

- Most common and aggressive form of brain cancer
- No current prevention approaches, and poor outcomes
  - Median survival 12mo, 3-5% 5yr survival
- Exhibits heterogeneous physiological and morphological features as it proliferates
- Investigating these heterogeneities and relating them to clinical/genetic outcomes can lead to the development of personalized treatment strategies.

Our Goal:

Assess how variability in tumor image intensities is associated with demographic, clinical, and genetic factors

Glioblastoma Images



- Presurgical T1-weighted post-contrast MRI images from GBM patients
- **Radiomics**: compute features summarizing tumor image characteristics and relate to clinical outcomes (100s of different features)
- *Histogram features*: Summaries computed from pixel intensity distributions (e.g. mean, variance, skewness, Q05, Q95)

Modeling Distributions

The typical approach is to fit separate regression analyses to each radiomic feature, which has some major drawbacks:

- Multiple testing problems
- May miss distributional differences not contained in pre-chosen summaries.

Modeling Distributions

The typical approach is to fit separate regression analyses to each radiomic feature, which has some major drawbacks:

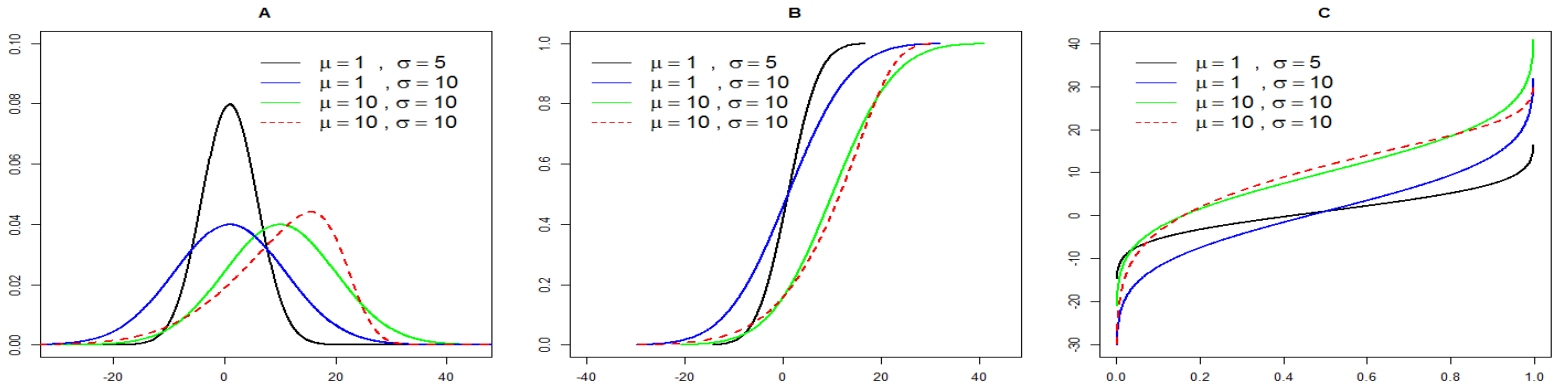
- Multiple testing problems
- May miss distributional differences not contained in pre-chosen summaries.

Alternative Approach

Instead of just modeling the extracted summaries, model the entire distribution of pixel intensities (as functional data).

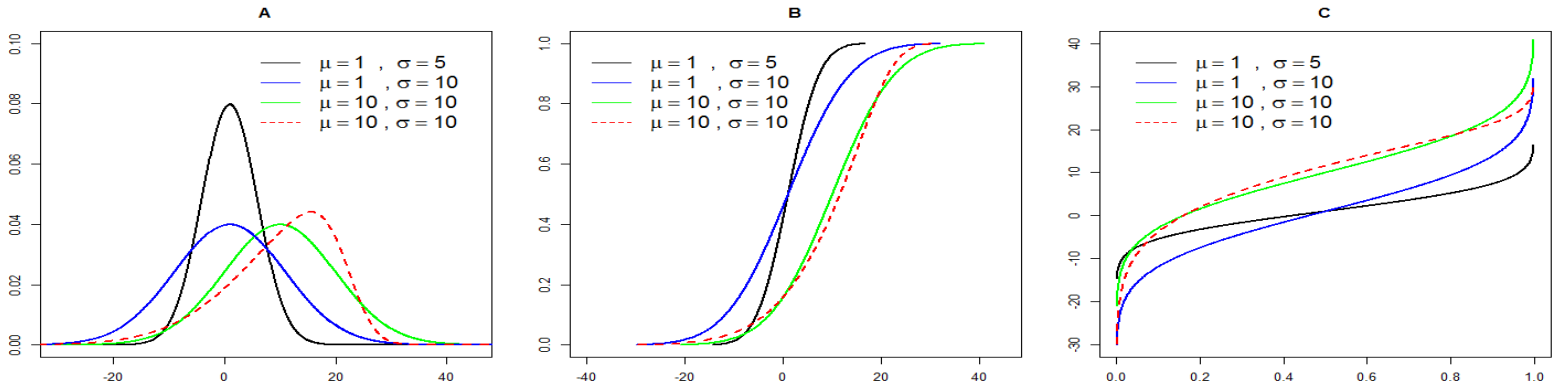
Modeling Distributions

- Various choices to represent pixel intensity distributions: density, cumulative distribution, or quantile functions.



Modeling Distributions

- Various choices to represent pixel intensity distributions: density, cumulative distribution, or quantile functions.



- We choose to use the quantile function.
  The quantile function of Y on $p \in (0, 1)$, is defined as

Definition of the quantile function

$$Q_Y(p) = F_Y^{-1}(p) = \inf(y : F_Y(y) \geq p),$$

where $p = F_Y(y)$ is the proportion less than or equal to y.

Properties of Quantile Functions

Quantile functions have properties that make them useful here:

- Defined on a fixed domain, $p \in \mathcal{P} = (0, 1)$

Properties of Quantile Functions

Quantile functions have properties that make them useful here:

- Defined on a fixed domain, $p \in \mathcal{P} = (0, 1)$
- Straightforward to compute empirical estimates without choice of smoothing parameters

eDF

Let $Y_{(1)} \leq \cdots \leq Y_{(m)}$ be order statistics from a sample of size m. For $p \in [1/(m+1), m/(m+1)]$, the eQF is given by

$$\widehat{Q}_Y(p) = (1 - w) Y_{([(m+1)p])} + w Y_{([(m+1)p]+1)},$$

where w is a weight such that $(m+1)p = [(m+1)p] + w$.

Properties of Quantile Functions

Quantile functions have properties that make them useful here:

- Defined on a fixed domain, $p \in \mathcal{P} = (0, 1)$
- Straightforward to compute empirical estimates without choice of smoothing parameters
- Straightforward formulas to calculate distributional moments

Distributional Moments

$$\begin{aligned} \mu_Y &= \mathrm{E}(Y) = \int_0^1 Q_Y(p)dp \\ \sigma_Y^2 &= \mathrm{Var}(Y) = \int_0^1 (Q_Y(p) - \mu_Y)^2 \, dp \\ \xi_Y &= \mathrm{Skew}(Y) = \int_0^1 (Q_Y(p) - \mu_Y)^3 / \sigma_Y^3 dp \end{aligned}$$

Quantile functional regression

Approach: Regress eQF as functional response on covariates.

Quantile functional regression

Approach: Regress eQF as functional response on covariates.

1. For each subject $i = 1, \ldots, n$, construct the eQF $Q_i(p)$ from the order statistics of $Y_{ij}, j = 1, \ldots, m_i$.

Quantile functional regression

Approach: Regress eQF as functional response on covariates.

1. For each subject $i = 1, \ldots, n$, construct the eQF $Q_i(p)$ from the order statistics of $Y_{ij}, j = 1, \ldots, m_i$.
2. Regress $Q_i(p)$ on covariates $x_{ia}, a = 1, \ldots, A$, each with regression coefficients $\beta_a(p)$ defined on $p \in \mathcal{P} = (0, 1)$.

Quantile Functional Regression Model

$$Q_i(p) = \beta_0(p) + \sum_{a=1}^{A} x_{ia}\beta_a(p) + E_i(p)$$

Quantile functional regression

Approach: Regress eQF as functional response on covariates.

1. For each subject $i = 1, \ldots, n$, construct the eQF $Q_i(p)$ from the order statistics of $Y_{ij}, j = 1, \ldots, m_i$.
2. Regress $Q_i(p)$ on covariates $x_{ia}, a = 1, \ldots, A$, each with regression coefficients $\beta_a(p)$ defined on $p \in \mathcal{P} = (0, 1)$.

Quantile Functional Regression Model

$$Q_i(p) = \beta_0(p) + \sum_{a=1}^{A} x_{ia}\beta_a(p) + E_i(p)$$

3. Test for significantly associated covariates: $H_0 : \beta_a(p) \equiv 0$.

Quantile functional regression

Approach: Regress eQF as functional response on covariates.

1. For each subject $i = 1, \ldots, n$, construct the eQF $Q_i(p)$ from the order statistics of $Y_{ij}, j = 1, \ldots, m_i$.
2. Regress $Q_i(p)$ on covariates $x_{ia}, a = 1, \ldots, A$, each with regression coefficients $\beta_a(p)$ defined on $p \in \mathcal{P} = (0, 1)$.

Quantile Functional Regression Model

$$Q_i(p) = \beta_0(p) + \sum_{a=1}^{A} x_{ia}\beta_a(p) + E_i(p)$$

3. Test for significantly associated covariates: $H_0 : \beta_a(p) \equiv 0$.
4. Key point: can characterize the significant distributional differences e.g. range of p, mean, variance, skewness, "Gaussianness"

Types of Quantile and Functional Regression

Response $(\cdot)$	**Objective function** $E((\cdot)\|X)$	**Objective function** $F^{-1}_{(\cdot)}(p\|X)$
scalar Y	classic regression	quantile regression
function $Y(t)$	functional regression	functional quantile regression
quantile function $F^{-1}(p)$	quantile functional regression*	quantile functional quantile regression

Types of Quantile and Functional Regression

Response $(\cdot)$	**Objective function** $E((\cdot)\|X)$	**Objective function** $F^{-1}_{(\cdot)}(p\|X)$
scalar Y	classic regression	quantile regression
function $Y(t)$	functional regression	functional quantile regression
quantile function $F^{-1}(p)$	quantile functional regression*	quantile functional quantile regression

- **Classic regression**: $E(Y|X)$

Types of Quantile and Functional Regression

Response $(\cdot)$	**Objective function** $E((\cdot)\|X)$	**Objective function** $F^{-1}_{(\cdot)}(p\|X)$
scalar Y	classic regression	quantile regression
function $Y(t)$	functional regression	functional quantile regression
quantile function $F^{-1}(p)$	quantile functional regression*	quantile functional quantile regression

- **Classic regression**: $\mathrm{E}(Y|X)$
- **Quantile regression**: $F_Y^{-1}(p|X)$
  e.g. He and Liang 2000; Koenker 2005

Types of Quantile and Functional Regression

Response $(\cdot)$	**Objective function** $E((\cdot)\|X)$	**Objective function** $F^{-1}_{(\cdot)}(p\|X)$
scalar Y	classic regression	quantile regression
function $Y(t)$	functional regression	functional quantile regression
quantile function $F^{-1}(p)$	quantile functional regression*	quantile functional quantile regression

- **Classic regression**: $\mathrm{E}(Y|X)$
- **Quantile regression**: $F_Y^{-1}(p|X)$
  e.g. He and Liang 2000; Koenker 2005
- **Functional regression**: $\mathrm{E}\{Y(t)|X\}$
  See review article by Morris (2015)

Types of Quantile and Functional Regression

Response $(\cdot)$	**Objective function** $E((\cdot)\|X)$	**Objective function** $F^{-1}_{(\cdot)}(p\|X)$
scalar Y	classic regression	quantile regression
function $Y(t)$	functional regression	functional quantile regression
quantile function $F^{-1}(p)$	quantile functional regression*	quantile functional quantile regression

- **Classic regression**: $\mathrm{E}(Y|X)$
- **Quantile regression**: $F_Y^{-1}(p|X)$
  e.g. He and Liang 2000; Koenker 2005
- **Functional regression**: $\mathrm{E}\{Y(t)|X\}$
  See review article by Morris (2015)
- **Functional quantile regression**: $F^{-1}_{Y(t)}(p|X)$
  e.g. Brockhaus et al. (2015), Liu, Li, Morris (2017)

Types of Quantile and Functional Regression

Response $(\cdot)$	**Objective function** $E((\cdot)\|X)$	**Objective function** $F^{-1}_{(\cdot)}(p\|X)$
scalar Y	classic regression	quantile regression
function $Y(t)$	functional regression	functional quantile regression
quantile function $F^{-1}(p)$	quantile functional regression*	quantile functional quantile regression

- **Classic regression**: $\mathrm{E}(Y|X)$
- **Quantile regression**: $F_Y^{-1}(p|X)$
  e.g. He and Liang 2000; Koenker 2005
- **Functional regression**: $\mathrm{E}\{Y(t)|X\}$
  See review article by Morris (2015)
- **Functional quantile regression**: $F^{-1}_{Y(t)}(p|X)$
  e.g. Brockhaus et al. (2015), Liu, Li, Morris (2017)
- **Quantile functional regression**: $\mathrm{E}\{F_Y^{-1}(p)|X\}$
  Expected quantile function given covariates *our focus*

Quantile Functional Regression

Quantile Functional Regression Model

$$Q_i(p) = \beta_0(p) + \sum_{a=1}^{A} x_{ia}\beta_a(p) + E_i(p)$$

Naive approach: compute independent regressions for each p

- fail to borrow strength over $p \rightarrow$ wiggly, inefficient $\widehat{\beta}_a(p)$.
- ignore correlation over p in $E_i(p) \rightarrow$ loss of inferential power.

Quantile Functional Regression

Quantile Functional Regression Model

$$Q_i(p) = \beta_0(p) + \sum_{a=1}^{A} x_{ia}\beta_a(p) + E_i(p)$$

Naive approach: compute independent regressions for each p

- fail to borrow strength over $p \rightarrow$ wiggly, inefficient $\widehat{\beta}_a(p)$.
- ignore correlation over p in $E_i(p) \rightarrow$ loss of inferential power.

Functional regression approach: Use *basis functions* representations to account for correlation (across p)

- $\beta_a(p)$ regularized via L1/L2 penalization of basis coefficients.
- Basis functions induce correlation across p in Cov$\{E_i(p)\}$.
- Common bases: splines, PC, Fourier bases, wavelets

Quantile Functional Regression

Quantile Functional Regression Model

$$Q_i(p) = \beta_0(p) + \sum_{a=1}^{A} x_{ia}\beta_a(p) + E_i(p)$$

Naive approach: compute independent regressions for each p

- fail to borrow strength over $p \rightarrow$ wiggly, inefficient $\widehat{\beta}_a(p)$.
- ignore correlation over p in $E_i(p) \rightarrow$ loss of inferential power.

Functional regression approach: Use *basis functions* representations to account for correlation (across p)

- $\beta_a(p)$ regularized via L1/L2 penalization of basis coefficients.
- Basis functions induce correlation across p in Cov$\{E_i(p)\}$.
- Common bases: splines, PC, Fourier bases, wavelets
- Doesn't work for quantile functions!

Quantile Functional Regression

Quantile Functional Regression Model

$$Q_i(p) = \beta_0(p) + \sum_{a=1}^{A} x_{ia}\beta_a(p) + E_i(p)$$

Naive approach: compute independent regressions for each p

- fail to borrow strength over $p \rightarrow$ wiggly, inefficient $\widehat{\beta}_a(p)$.
- ignore correlation over p in $E_i(p) \rightarrow$ loss of inferential power.

Functional regression approach: Use *basis functions* representations to account for correlation (across p)

- $\beta_a(p)$ regularized via L1/L2 penalization of basis coefficients.
- Basis functions induce correlation across p in Cov$\{E_i(p)\}$.
- Common bases: splines, PC, Fourier bases, wavelets
- Doesn't work for quantile functions!

Here, we introduce new custom basis functions *quantlets*.

Construction of Quantlet Basis Functions

Multi-step process to derive custom *quantlet* basis functions:

1. Construct overcomplete dictionary

Details of Step

Gaussian bases: $\psi_0(p) = 1$ for $p \in (0,1)$, $\psi_1(p) = \Phi^{-1}(p)$.
Beta CDF bases: $\psi_k(p) = F_{\theta_k}(p)$ for $k = 2, \ldots, K_0$
Overcomplete dictionary: $\mathcal{D}^0 = \{\psi_k, k = 0, \ldots, K_0\}$

Construction of Quantlet Basis Functions

Multi-step process to derive custom *quantlet* basis functions:

1. Construct overcomplete dictionary
2. Choose sparse set of dictionary elements for each subject.

Details of Step

For each subject, use penalized regression (e.g. lasso) to find a sparse subset of dictionary elements.

$$|Q_i(p) - \sum_{k \in \mathcal{D}^0} \psi_k(p) Q_{ik}^O|_2^2 + \lambda_i \sum_{k \in \mathcal{D}^0} |Q_{ik}^O|_1$$

Obtain $\mathcal{D}_i = \{\psi_k(p) \in \mathcal{D}^0 : Q_{ik}^0 \neq 0\}$.

Construction of Quantlet Basis Functions

Multi-step process to derive custom *quantlet* basis functions:

1. Construct overcomplete dictionary
2. Choose sparse set of dictionary elements for each subject.
3. Take union set, and then find subset that is *near-lossless*.

Details of Step

Union set: $\mathcal{D}^U = \cup_{i=1}^n \mathcal{D}_i$
Cardinality $\mathcal{C}$ set: $\mathcal{D}^{\mathcal{C}} = \{\psi_k(p), k : \sum_{i=1}^n I(Q_{ik}^0 \neq 0) \geq \mathcal{C}\}$
Lossless measure: Cross-validated concordance coefficient:

$$\rho_i^{\mathcal{C}} = \text{Concordance}\{Q_i(p), \widehat{Q}_i^{\mathcal{C}}(p)\} \in (0,1)$$

Plot $\rho_0^{\mathcal{C}} = \min_i\{\rho_i^{\mathcal{C}}\}$ vs. $\mathcal{C}$ and choose $\mathcal{C} : \rho_0^{\mathcal{C}} < \epsilon$
Near-lossless set: $\mathcal{D}^{\epsilon} = \{\mathcal{D}^{\mathcal{C}} \text{ with } \mathcal{C} = \min(\mathcal{C} : \rho_0^{\mathcal{C}} < \epsilon)\}$

Construction of Quantlet Basis Functions

Multi-step process to derive custom *quantlet* basis functions:

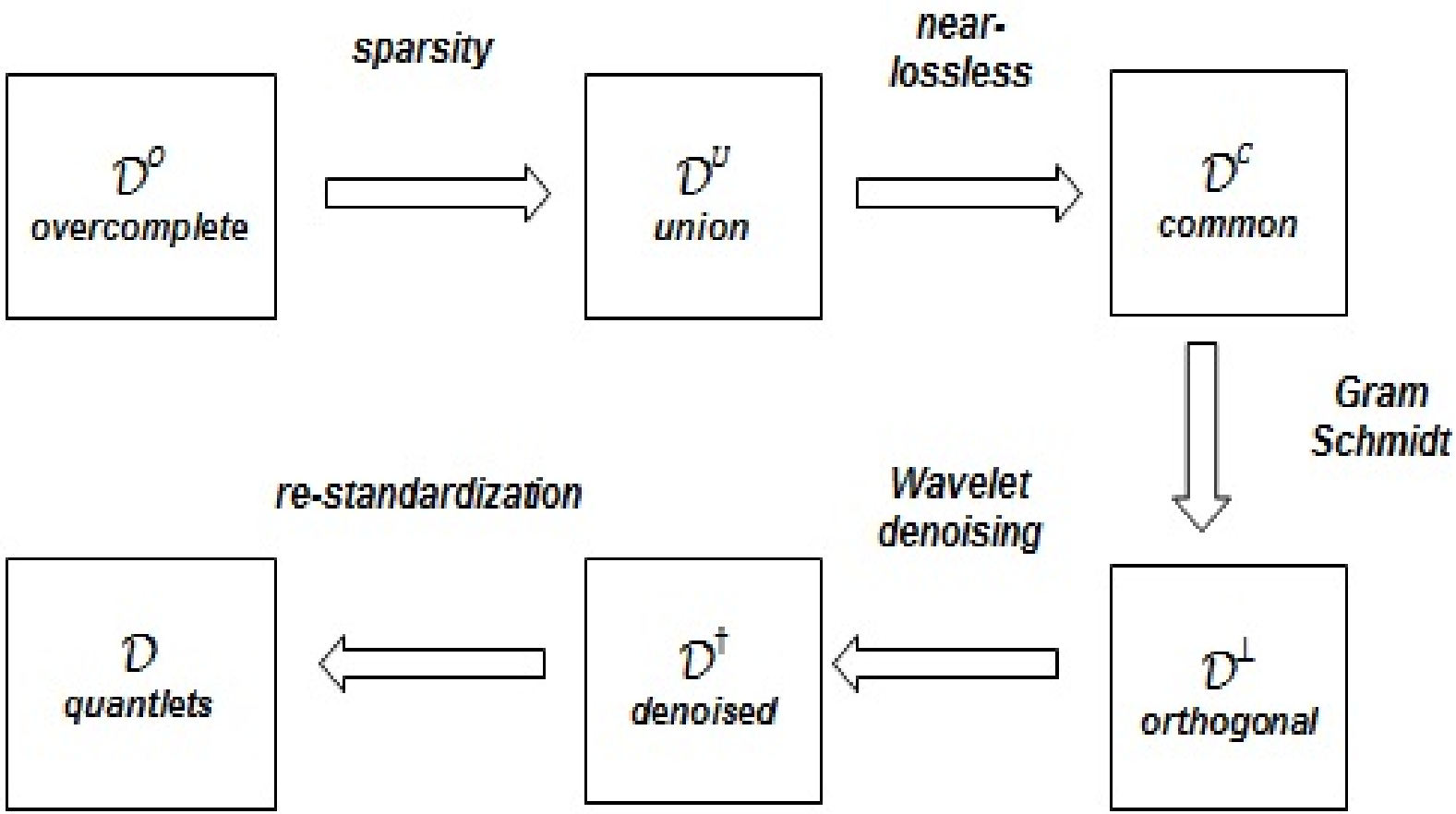
1. Construct overcomplete dictionary
2. Choose sparse set of dictionary elements for each subject.
3. Take union set, and then find subset that is *near-lossless*.
4. Orthogonalize this subset, regularize, and re-standardize.

Details of Step

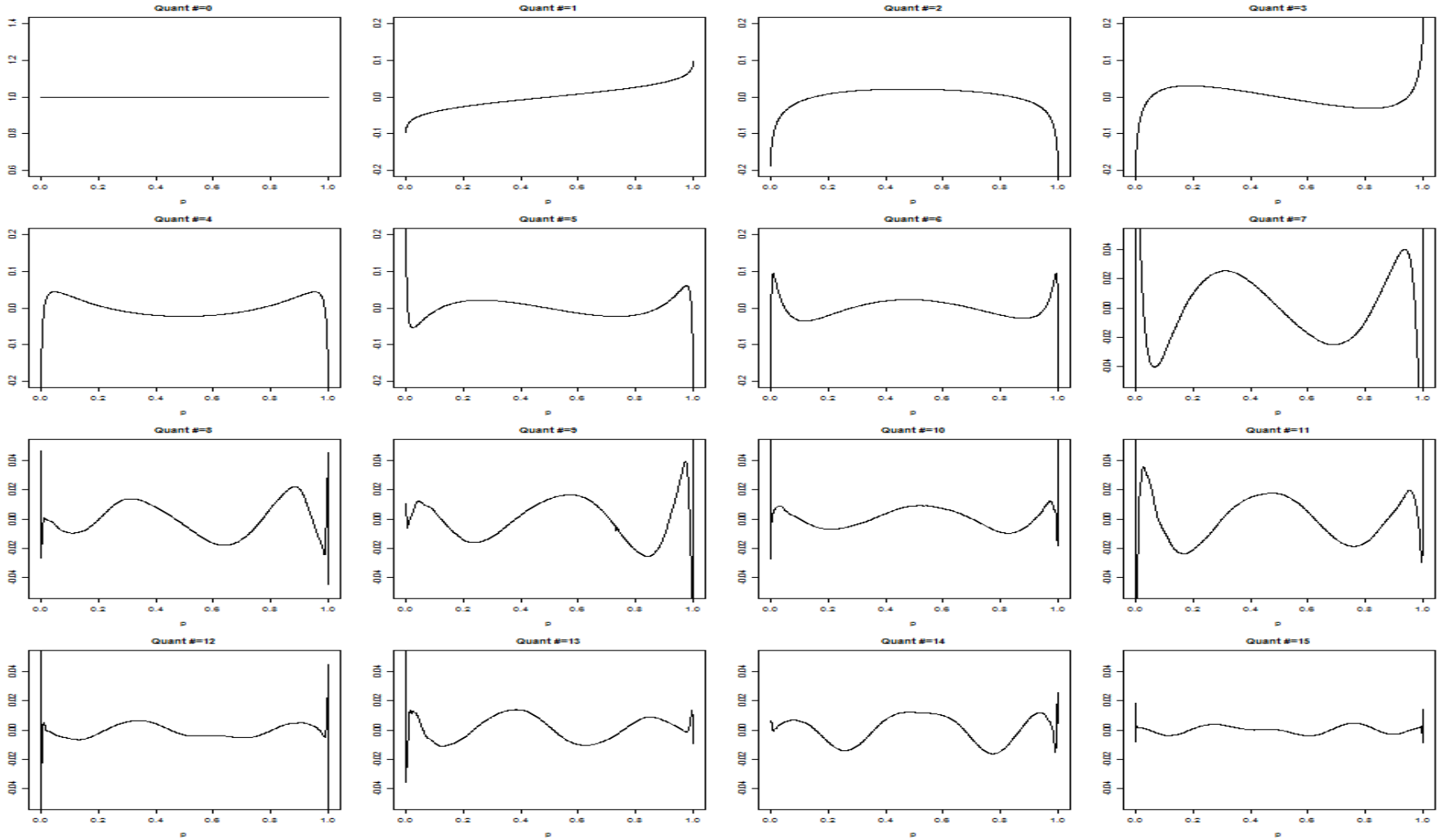
Orthogonal set: $\mathcal{D}^{\perp} = \{\psi_k^{\perp}, k = 0, \ldots, K\} = \text{Gram-Schmidt}(\mathcal{D}^{\epsilon})$
Regularize $\psi^{\perp}$ via wavelet denoising and then renormalize.

Resulting bases are called *quantlets:* $\mathcal{D} = \{\xi_k(p), k = 0, \ldots, K\}$

Construction of Quantlets



First 16 Quantlets for GBM Data



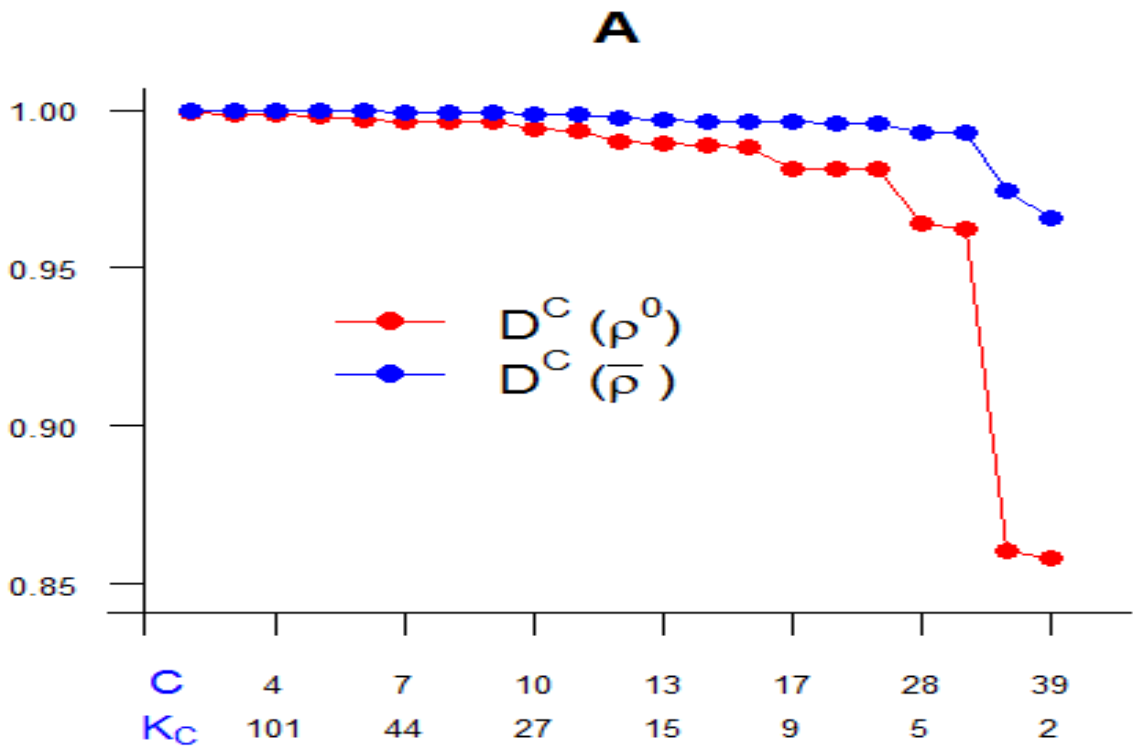
First two are Gaussian quantiles: mean & variance

Properties of Quantlets

- **Empirically defined:** adapts to characteristics of given data.

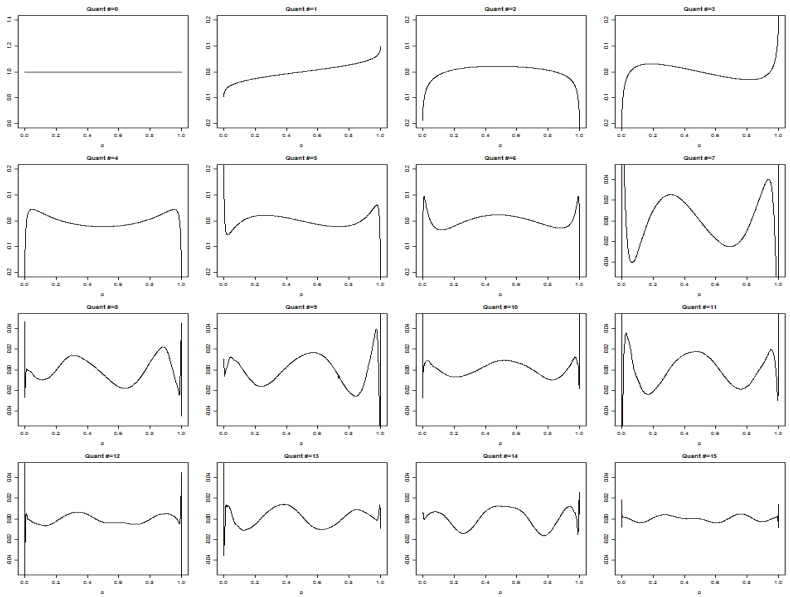
Properties of Quantlets

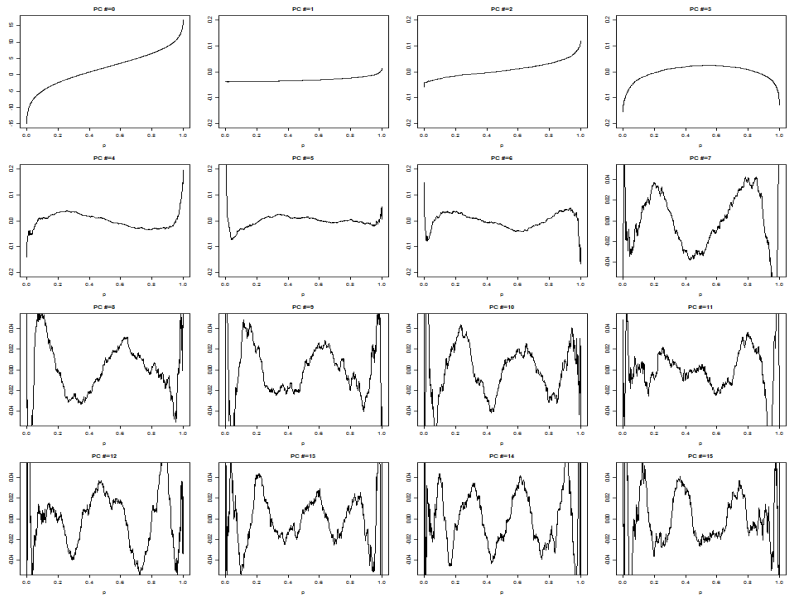
- **Empirically defined:** adapts to characteristics of given data.
- **Near-lossless:** rich enough to capture structure in each eQF.



Properties of Quantlets

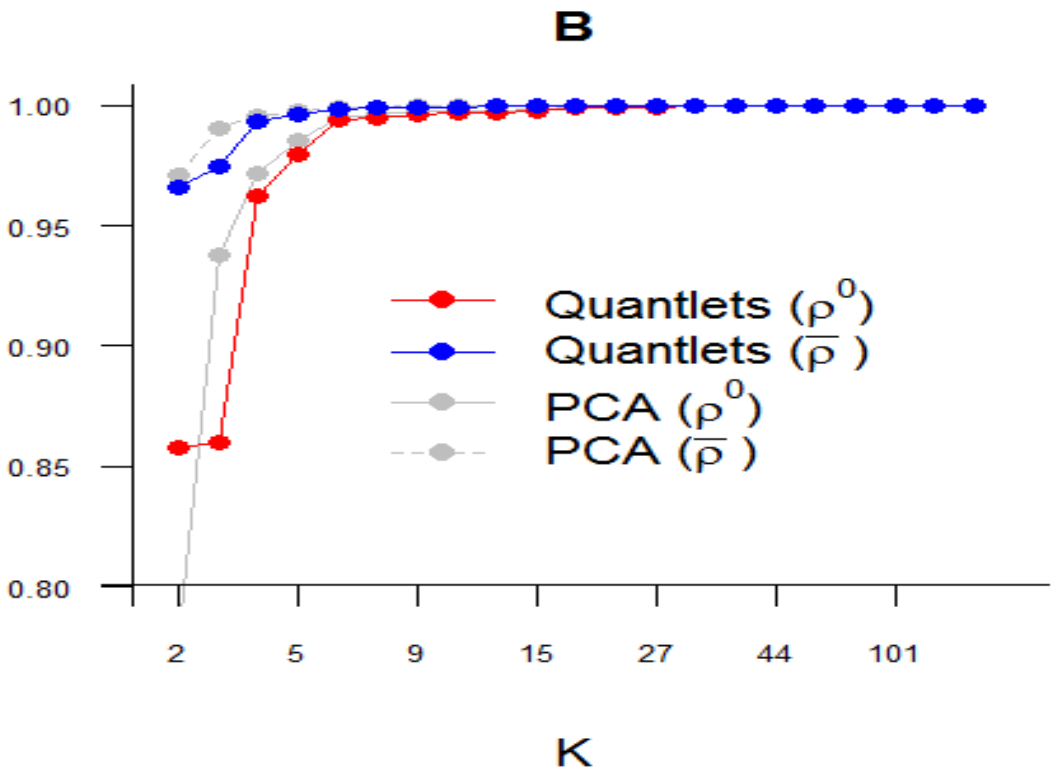
- **Empirically defined:** adapts to characteristics of given data.
- **Near-lossless:** rich enough to capture structure in each eQF.
- **Regularity:** denoising removes wiggles $\rightarrow$ smooth quantlets.





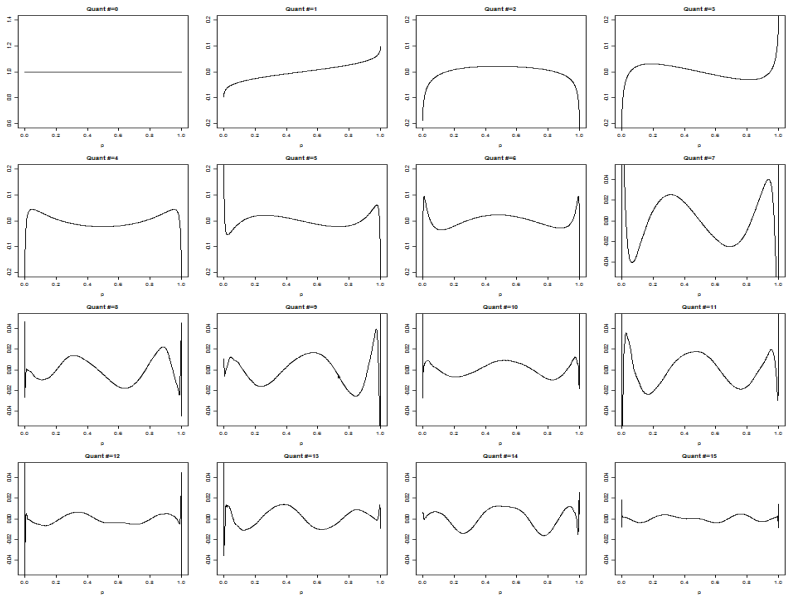
Properties of Quantlets

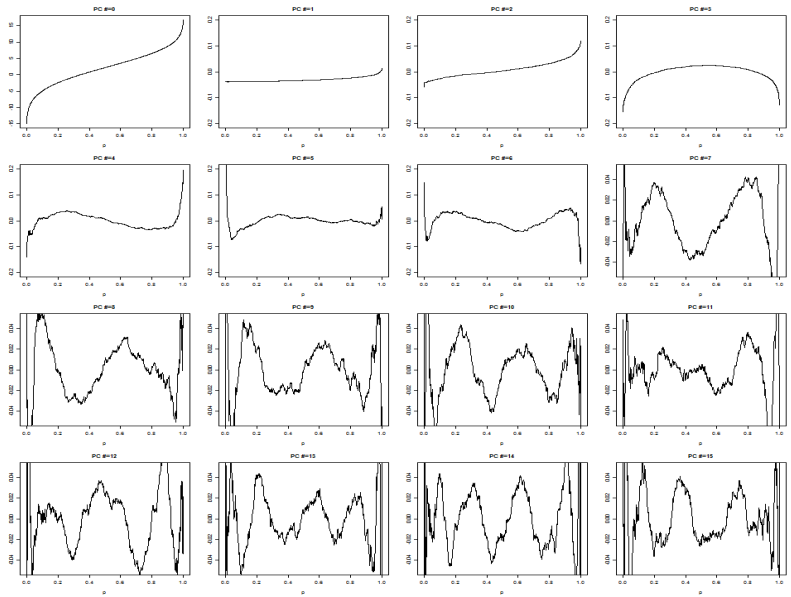
- **Empirically defined:** adapts to characteristics of given data.
- **Near-lossless:** rich enough to capture structure in each eQF.
- **Regularity:** denoising removes wiggles $\rightarrow$ smooth quantlets.
- **Sparsity:** tends to produce low dimensional basis.



Properties of Quantlets

- **Empirically defined:** adapts to characteristics of given data.
- **Near-lossless:** rich enough to capture structure in each eQF.
- **Regularity:** denoising removes wiggles $\rightarrow$ smooth quantlets.
- **Sparsity:** tends to produce low dimensional basis.
- **Interpretability:** first two bases measure Gaussianity





Basis Transform Modeling Approach

Data Space Model

$$Q_i(p) = X_i^T B(p) + E_i(p),$$

where $B(p) = (\beta_1(p), \ldots, \beta_A(p))^T$ and $E_i(p)$ is a noise process.

1. Compute quantlet basis coefficients

Computing Quantlet Coefficients

Let $\boldsymbol{Q_i} = [Q_i(p_1), \ldots, Q_i(p_{m_i})]$ with $p_j = j/(m_i + 1)$
Let $\boldsymbol{\Psi_i}$ be $K \times m_i$ matrix with elements $\boldsymbol{\psi_i}(k, j) = \psi_k(p_j)$
Quantlet coefficients: $\boldsymbol{Q_i^* = Q_i \Psi_i^-}$ where $\boldsymbol{\Psi_i^- = \Psi_i^T (\Psi_i \Psi_i^T)^{-1}}$.

Basis Transform Modeling Approach

Data Space Model

$$Q_i(p) = X_i^T B(p) + E_i(p),$$

where $B(p) = (\beta_1(p), \ldots, \beta_A(p))^T$ and $E_i(p)$ is a noise process.

1. Compute quantlet basis coefficients
2. Fit quantlet space model

Quantlet Space Model

$$\boldsymbol{Q}^* = \boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E}^*$$

where $Q_i(p_j) = \sum_{k=1}^{K} Q_{ik}^* \psi_k(p_j)$ and $\beta_a(p) = \sum_{k=1}^{K} B_{ak}^* \psi_k(p)$,
$E_i(p) = \sum_{k=1}^{K} E_{ik}^* \psi_k(p)$, and $\{p_1, \ldots, p_J\} \in (0, 1)$.
$\boldsymbol{E_i^*} \sim \text{MVN}(0, \Sigma^*)$ where Σ^* is $K \times K$ covariance matrix.

Basis Transform Modeling Approach

Data Space Model

$$Q_i(p) = X_i^T B(p) + E_i(p),$$

where $B(p) = (\beta_1(p), \ldots, \beta_A(p))^T$ and $E_i(p)$ is a noise process.

1. Compute quantlet basis coefficients
2. Fit quantlet space model
3. Transform results back to data space for inference

Transform Results to Data Space

$\beta_a(p) = \sum_{k=1}^{K} B^*_{ak} \psi_k(p)$, and then perform desired inference.

Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.

Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.
  - Sparsity prior on B^*_{ak} to regularize $\beta_a(p)$. (spike Gaussian-slab)

Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.
  - Sparsity prior on B^*_{ak} to regularize $\beta_a(p)$. (spike Gaussian-slab)
  - Vague proper prior on covariance parameters.

Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.
  - Sparsity prior on B^*_{ak} to regularize $\beta_a(p)$. (spike Gaussian-slab)
  - Vague proper prior on covariance parameters.
  - EBayes or hyperpriors on sparsity hyperparameters.

Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.
  - Sparsity prior on B^*_{ak} to regularize $\beta_a(p)$. (spike Gaussian-slab)
  - Vague proper prior on covariance parameters.
  - EBayes or hyperpriors on sparsity hyperparameters.
- MCMC used to update parameters in the quantlet space model.

Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.
  - Sparsity prior on B^*_{ak} to regularize $\beta_a(p)$. (spike Gaussian-slab)
  - Vague proper prior on covariance parameters.
  - EBayes or hyperpriors on sparsity hyperparameters.
- MCMC used to update parameters in the quantlet space model.
  - Complete conditional for B^*_{ak} is mixture of δ_0 and Gaussian.

Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.
  - Sparsity prior on B^*_{ak} to regularize $\beta_a(p)$. (spike Gaussian-slab)
  - Vague proper prior on covariance parameters.
  - EBayes or hyperpriors on sparsity hyperparameters.
- MCMC used to update parameters in the quantlet space model.
  - Complete conditional for B^*_{ak} is mixture of δ_0 and Gaussian.
  - Covariance parameters have conjugate complete conditionals.

Bayesian Modeling

- We use a Bayesian modeling approach to fit this model.
  - Sparsity prior on B^*_{ak} to regularize $\beta_a(p)$. (spike Gaussian-slab)
  - Vague proper prior on covariance parameters.
  - EBayes or hyperpriors on sparsity hyperparameters.
- MCMC used to update parameters in the quantlet space model.
  - Complete conditional for B^*_{ak} is mixture of δ_0 and Gaussian.
  - Covariance parameters have conjugate complete conditionals.
- Posterior samples transformed back to original data space to get posterior samples of $\beta_a(p)$ on desired grid of p.

Recommended Sequence of Bayesian Inference

1. Construct 95% joint credible bands for each predictor.

$100(1-\alpha)\%$ Joint Credible Band (Ruppert/Wand/Carroll 2003)

$$J_{a,\alpha}(p) = \hat{\beta}_a(p) \pm q_{1-\alpha}\left[\widehat{\text{StDev}}\{\hat{\beta}_a(p)\}\right]$$

where $q_{1-\alpha}$ is $(1-\alpha)$ quantile of:

$$Z_a^{(m)} = \max_{p\in\mathcal{P}}\left|\frac{\beta_a^{(m)}(p) - \hat{\beta}_a(p)}{\widehat{\text{St.Dev}}\{\hat{\beta}_a(p)\}}\right|$$

.

Recommended Sequence of Bayesian Inference

1. Construct 95% joint credible bands for each predictor.
2. Calculate global Bayesian p-value for each predictor.

Global Bayesian P-value (Meyer et al. 2015)

To assess $H_0 : \beta_a(p) \equiv 0$, we compute:

$$P_{a,\text{Bayes}} = \min\{\alpha : 0 \notin J_{a,\alpha}(p) \text{ for some } p \in \mathcal{P}\},$$

and conclude $\beta_a(p)$ differs from 0 whenever $P_{a,\text{Bayes}} < \alpha$.

Recommended Sequence of Bayesian Inference

1. Construct 95% joint credible bands for each predictor.
2. Calculate global Bayesian p-value for each predictor.
3. For significant predictors, flag $\{p : P_{a,\text{SimBaS}} < \alpha\}$.

Simultaneous Band Scores (SimBaS, Meyer et al. 2015)

$$\begin{aligned} P_{a,\text{SimBas}}(p) &= \min\{\alpha : 0 \notin J_{a,\alpha}(p)\} \\ &= M^{-1} \sum_{m=1}^{M} I\left\{ \left| \frac{\hat{\beta}_a(p)}{\widehat{\text{StDev}}\{\hat{\beta}_a(p)\}} \right| \leq Z_a^{(m)} \right\} \end{aligned}$$

Recommended Sequence of Bayesian Inference

1. Construct 95% joint credible bands for each predictor.
2. Calculate global Bayesian p-value for each predictor.
3. For significant predictors, flag $\{p : P_{a,\text{SimBaS}} < \alpha\}$.
4. For significant predictors, assess which moments differ.

Probability scores for moments

$$\mu_{\boldsymbol{X}}^{(m)} = \int_0^1 \boldsymbol{X}^T \beta^{(m)}(p) dp$$

$$P_{\mu_1-\mu_2} = 2 * \min \left\{ M^{-1} \sum_{m=1}^{M} I\left(\mu_{\boldsymbol{X}_1}^{(m)} - \mu_{\boldsymbol{X}_2}^{(m)} > 0\right), \right.$$
$$\left. M^{-1} \sum_{m=1}^{M} I\left(\mu_{\boldsymbol{X}_1}^{(m)} - \mu_{\boldsymbol{X}_2}^{(m)} < 0\right) \right\}$$

Simulation

Figure: 4 groups: mean distributions are N(1,5), N(3,5), N(1,6.5), and skewed normal with mean 1 and variance 5.

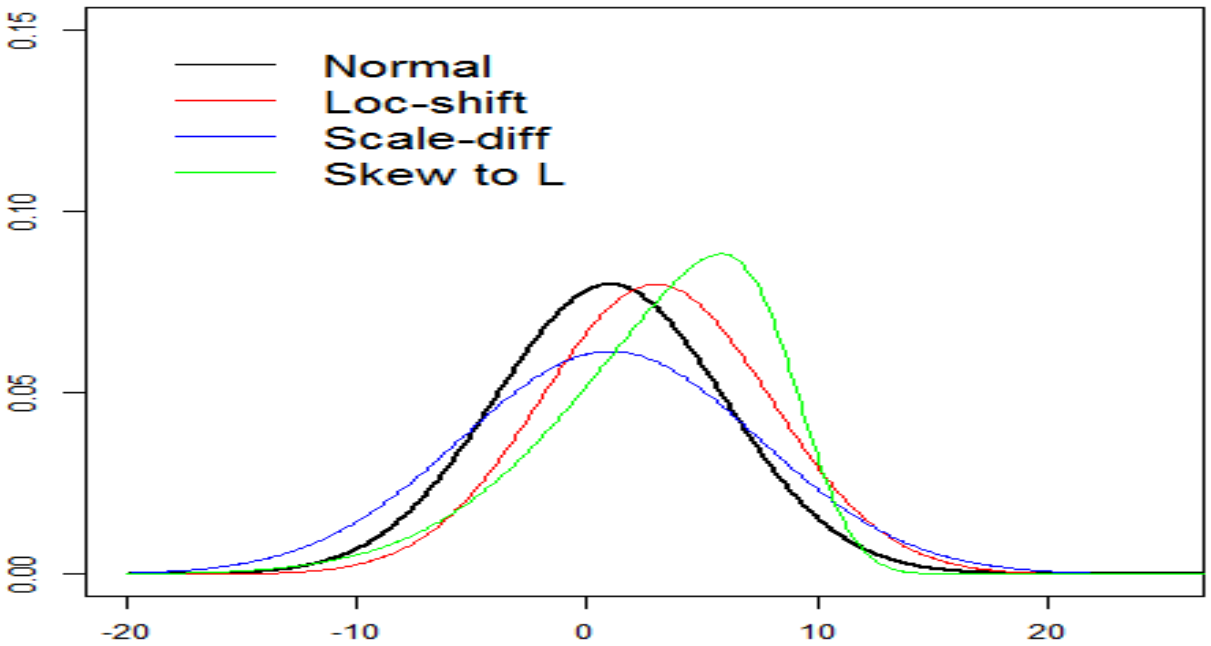
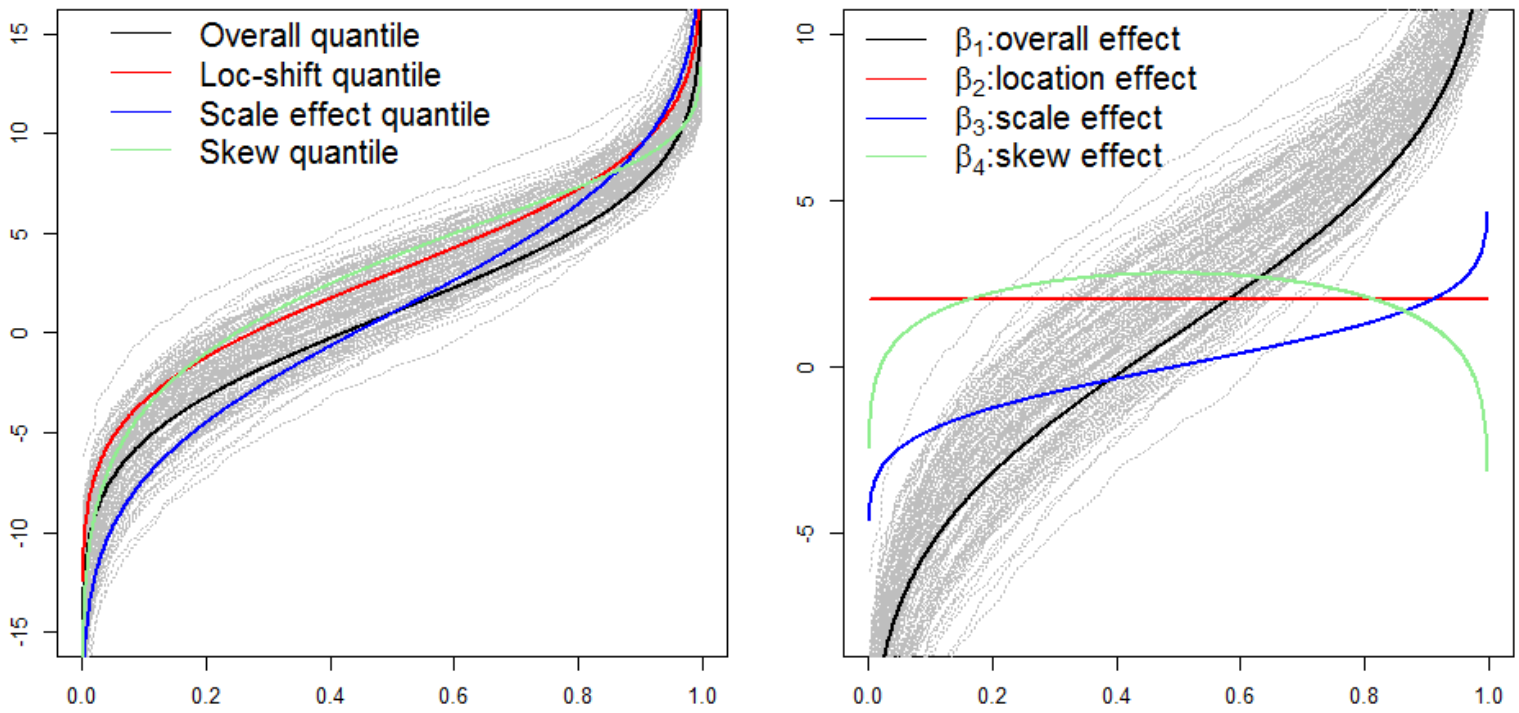


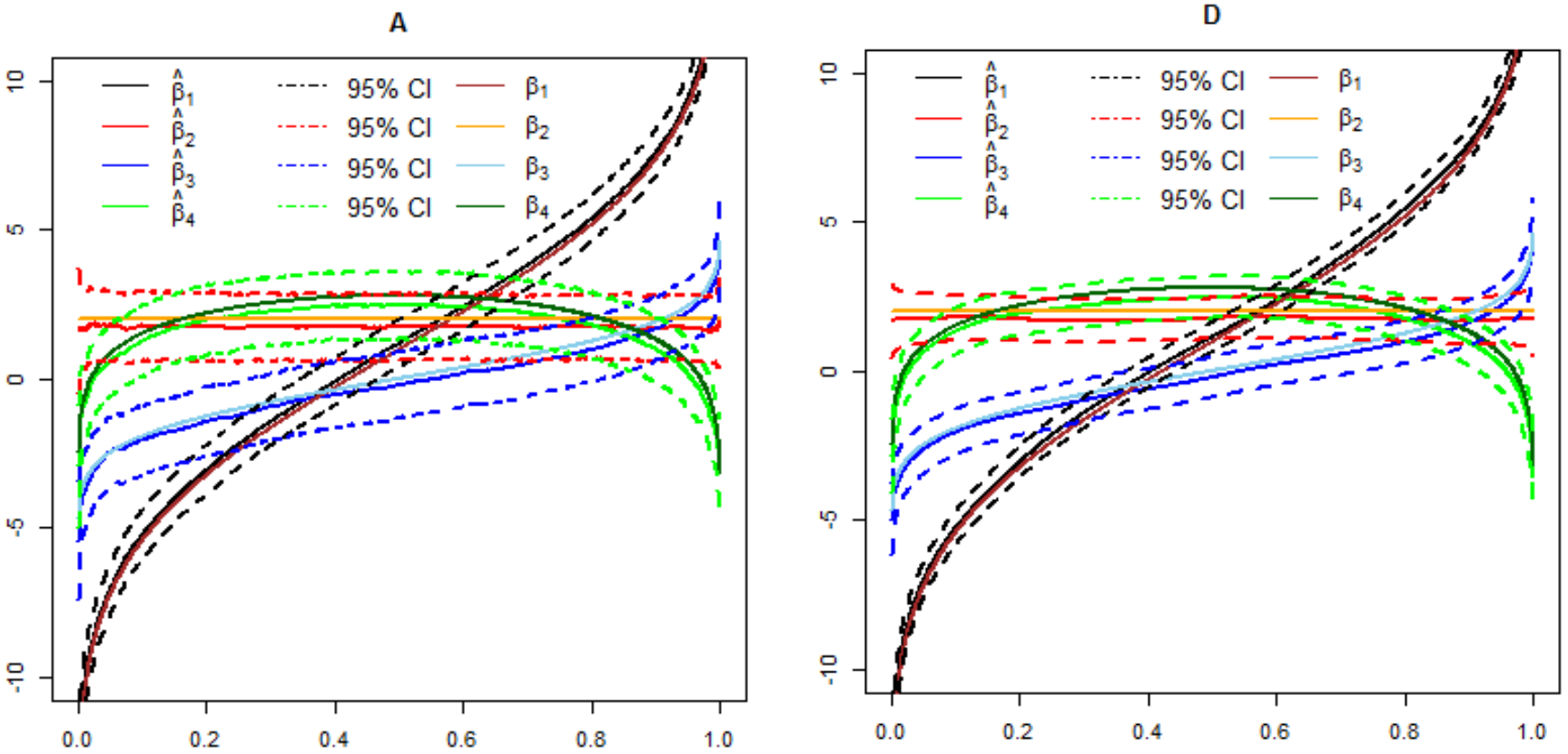
Figure: **Simulated Data**. $\beta_a(p)$ are location, scale, and skewness shifts.



- $Y_{ij}(p) = Q_{ij}(p) + \epsilon_{ij}(p)$ on $1,024$ grid points $\{p_1, \ldots, p_{1024}\}$.
- $\epsilon_{ij}(p)$ follows AR(1) process to approximate biological variability within groups.

Simulation Results

Figure: Results of the simulation: estimations and 95% joint CI (A=Naive *one-p-at-a-time* method; D=*quantlets* with regularization)



Simulation Results

Table: Area and coverage for the joint 95% credible intervals.

Type	**A** (naive)	**B** (PCA)	**C** (no reg.)	**D** (regularized)
$\beta_1(p)$	1.603 (1.000)	1.092 (0.999)	1.186 (1.000)	1.069 (1.000)
$\beta_2(p)$	2.246 (1.000)	1.551 (1.000)	1.706 (1.000)	1.465 (1.000)
$\beta_3(p)$	2.242 (1.000)	1.599 (1.000)	1.717 (1.000)	1.457 (1.000)
$\beta_4(p)$	2.281 (1.000)	1.583 (1.000)	1.651 (1.000)	1.499 (1.000)

Table: Probability scores for differences in mean, variance, and skewness.

$\mathbf{H_0}$	**True**	**A**	**B**	**C**	**D**	**E** (feature)	**F** (Gau
$\mu_1 = \mu_3$	$\mu_1 = \mu_3$	0.001	0.193	0.211	0.217	0.205	0.21
$\mu_2 = \mu_4$	$\mu_2 = \mu_4$	0.001	0.447	0.465	0.445	0.438	0.46
$\sigma_1 = \sigma_3$	$\sigma_1 \neq \sigma_3$	0.001	0.001	0.001	0.001	0.001	0.00
$\sigma_2 = \sigma_4$	$\sigma_2 = \sigma_4$	0.002	0.420	0.334	0.331	0.187	0.01
$\xi_1 = \xi_3$	$\xi_1 = \xi_3$	0.374	0.498	0.488	0.479	0.389	0.49
$\xi_2 = \xi_4$	$\xi_2 \neq \xi_4$	0.001	0.001	0.001	0.001	0.001	0.50

Simulation Results

Table: Area and coverage for the joint 95% credible intervals.

Type	**A** (naive)	**B** (PCA)	**C** (no reg.)	**D** (regularized)
$\beta_1(p)$	1.603 (1.000)	1.092 (0.999)	1.186 (1.000)	1.069 (1.000)
$\beta_2(p)$	2.246 (1.000)	1.551 (1.000)	1.706 (1.000)	1.465 (1.000)
$\beta_3(p)$	2.242 (1.000)	1.599 (1.000)	1.717 (1.000)	1.457 (1.000)
$\beta_4(p)$	2.281 (1.000)	1.583 (1.000)	1.651 (1.000)	1.499 (1.000)

Table: Probability scores for differences in mean, variance, and skewness.

$\mathbf{H_0}$	**True**	**A**	**B**	**C**	**D**	**E** (feature)	**F** (Gau
$\mu_1 = \mu_3$	$\mu_1 = \mu_3$	0.001	0.193	0.211	0.217	0.205	0.21
$\mu_2 = \mu_4$	$\mu_2 = \mu_4$	0.001	0.447	0.465	0.445	0.438	0.46
$\sigma_1 = \sigma_3$	$\sigma_1 \neq \sigma_3$	0.001	0.001	0.001	0.001	0.001	0.00
$\sigma_2 = \sigma_4$	$\sigma_2 = \sigma_4$	0.002	0.420	0.334	0.331	0.187	0.01
$\xi_1 = \xi_3$	$\xi_1 = \xi_3$	0.374	0.498	0.488	0.479	0.389	0.49
$\xi_2 = \xi_4$	$\xi_2 \neq \xi_4$	0.001	0.001	0.001	0.001	0.001	0.50

Bottomline: much better coverage and power

GBM Data Analysis

Response: T1 MRI images from 64 patients in glioblastoma (GBM) study, Y_{ij}=intensity of pixel j from subject i, $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, with m_i ranging from 371 to 3421.

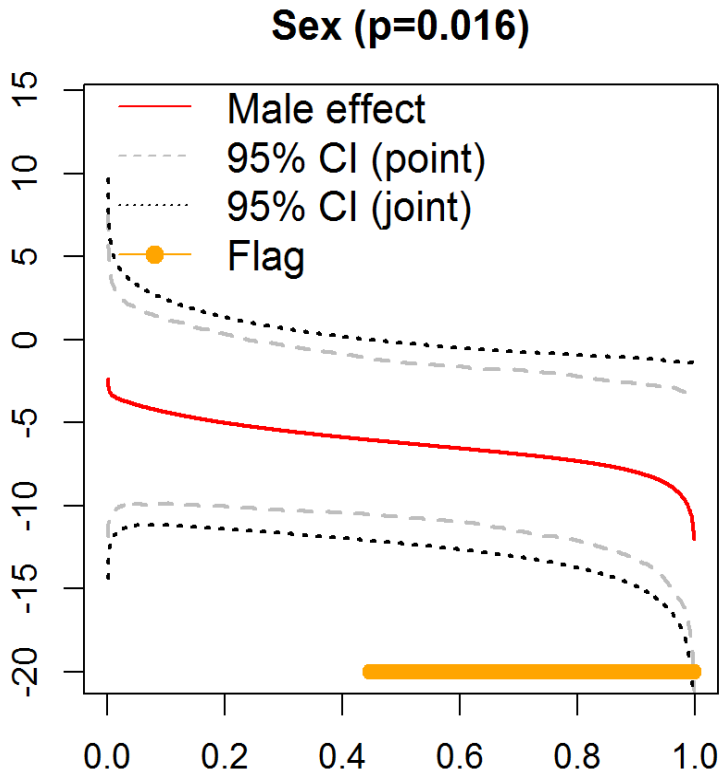
Covariates:

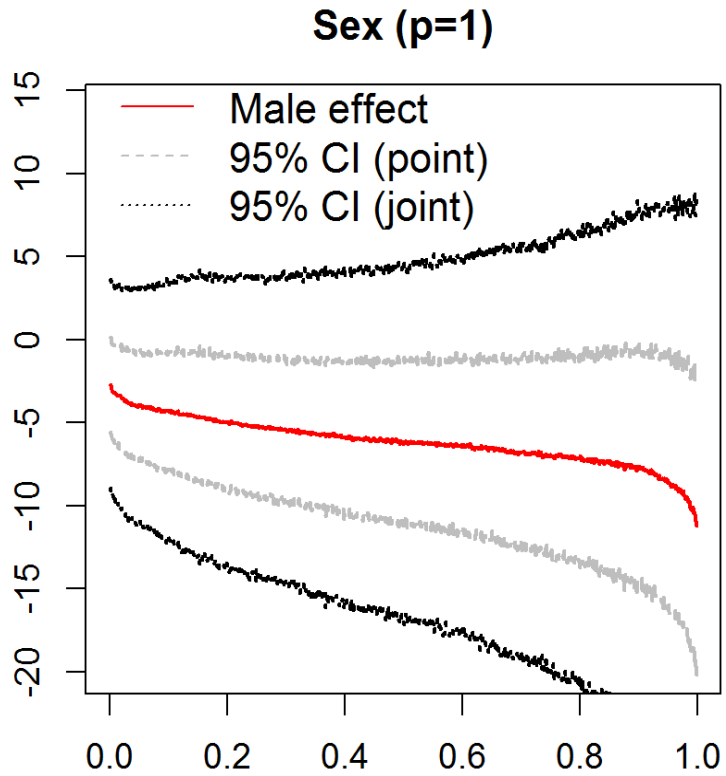
- **Demographic variables:** *sex* (21 F/43M) & *age* (56.5yr)
- **GBM subtype:** *mesenchymal* (30 mes./34 other)
- **Clinical outcome:** *survival* ($>$ 12m/$<$ 12m)
- **Genetic alterations:** *DDIT3*(6m/58wt) & *EGFR*(24m/58wt)

Model

$$\begin{aligned} Q_i(p|X_i) =& \beta_0(p) + x_{\text{sex},i}\beta_{\text{sex}}(p) + x_{\text{age},i}\beta_{\text{age}}(p) + x_{\text{surv,i}}\beta_{\text{surv}}(p) \\ &+ x_{\text{Mes},i}\beta_{\text{Mes}}(p) + x_{\text{DDIT3},i}\beta_{\text{DDIT3}}(p) \\ &+ x_{\text{EGFR},i}\beta_{\text{EGFR}}(p) + E_i(p). \end{aligned}$$

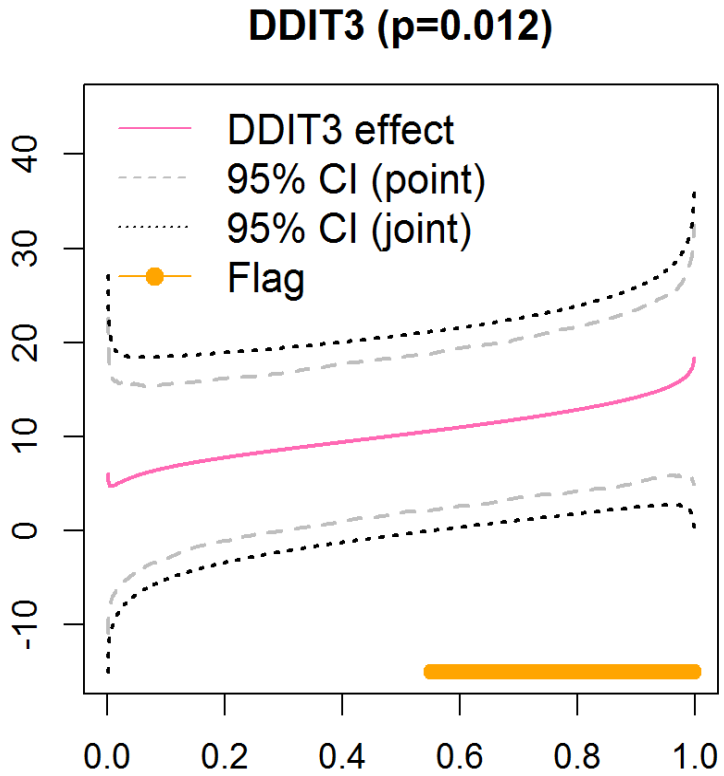
GBM Results

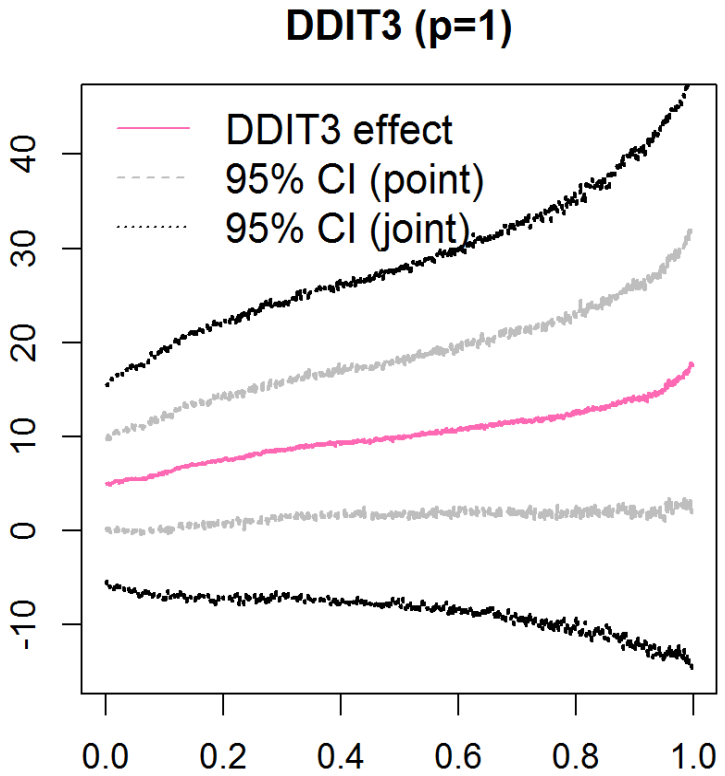




- $P_{\text{sex},\mu} = 0.004,\ P_{\text{sex},\sigma^2} = 0.121,\ P_{\text{sex},\xi} = 0.51$

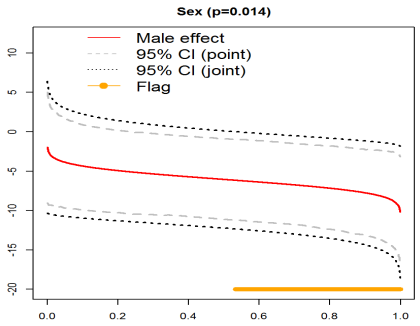
GBM Results

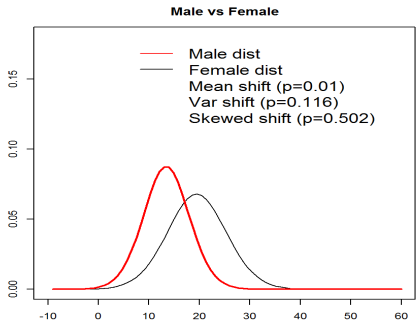


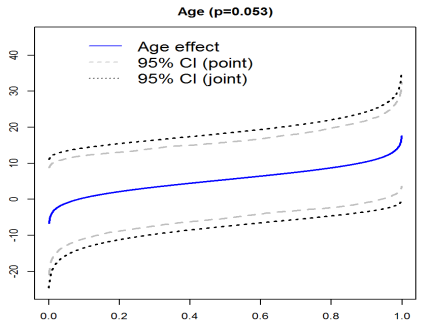


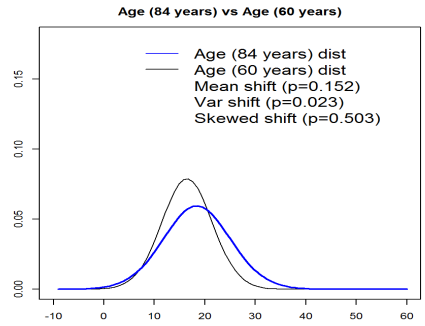
- $P_{\text{DDIT3},\mu} = 0.008$, $P_{\text{DDIT3},\sigma^2} = 0.023$, $P_{\text{DDIT3},\xi} = 0.468$

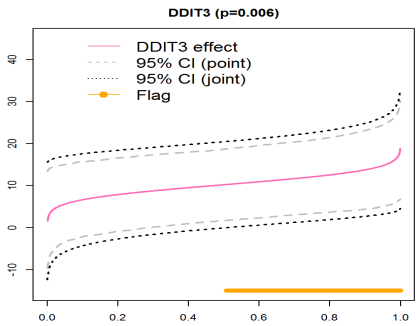
Full Results

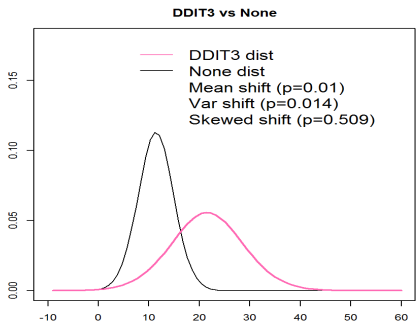


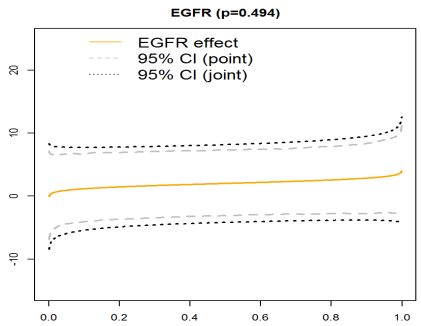


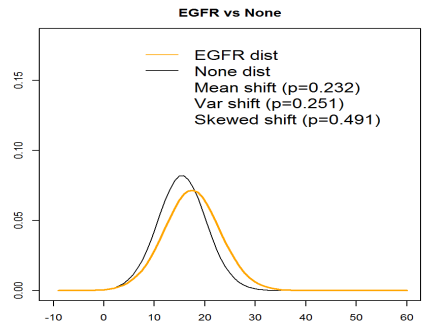


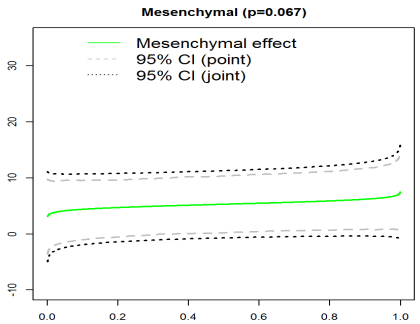


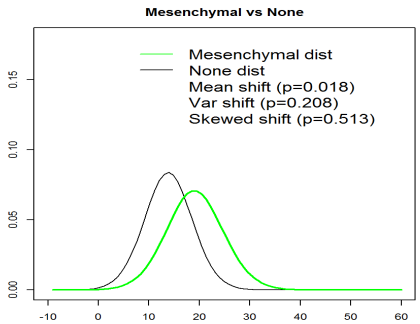


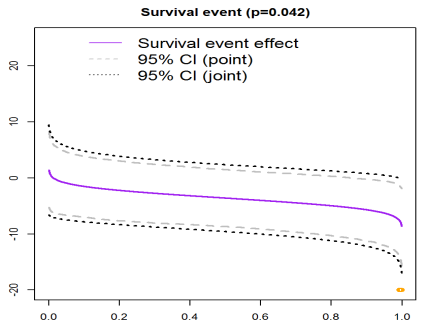


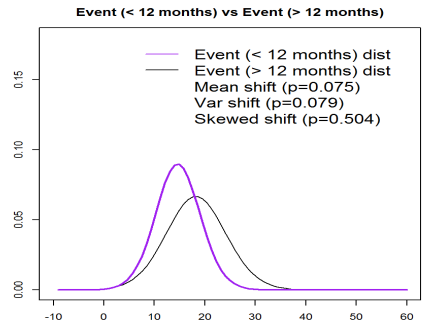












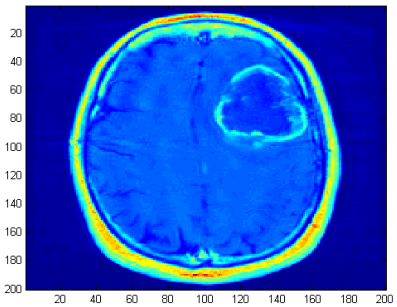
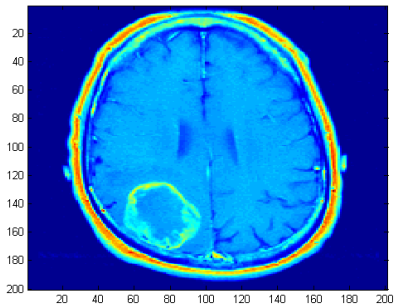
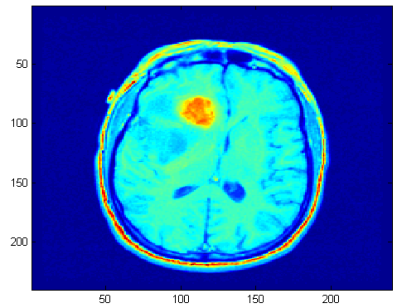
Summary

- General approach to regress distributions on covariates
- Useful in many settings (e.g. activity data, climate data)
- Introduce quantlets basis functions that are sparse, regularized, near-lossless, empirically determined, and interpretable and lead to efficient regression.
- Bayesian framework yields global and local tests that adjust for multiple testing.
  - Greater power than naive *one-p-at-a-time* approach
  - No power loss compared with feature extraction.

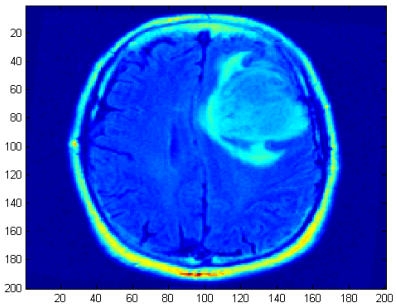
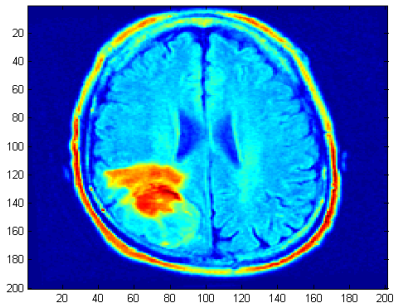
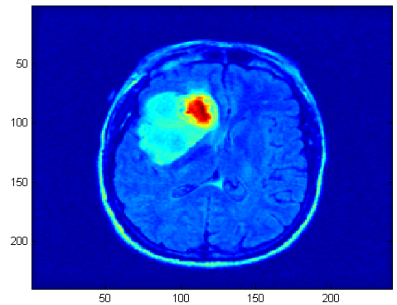
Clustering-based approaches

MRI modalities

- T1-post contrast

- T2-FLAIR

Motivation for DEMARCATE

Studies using tumor intensity values have been conducted before.

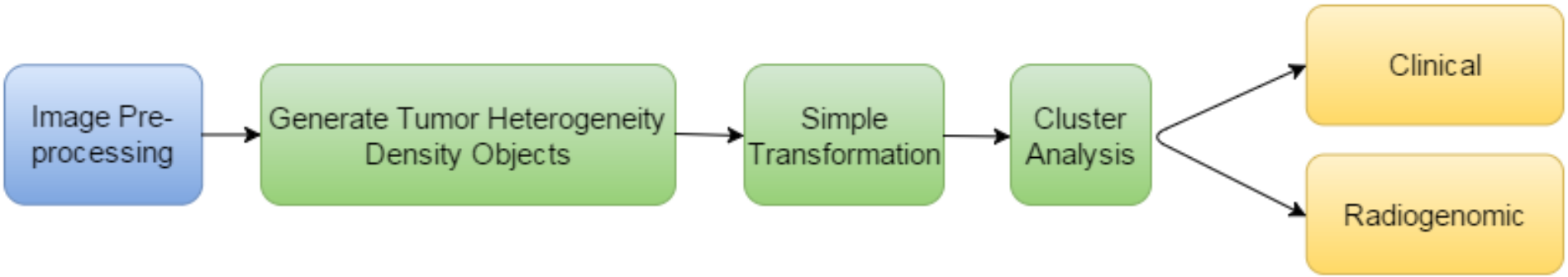
- ISSUES with previous studies:
  - Choice of number and location of summary features subjective.
  - Fail to capture small-scale and sensitive changes in the tumor.
  - Significant loss in statistical information.
- Our proposed SOLUTION:
  - Use full density!

Motivation for DEMARCATE

Studies using tumor intensity values have been conducted before.

- ISSUES with previous studies:
  - Choice of number and location of summary features subjective.
  - Fail to capture small-scale and sensitive changes in the tumor.
  - Significant loss in statistical information.
- Our proposed SOLUTION:
  - Use full density!

DEMARCATE:**DE**nsity-based **MA**gnetic **R**esonance image **C**lustering for **A**ssessing **T**umor h**E**terogeneity

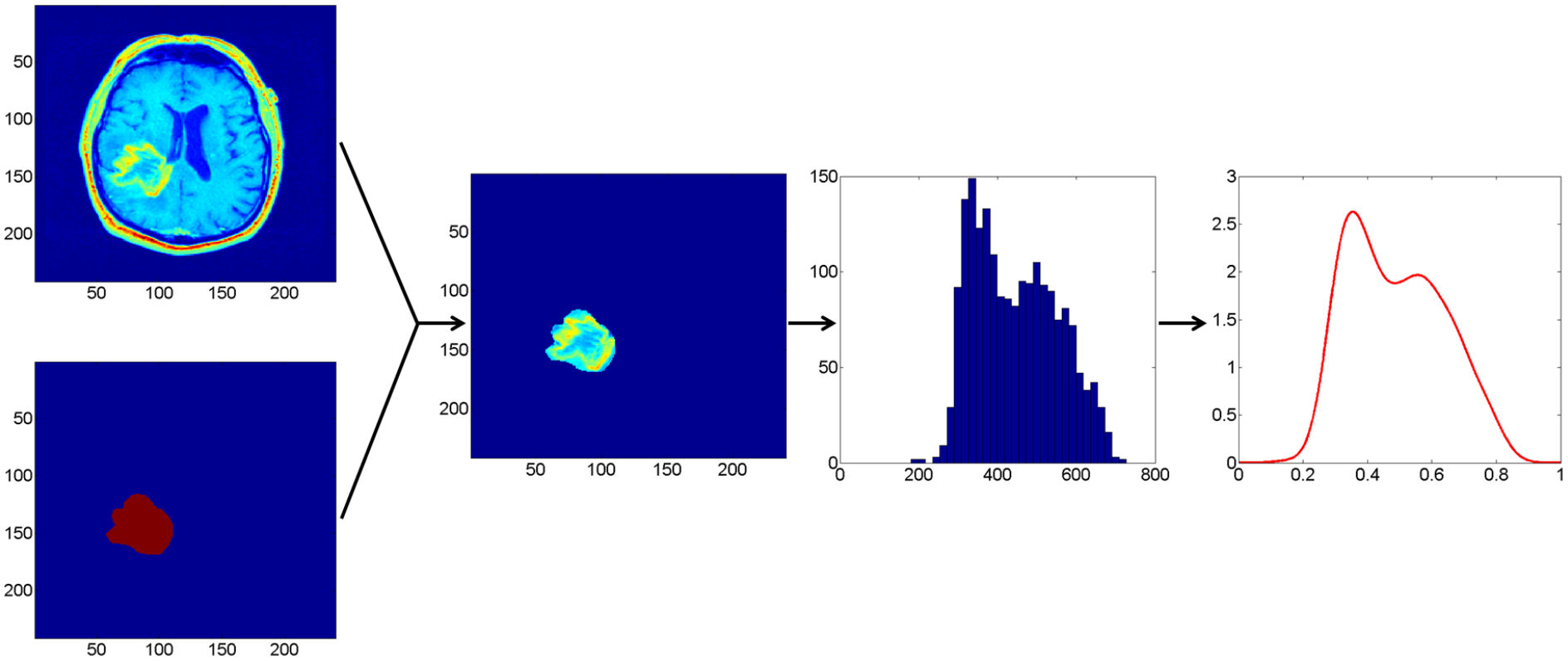


Abhijoy Saha (OSU)

The Cancer Genome Atlas (TCGA) GBM dataset

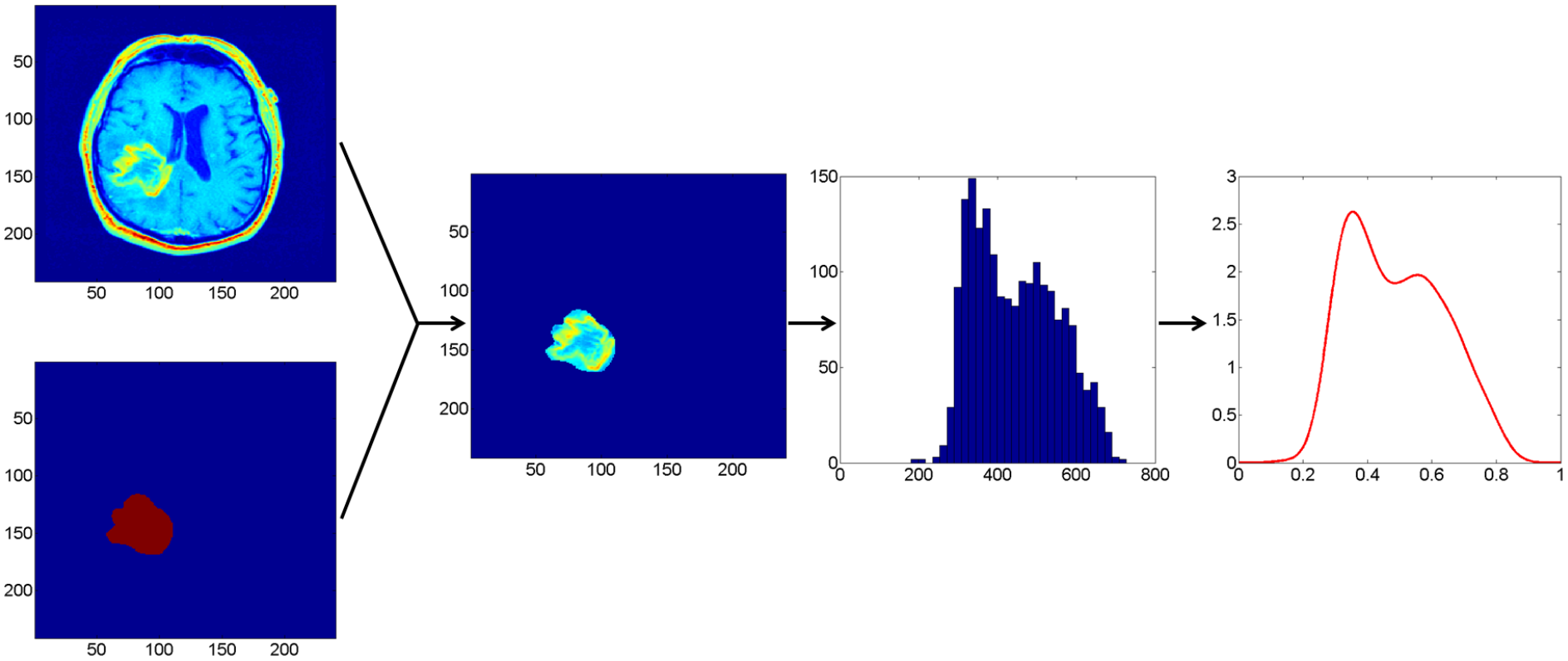
- Sample size: 64 subjects
- Imaging data: MRI obtained from The Cancer Imaging Archive (TCIA)
  - pre-surgical **T1**-weighted **post contrast**
  - **T2**-weighted fluid-attenuated inversion recovery (**FLAIR**)
- Clinical and genomic covariates:
  - Gender
  - Survival Time (in months)
  - Age (in years)
  - *Tumor Subtype* (Classical, Mesenchymal, Neural and Proneural)
  - *Gene Mutation Status*

Generation of THDP



Tumor **H**eterogeneity **D**ensity **P**rofile (**THDP**)

Generation of THDP



Tumor **H**eterogeneity **D**ensity **P**rofile (**THDP**)

Captures small-scale changes in tumors; build clustering models on density-space

Space of THDPs

Let $\mathcal{P}$ denote the space of THDPs:

$$\mathcal{P} = \{f : [0,1] \to \mathbb{R}_{\geq 0} \mid \int_0^1 f(t)dt = 1\}.$$

For any point $f \in \mathcal{P}$, the tangent space at that point is defined as

$$T_f(\mathcal{P}) = \{\delta f : [0,1] \to \mathbb{R} \mid \int_0^1 \delta f(t) f(t) dt = 0\}.$$

This tangent space will be used to define a suitable intrinsic metric between two THDPs on $\mathcal{P}$: **Fisher–Rao (FR) Riemannian metric**.

Fisher–Rao Riemannian metric

For any point f in $\mathcal{P}$ and two tangent vectors $\delta f_1, \delta f_2 \in T_f(\mathcal{P})$, the nonparametric version of the FR metric is

$$\langle \delta f_1, \delta f_2 \rangle = \int_0^1 \delta f_1(t) \delta f_2(t) \frac{1}{f(t)} dt. \tag{1}$$

Fisher–Rao Riemannian metric

For any point f in $\mathcal{P}$ and two tangent vectors $\delta f_1, \delta f_2 \in T_f(\mathcal{P})$, the nonparametric version of the FR metric is

$$\langle \delta f_1, \delta f_2 \rangle = \int_0^1 \delta f_1(t) \delta f_2(t) \frac{1}{f(t)} dt. \tag{1}$$

- DRAWBACKS:
  - FR metric changes from point to point on the space of THDPs.
  - Computation of distances on $\mathcal{P}$ between these THDPs is cumbersome.
- SOLUTION:
  - Select an equivalent representation of the space in which the calculations become much simpler.

Square-root representation (SRT)

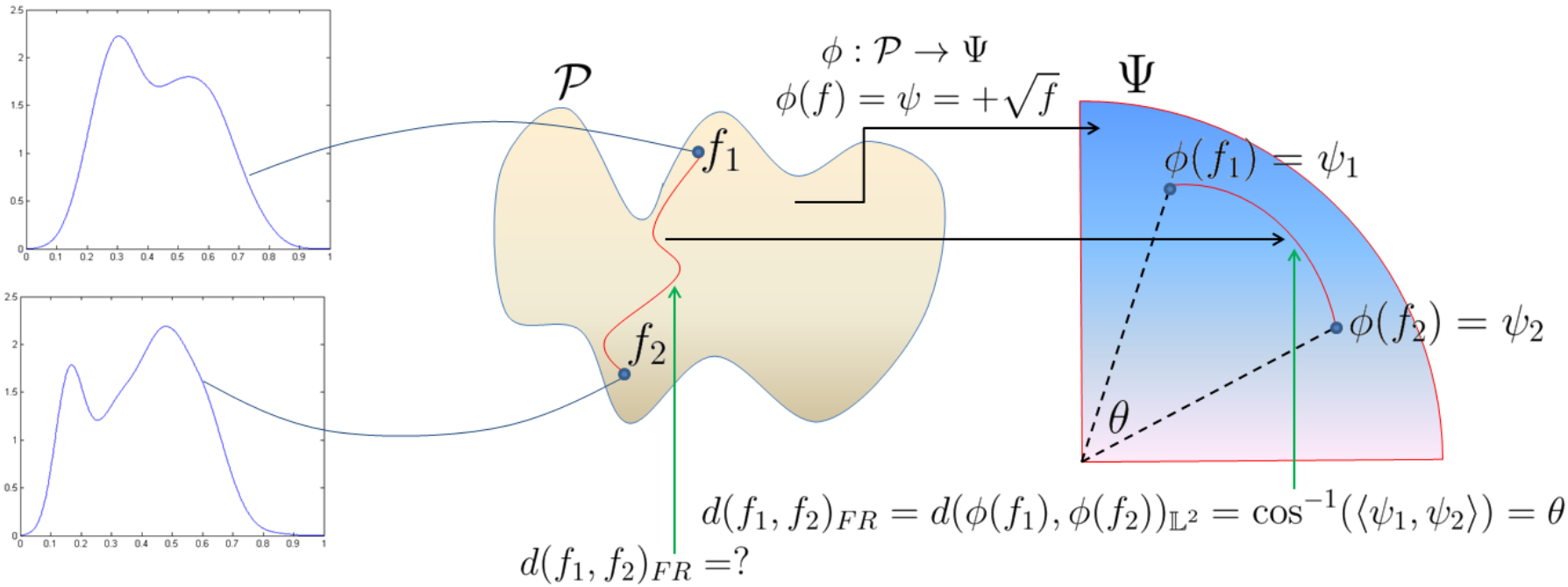
Define $\phi : \mathcal{P} \to \Psi$, where the square-root transform (SRT) of a THDP f is

$$\phi(f) = \psi = +\sqrt{f}.$$

- The space of SRT representations of THDPs is
$$\Psi = \{\psi : [0,1] \to \mathbb{R}_{\geq 0} \mid \int_0^1 \psi^2(t)dt = 1\}.$$
  - represents the positive orthant of the unit Hilbert sphere.
- $T_\psi(\Psi) = \{\delta\psi \mid \langle \delta\psi, \psi \rangle = 0\}$ denotes the tangent space at ψ.
- For any two vectors $\delta\psi_1, \delta\psi_2 \in T_\psi(\Psi)$, the FR metric becomes the standard $\mathbb{L}^2$ Riemannian metric:

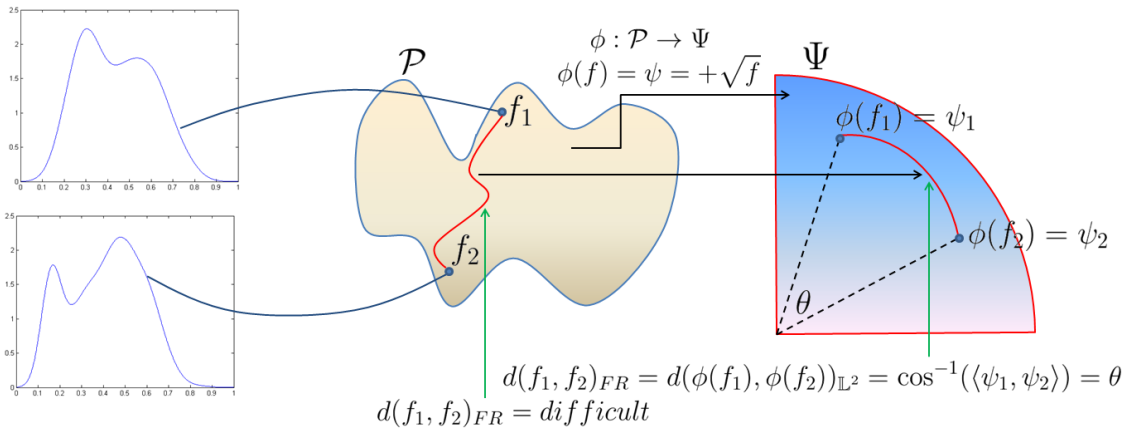
$$\langle \delta\psi_1, \delta\psi_2 \rangle = \int_0^1 \delta\psi_1(t)\delta\psi_2(t)dt. \tag{2}$$

Analysis on Ψ



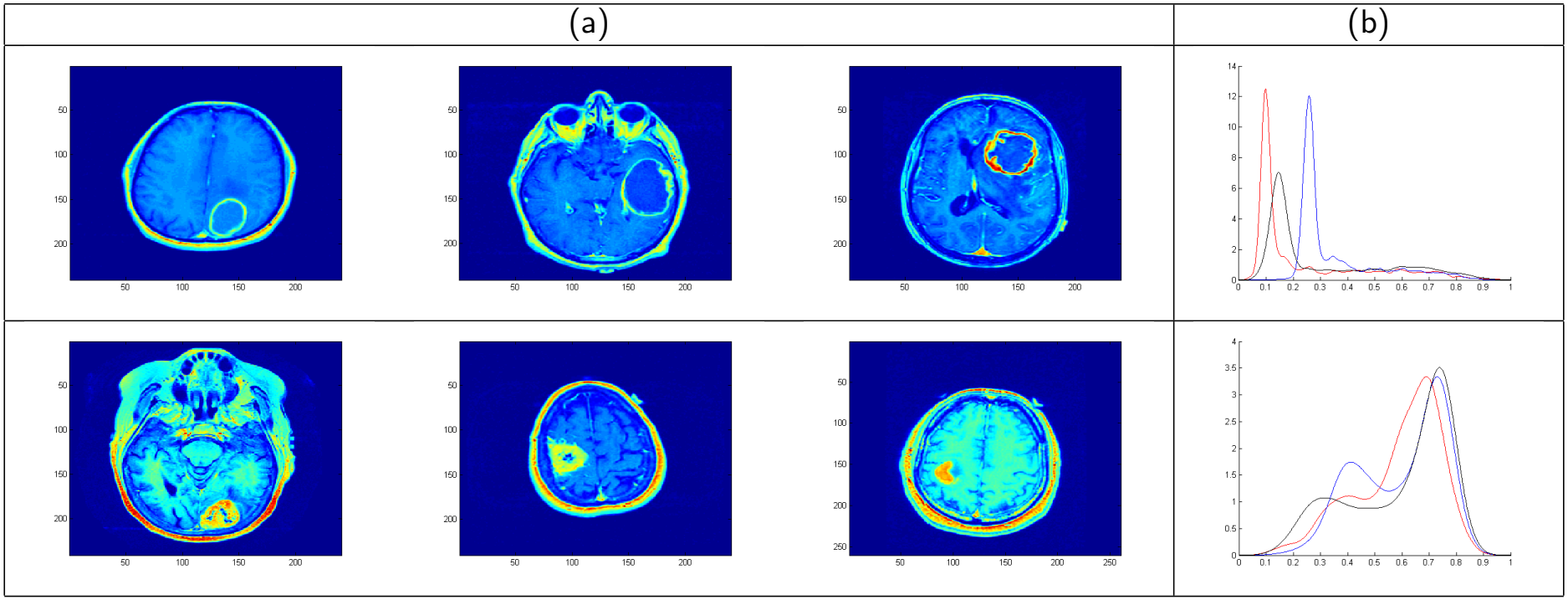
$$d_{FR}(f_1, f_2) = d_{\mathbb{L}^2}(\psi_1, \psi_2) = \cos^{-1}(\langle \psi_1, \psi_2 \rangle) = \theta.$$

Distance-based metrics for densities



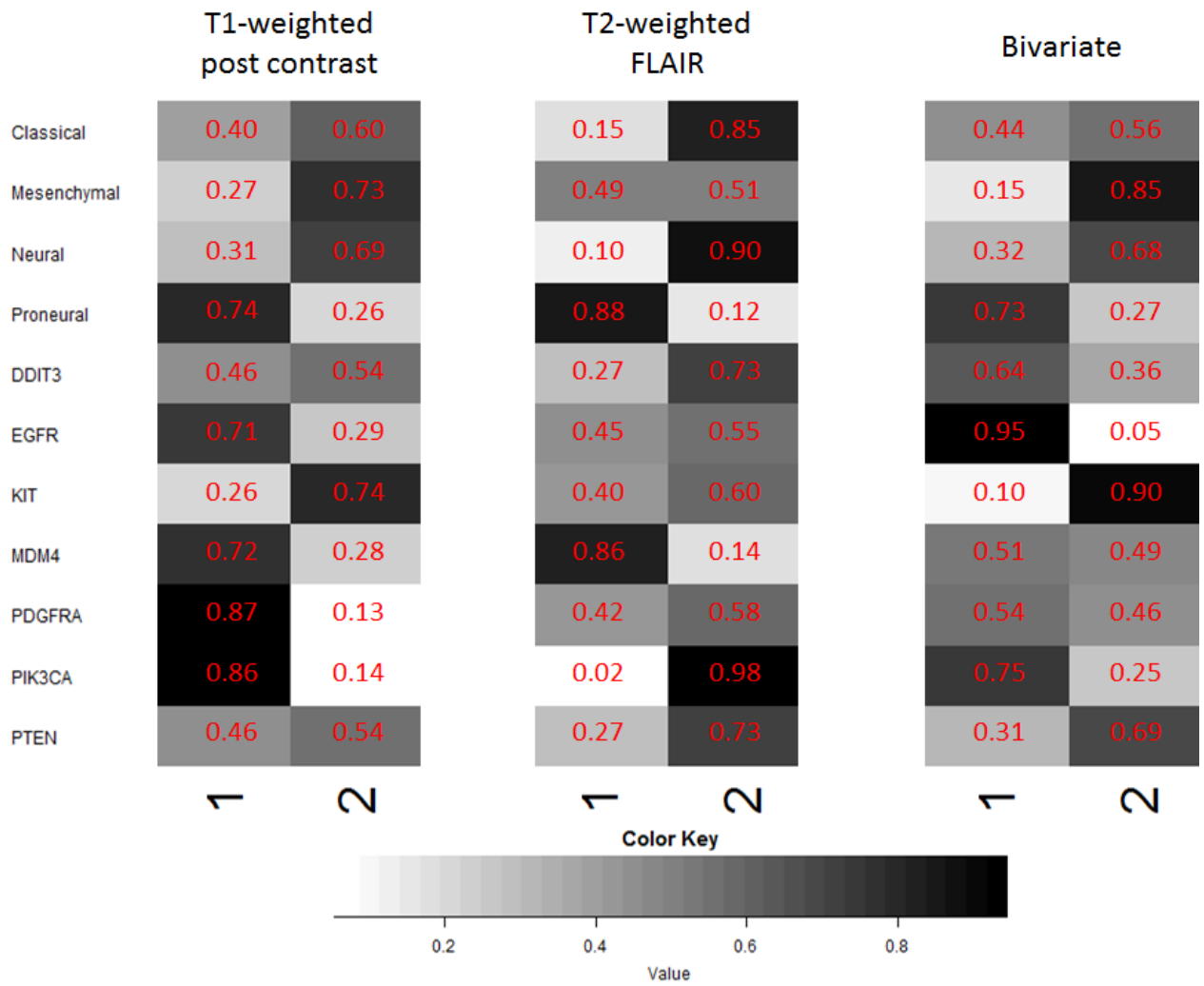
- Fisher–Rao (FR) Riemannian metric
- FR metric reduces to the standard $\mathbb{L}^2$ metric – allows explicit computation of geodesic paths and distances between densities; analytically and computationally efficient manner.
- FR metric used cluster the images

GBM Data example



2 "significant" clusters with marked differences in tumor morphology; existence of "ring-like" structure; also different in genomic characteristics and prognostic clinical outcomes.

Genomic characteristics



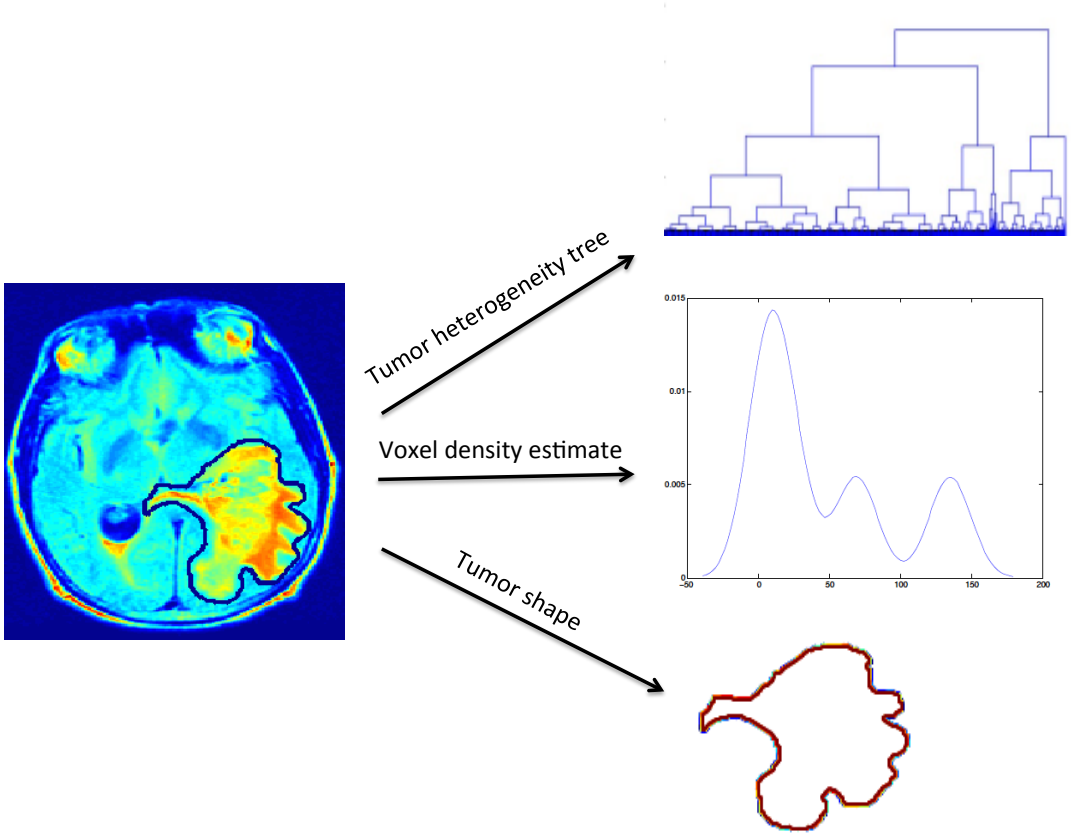
Enrichment plots for tumor subtype and genomic covariates for the T1-weighted post contrast MRI (left) and the T2-weighted FLAIR MRI (right)

Radiogenomic and clinical associations

Study notable associations between cluster partitions and external covariates (*tumor subtype*, *driver gene mutation* and *age of subject*):

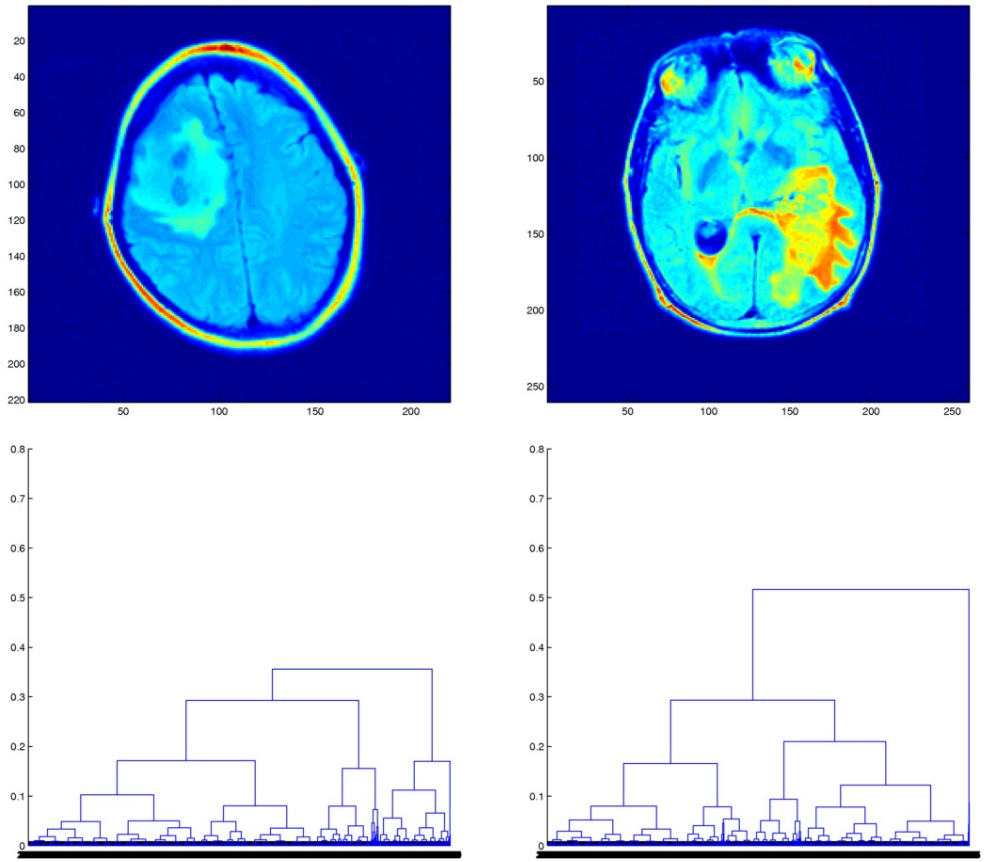
- T1-post contrast
  - *Proneural* subtype and *PDGFRA* are enriched in the same cluster.
  - *Mesenchymal* subtype and *PTEN* are enriched in the same cluster.
  - Younger patients in the *proneural* enriched cluster (**52.5** years as opposed to **59** years).
- T2-FLAIR
  - *Classical* subtype and *EGFR* are enriched in the same cluster.
  - *Neural* subtype and many of the driver genes including *DDIT3, EGFR, KIT, PDGFRA, PIK3CA, PTEN* are enriched in the same cluster.
  - Younger patients in the *proneural* enriched cluster (**51.3** years as opposed to **61.1** years).

Tree-based Characterizations



Statistical Analyses of Tree Structured data (Bharath et al; JASA, 2017)

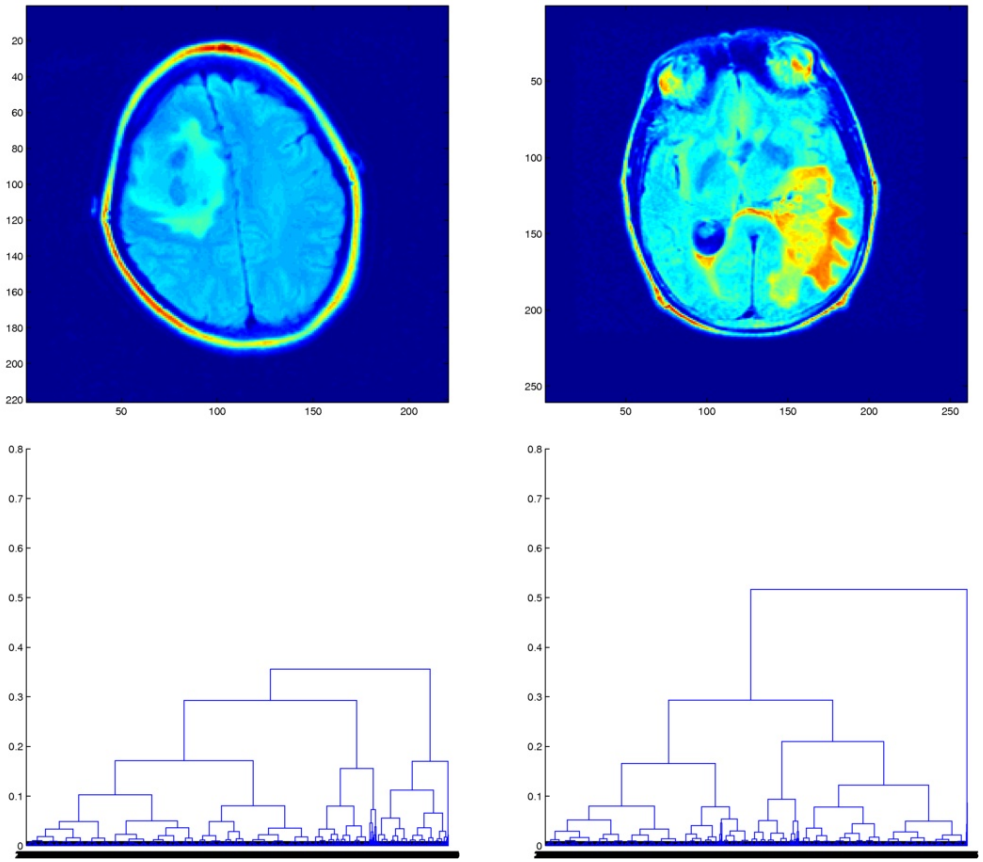
Tree-based representations



T1 MRI images. Top row left: Long survivor ($\sim$ 60 months); Right: Short survivor ($\sim$ 1 month)

Bottom row: Hierarchical clustering of voxel-wise image intensities.

Tree-based representations



T1 MRI images. Top row left: Long survivor ($\sim$ 60 months); Right: Short survivor ($\sim$ 1 month)

Bottom row: Hierarchical clustering of voxel-wise image intensities.

Tumor Heterogeneity manifested as topology of tree (e.g. height, path length, branching structure)

Statistical analyses of tree-structured data

- Analyses of non-Euclidean objects; statistical atoms are now observed "trees"
- Trees have unique topological features: height, branching structure, number of nodes etc.

Statistical analyses of tree-structured data

- Analyses of non-Euclidean objects; statistical atoms are now observed "trees"
- Trees have unique topological features: height, branching structure, number of nodes etc.
- Build probability models on trees; explicit likelihoods for generating tree-structured data; based on conditional Galton-Watson trees (David Aldous, 90's seminal work)

Bharath et al, Statistical Tests For Large Tree-structured Data, JASA T&M (2017)

Representation

- Consider a representation of a tree τ_n with n vertices and $n-1$ edges:

$$\tau_n = (\mathcal{V}(\tau_n), \mathcal{E}(\tau_n)),$$

  where $\mathcal{V}(\tau_n) = (\text{root}, v_1, \ldots, v_{n-1})$ is the topological tree without edge lengths and $\mathcal{E}(\tau_n) = (e_1, \ldots, e_{n-1})$ is the edge-set.

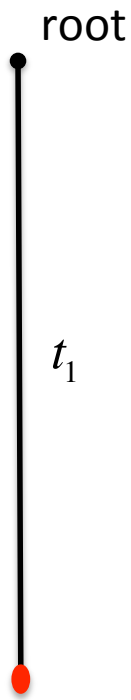
- In other words, τ_n is a point in $\mathcal{T}_n \times \mathbb{R}_+^{n-1}$ where $\mathcal{T}_n$ is the set of all combinatorial trees with n vertices.

Simple model for binary trees

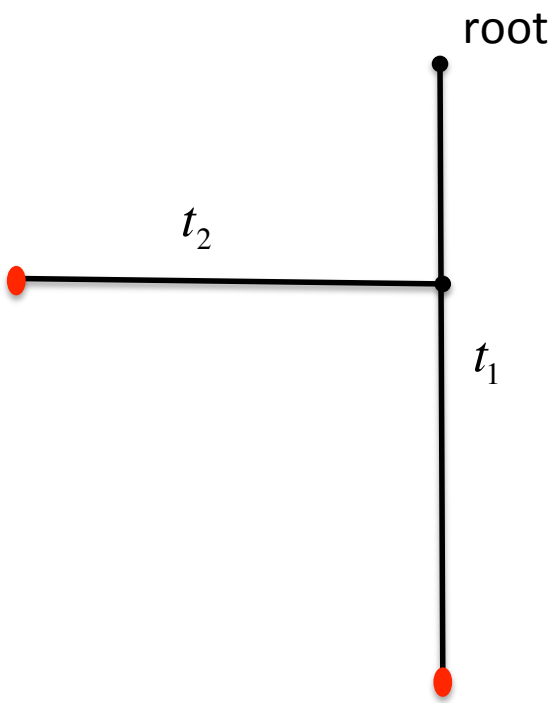
- Consider a non-homogeneous Poisson process with rate $\lambda(t) = \sigma^2 t$.
- Let $t_1, t_2, \ldots,$ be inter-event times.

Simple model for binary trees

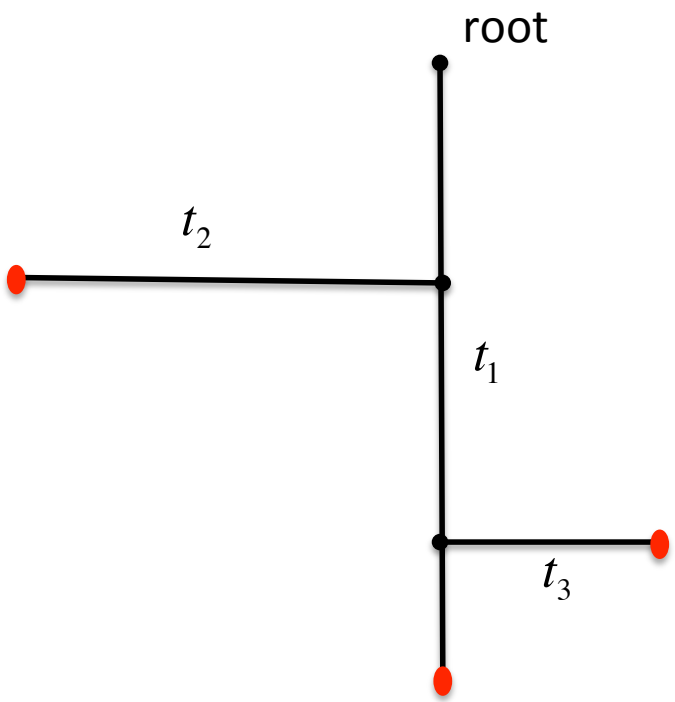
- Consider a non-homogeneous Poisson process with rate $\lambda(t) = \sigma^2 t$.
- Let $t_1, t_2, \ldots,$ be inter-event times.



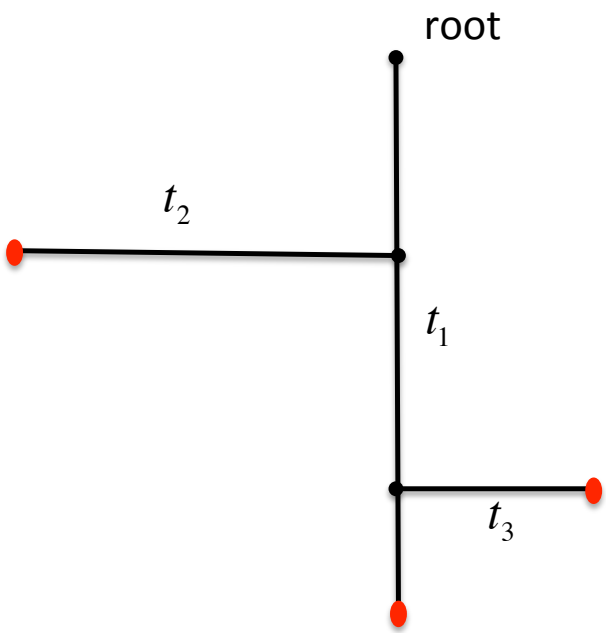
Model for binary trees

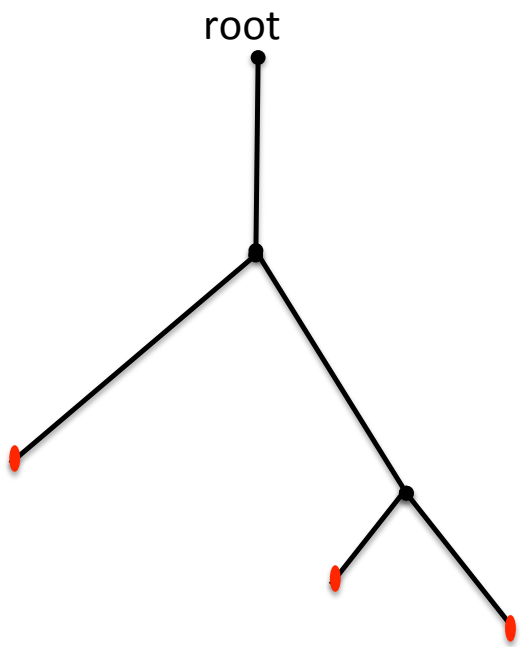


Model for binary trees



Model for binary trees





$t_1 + t_2 + t_3$ = total path length of binary tree

Model for binary trees

With n inter-event times t_i, a binary tree $\tau(n)$ with n leaves or terminal nodes, $2n$ vertices and $2n-1$ edges is constructed.

Proposition

From the properties of the Poisson process with rate t, $\tau(n)$ can be given the density

$$f(\tau(n)) = \left[\prod_{i=1}^{n-1} \frac{1}{2i-1}\right]^{-1} \frac{1}{2^{n-1}} s e^{-\frac{s^2}{2}}, \quad s = \sum_{i=1}^{2n-1} t_i$$

with respect to the product measure $\mu_n \otimes dx$ on $\mathcal{T}_{2n} \times \mathbb{R}_+^{2n-1}$, where μ_n is the uniform measure on all rooted binary trees on n leaves and dx is the Lebesgue measure on $\mathbb{R}_+^{2n-1}$.

Density for binary trees

$$f(\tau(n)) = \left[\prod_{i=1}^{n-1} \frac{1}{2i-1}\right]^{-1} \frac{1}{2^{n-1}} s e^{-\frac{s^2}{2}}, \quad s = \sum_{i=1}^{2n-1} t_i$$

- $f(\cdot)$ is impervious to labelling mechanism.
- Removal of a leaf from $\tau(n)$ results in a tree a with density $f(\tau(n-1))$.
- If the rate is θt for some $\theta > 0$, then $f(\cdot)$ retains interpretability of θ under marginalization.

Test for binary trees: one-sample

Suppose $\boldsymbol{\tau}(\boldsymbol{n}) = (\tau(n_1), \ldots, \tau(n_p))$ is an independent sample of binary trees from π_τ.

Theorem

Consider the critical function

$$\phi(\mathbf{n}, \alpha) = \begin{cases} 1 & \text{if } \sum_{i=1}^{p} s_i > \chi_{1-\alpha, 2\sum_{i=1}^{p} n_i}; \\ 0 & \text{otherwise.} \end{cases}$$

For the hypotheses $H_0 : \pi_\tau = f$ against $H_1 : \pi_\tau \neq f$, where f is the density from the non-homogeneous Poisson model, $E_{H_0}\phi(\mathbf{n}, \alpha) = \alpha$, and is invariant to the action of permutation group on leaf labels.

Test for binary trees: two-sample

Suppose $\boldsymbol{\tau}(\boldsymbol{n}) = (\tau(n_1), \ldots, \tau(n_p))$ and $\boldsymbol{\eta}(\boldsymbol{m}) = (\eta(m_1), \ldots, \eta(m_q))$ are independent samples of binary trees from π_τ and π_η respectively.

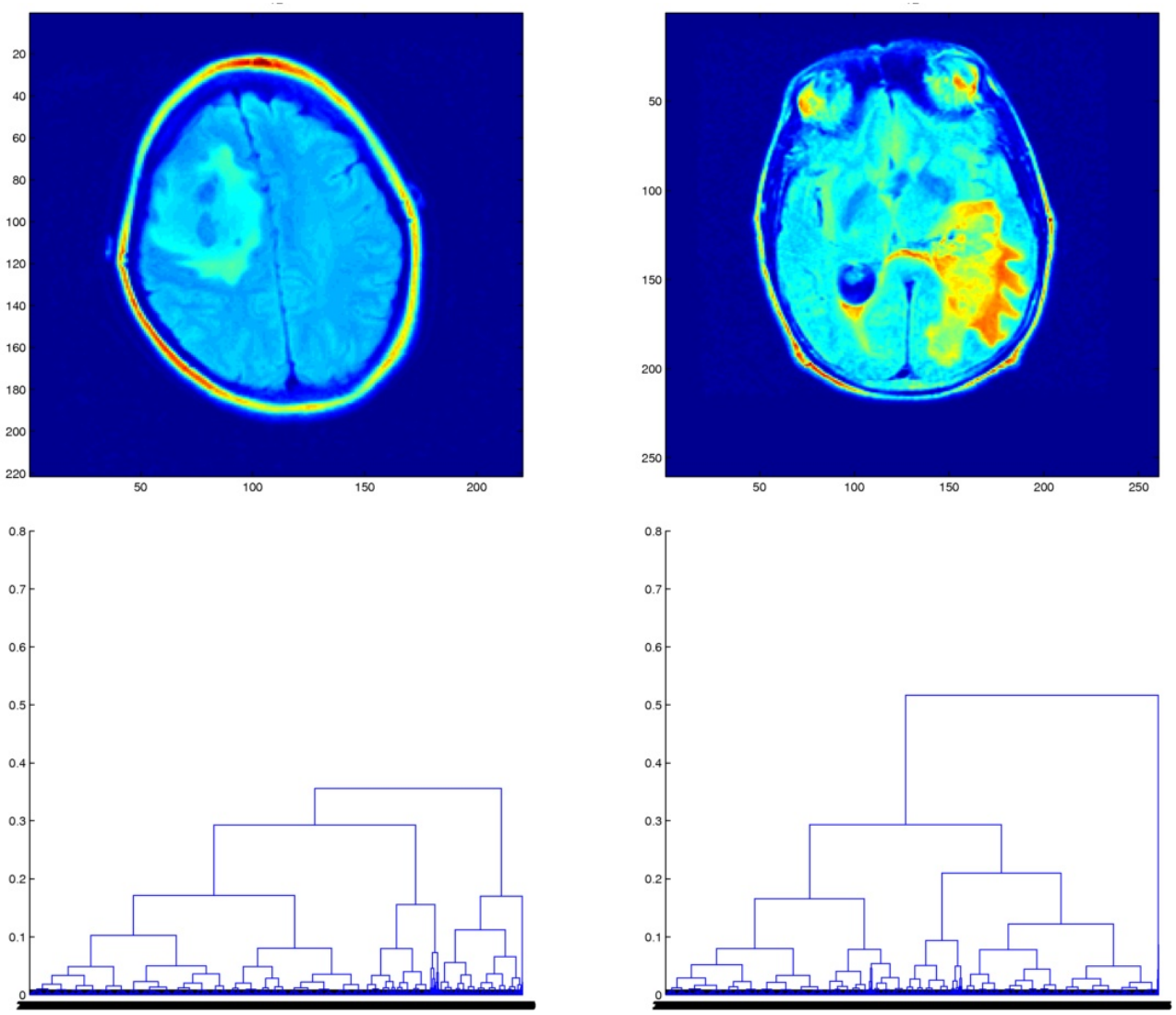
Theorem

Let r_j denote the sum of the branch lengths of $\eta(m_j)$, and without loss of generality assume that $\sum_{i=1}^{p} s_i > \sum_{j=1}^{q} r_j$. Then, the critical function

$$\psi(\mathbf{n}, \mathbf{m}, \alpha) = \begin{cases} 1 & \text{if } \frac{\sum_{i=1}^{p} s_i}{\sum_{j=1}^{q} r_j} > \left(\frac{\sum_{i=1}^{p} n_i}{\sum_{j=1}^{q} m_j} \right) F_{1-\alpha, 2\sum_{i=1}^{p} n_i, 2\sum_{j=1}^{q} m_j}; \\ 0 & \text{otherwise.} \end{cases}$$

For testing $H_0 : \pi_\tau = \pi_\eta = f$, $E_{H_0}\psi(\mathbf{n}, \mathbf{m}, \alpha) = \alpha$, and is invariant to the action of the permutation group on leaf labels.

Back to GBM data



Long surviving

Short surviving

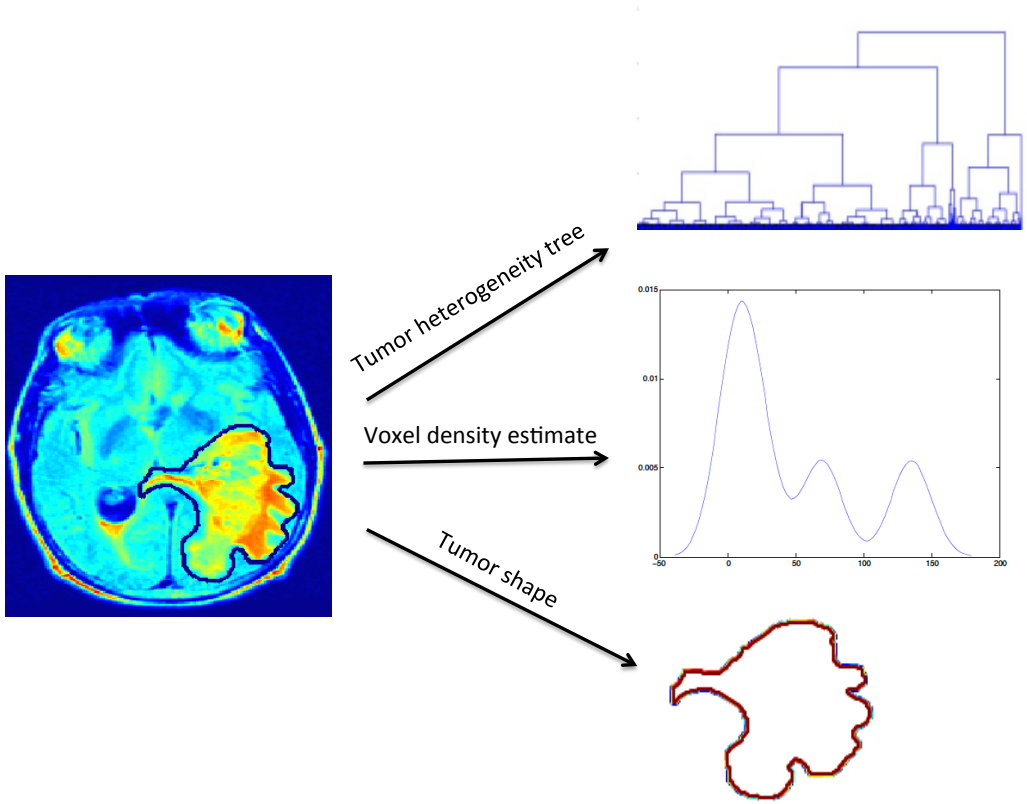
Two-sample test to detect heterogeneity

- Using the survival times, we created two groups of patients: those with survival times of utmost 12 months and those exceeding 12 months.
- The 12-month cut-off corresponded to a certain genetic classification— this was based on recommendations by neuroscientists.
- Differences in groups was detected by LCA-based test at 1% significance level.
- Naive Bayes classifier with the likelihood from LCA trees, provided 69% classification accuracy.

Two-sample test to detect heterogeneity

- Using the survival times, we created two groups of patients: those with survival times of utmost 12 months and those exceeding 12 months.
- The 12-month cut-off corresponded to a certain genetic classification— this was based on recommendations by neuroscientists.
- Differences in groups was detected by LCA-based test at 1% significance level.
- Naive Bayes classifier with the likelihood from LCA trees, provided 69% classification accuracy.
- Implications
  - Key finding: topology of trees changes "significantly" with survival and genomic variables; prospective prediction using MRI images
  - Allows embedding in more complex clustering and regression models

Shape-based Characterizations



Radiologic Image-based Statistical Shape Analysis of Brain Tumors (Bharath, Kurtek et al; JRSSC, 2018+)

Main Goals

- Shape Differences for functional estimation and regression when spatial correlation is present among curves.

Main Goals

- Shape Differences for functional estimation and regression when spatial correlation is present among curves.
- Shape Statistics: Given a collection of tumor shapes we want to generate summary statistics – mean, covariance, etc. – and study variability in tumor shape classes using principal component analysis.

Main Goals

- Shape Differences for functional estimation and regression when spatial correlation is present among curves.
- Shape Statistics: Given a collection of tumor shapes we want to generate summary statistics – mean, covariance, etc. – and study variability in tumor shape classes using principal component analysis.
- Stochastic Modeling: We want to develop statistical models that capture observed variability in tumor shapes. We also want to validate our models using random sampling.

Main Goals

- Shape Differences for functional estimation and regression when spatial correlation is present among curves.
- Shape Statistics: Given a collection of tumor shapes we want to generate summary statistics – mean, covariance, etc. – and study variability in tumor shape classes using principal component analysis.
- Stochastic Modeling: We want to develop statistical models that capture observed variability in tumor shapes. We also want to validate our models using random sampling.
- Statistical Inferences: We want to study classification, clustering, hypothesis testing, regression, etc. in the context of GBM.

(S. Kurtek)

Requirements

Require a representation of the tumor outlines and a proper metric on the space of their shapes.

- Key idea: represent tumors via their boundaries: parameterized curves.

Requirements

Require a representation of the tumor outlines and a proper metric on the space of their shapes.

- Key idea: represent tumors via their boundaries: parameterized curves.
  Desired properties of the shape metric:

Requirements

Require a representation of the tumor outlines and a proper metric on the space of their shapes.

- Key idea: represent tumors via their boundaries: parameterized curves. Desired properties of the shape metric:

- Interpretation: The metric should have an intuitive interpretation, and "measure" the types of changes in shape that are "important" for tumor development

Requirements

Require a representation of the tumor outlines and a proper metric on the space of their shapes.

- Key idea: represent tumors via their boundaries: parameterized curves. Desired properties of the shape metric:
- Interpretation: The metric should have an intuitive interpretation, and "measure" the types of changes in shape that are "important" for tumor development
- Invariance: The metric should be "preserved" by certain transformations: translation, scale, rotation and re- parameterization – "preserved" $\rightarrow$ same transformation acts on two objects, the distance between them remains unchanged: the transformations act by isometries.

Requirements

Require a representation of the tumor outlines and a proper metric on the space of their shapes.

- Key idea: represent tumors via their boundaries: parameterized curves. Desired properties of the shape metric:
- Interpretation: The metric should have an intuitive interpretation, and "measure" the types of changes in shape that are "important" for tumor development
- Invariance: The metric should be "preserved" by certain transformations: translation, scale, rotation and re- parameterization – "preserved" $\rightarrow$ same transformation acts on two objects, the distance between them remains unchanged: the transformations act by isometries.
- Efficiency: Calculations involving the metric should be computationally feasible

(S. Kurtek)

BASIC SETUP

- Let an absolutely continuous, parameterized curve be given by:

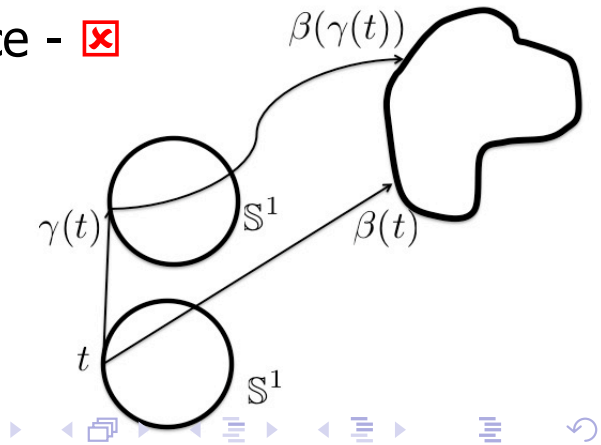
$$\beta : \mathbb{S}^1 \to \mathbb{R}^2$$

- Define the group of re-parameterizations, Γ, to be the set of all diffeomorphisms of $\mathbb{S}^1$.
- For a $\gamma \in \Gamma, \beta(\gamma(t))$ denotes the re-parameterized curve.
- If we change the parameterization of two curves in the same way, the $\mathbb{L}^2$ distance between them changes although their shapes remain the same.

$$\|\beta_1 - \beta_2\| \neq \|\beta_1 \circ \gamma - \beta_2 \circ \gamma\|$$

Desiderata: Interpretation - ?; Efficiency - ☑; Invariance - ☒

$\Longrightarrow$ Need a different framework.



ELASTIC RIEMANNIAN METRIC

- For a $\beta(t),\ t \in \mathbb{S}^1$ let $p(t) = |\dot{\beta}(t)|$ and $\theta(t) = \dot{\beta}(t)/|\dot{\beta}(t)|$.
- Define two tangent vectors $(\delta p_i, \delta\theta_i),\ i = 1, 2$ in the tangent space at (p, θ).

- Elastic Riemannian metric (a,b>0):

$$\langle\langle(\delta p_1, \delta\theta_1), (\delta p_2, \delta\theta_2)\rangle\rangle_{(p,\theta)} = a \int_{\mathbb{S}^1} \frac{\delta p_1(t)\delta p_2(t)}{p(t)} dt + b \int_{\mathbb{S}^1} \delta\theta_1(t)^T \delta\theta_2(t) p(t) dt$$

- Properties:

1. First term measures stretching while second term measures bending.
2. Difficult to work with computationally.
3. Invariant to re-parameterizations.

Desiderata: Interpretation - ☑; Efficiency - ☒; Invariance - ☑

Interesting note: This metric is closely related to the nonparametric Fisher-Rao statistical metric.

SRVF REPRESENTATION OF CURVES

- For a $\beta(t),\ t \in \mathbb{S}^1$ define the square-root velocity function (SRVF) as:

$$q(t) = \frac{\dot{\beta}(t)}{\sqrt{|\dot{\beta}(t)|}}$$

- Properties:

1. The elastic metric with a=1/4 and b=1 becomes the standard $\mathbb{L}^2$ metric and retains all of the invariance properties:

$$\|q_1 - q_2\| = \|(q_1, \gamma) - (q_2, \gamma)\|$$

2. Translation variability is automatically removed.

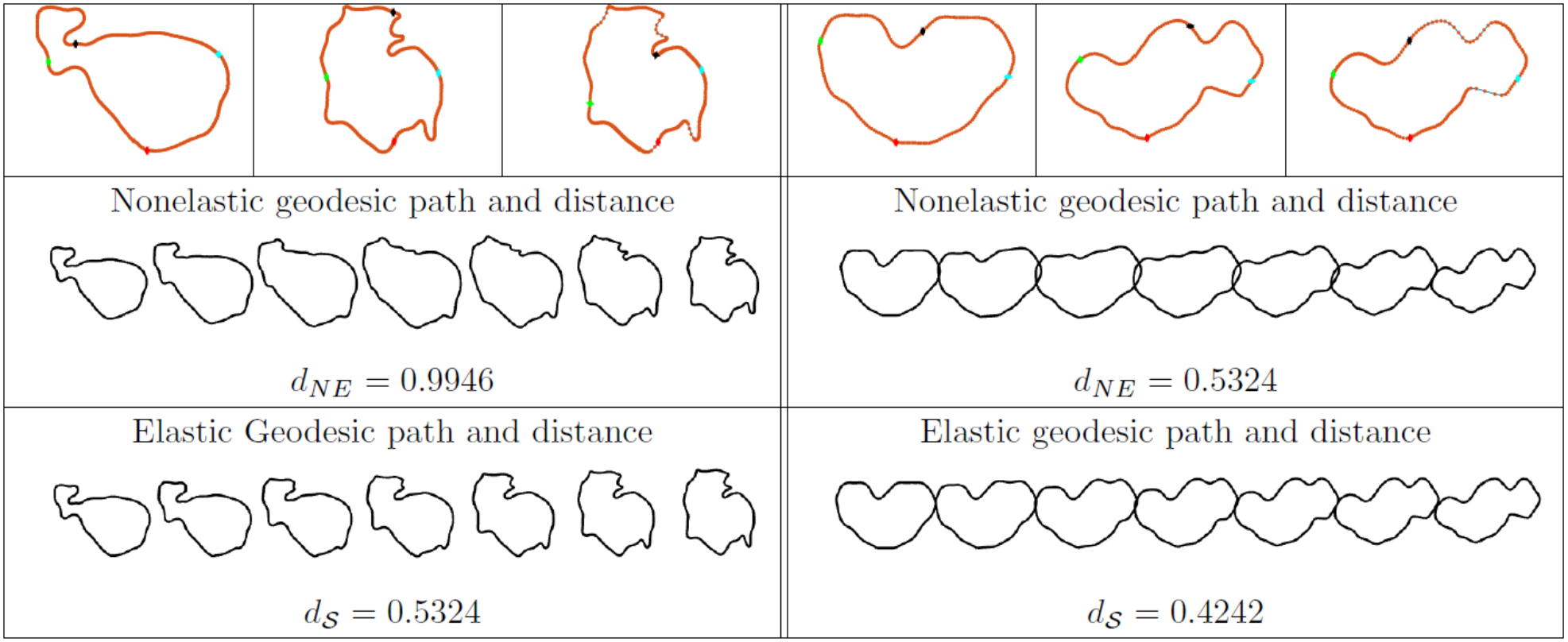
3. Curves are scaled to a fixed length – removes scaling: sphere.

$$\|q\|^2 = \int_{\mathbb{S}^1} |\dot{\beta}(t)| dt = 1$$

Desiderata: Interpretation - ☑; Efficiency - ☑; Invariance - ☑

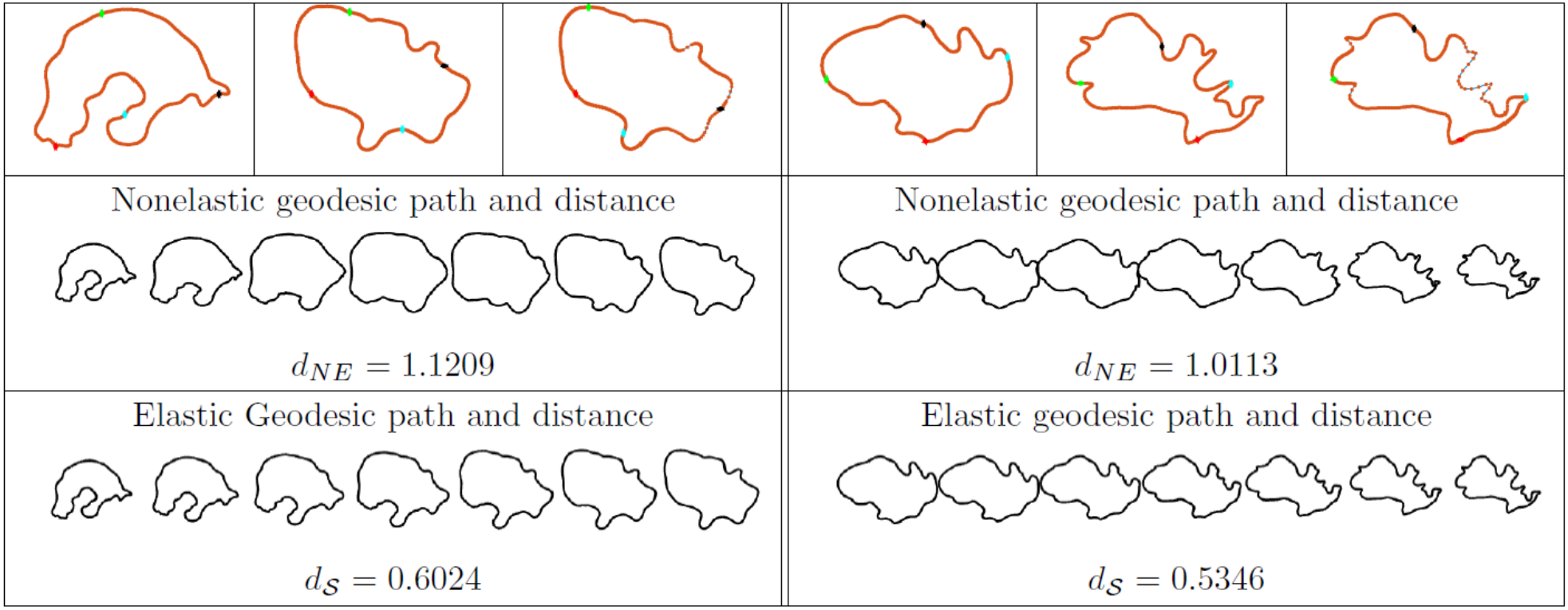
COMPARISON OF TUMOR SHAPES

- Comparison of T1 tumor shapes.
- Left panel: patients with survival times of 14.3 (left) and 29.2 (right) months.
- Right panel: patients with survival times of 8.8 (left) and 48.6 (right) months.



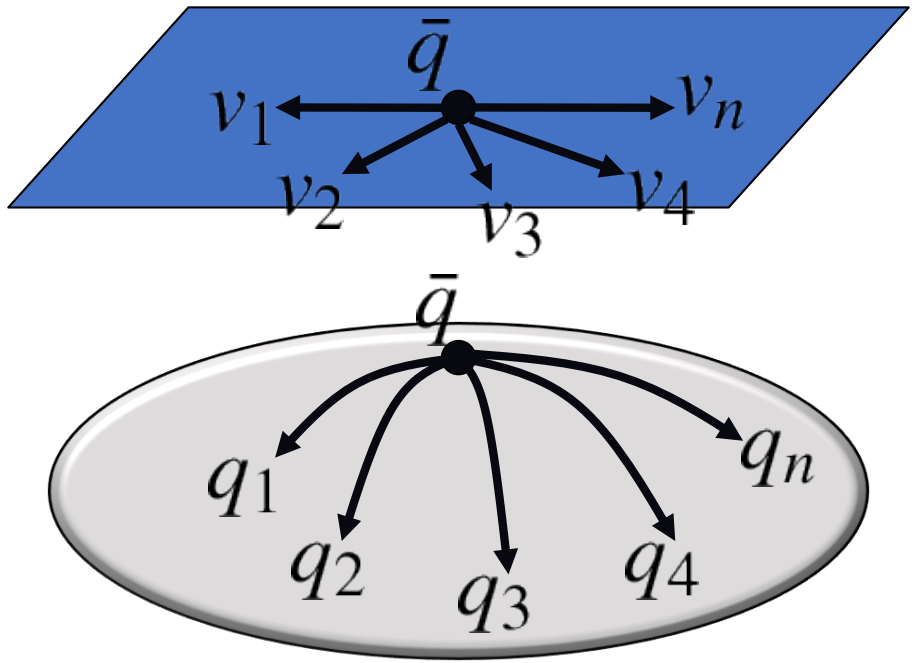
COMPARISON OF TUMOR SHAPES

- Comparison of T2 tumor shapes.
- Left panel: patients with survival times of 2.69 (left) and 13.3 (right) months.
- Right panel: patients with survival times of 6.14 (left) and 0.72 (right) months.



SAMPLE STATISTICS OF SHAPES

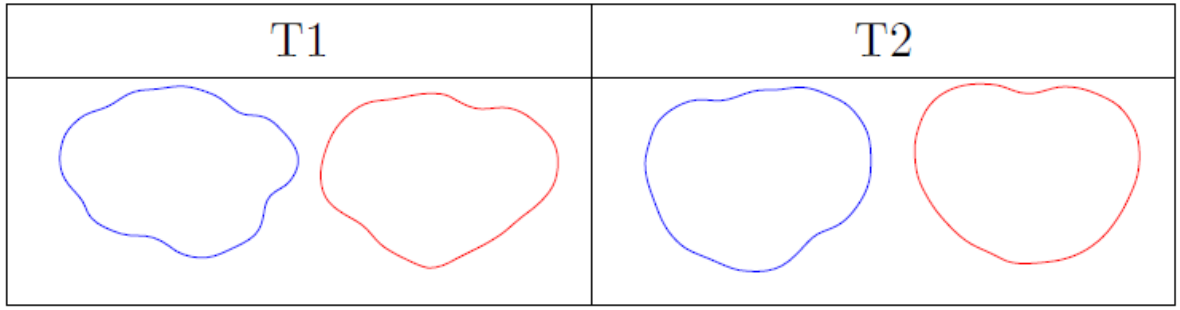
Karcher Mean: $[\bar{q}] = \arg\min_{[q]\in\mathcal{S}} \sum_{i=1}^{n} d_s([q],[q_i])^2$



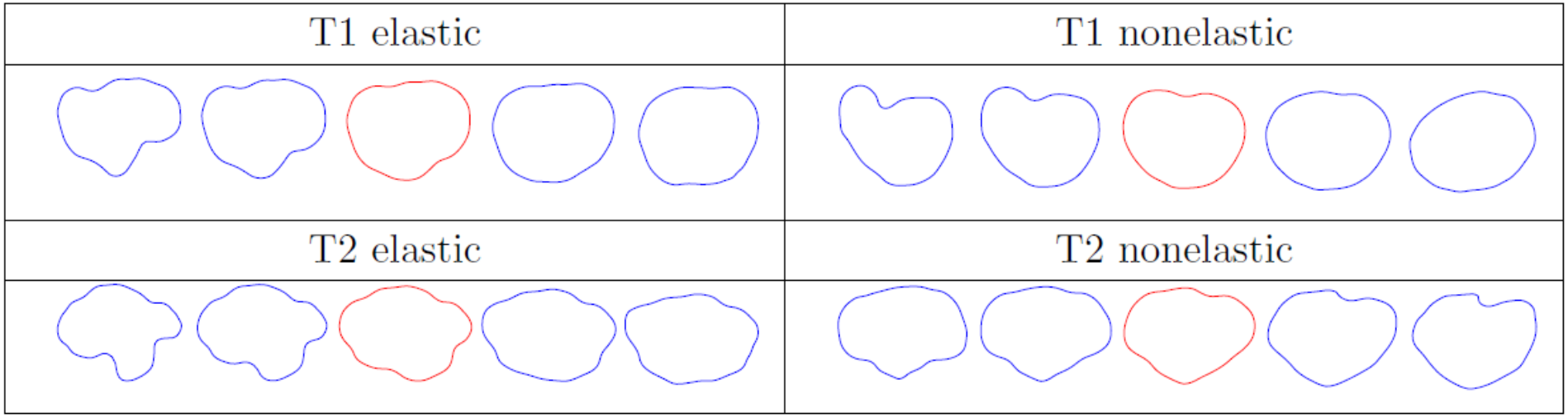
Shape variations are studied in the tangent space (using Karcher covariance) via principal component analysis (PCA).

ELASTIC VS. NONELASTIC SUMMARIES

- Comparison of elastic (blue) vs. nonelastic (red) sample average tumor shape.

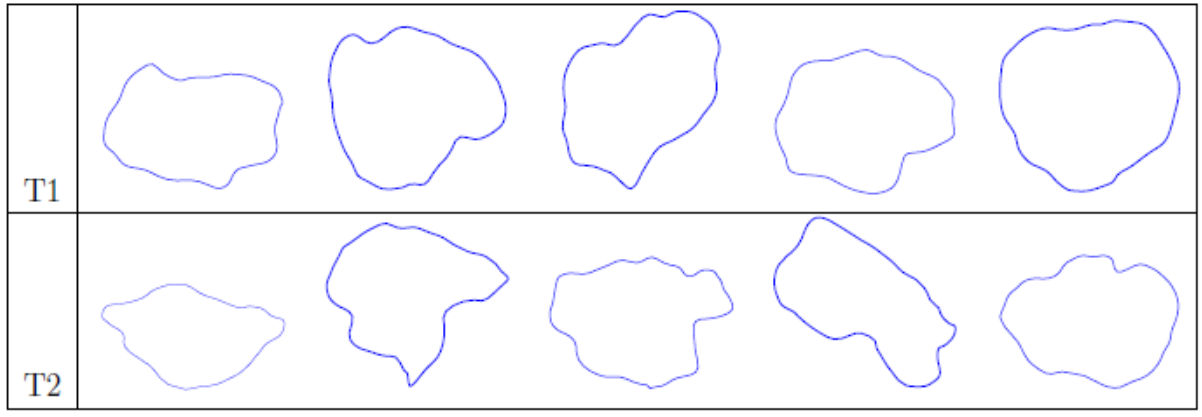


- Principal direction of variability based on elastic vs. nonelastic PCA (sample average is highlighted in red).



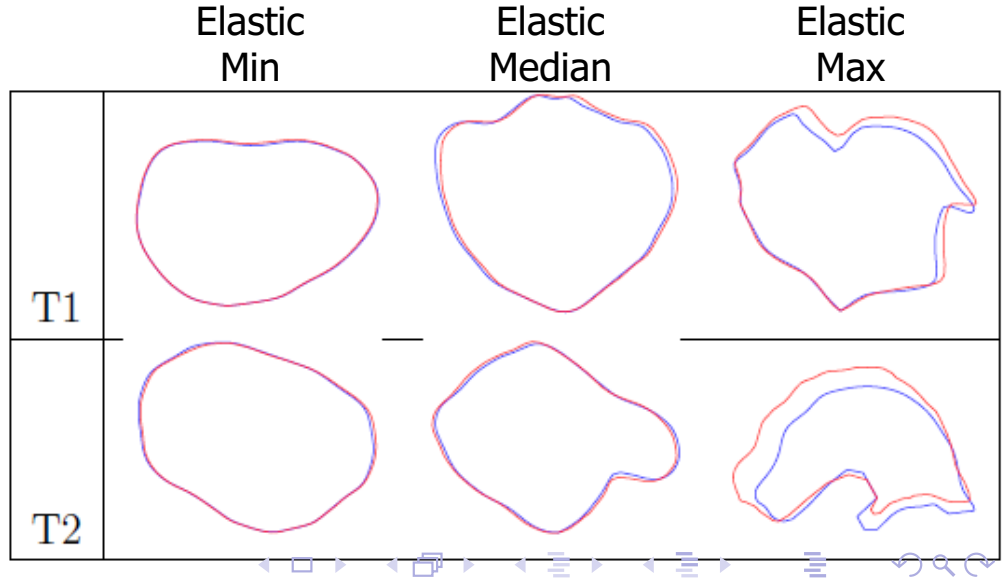
ELASTIC VS. NONELASTIC SUMMARIES

- Simulated tumor shapes via elastic PCA basis.



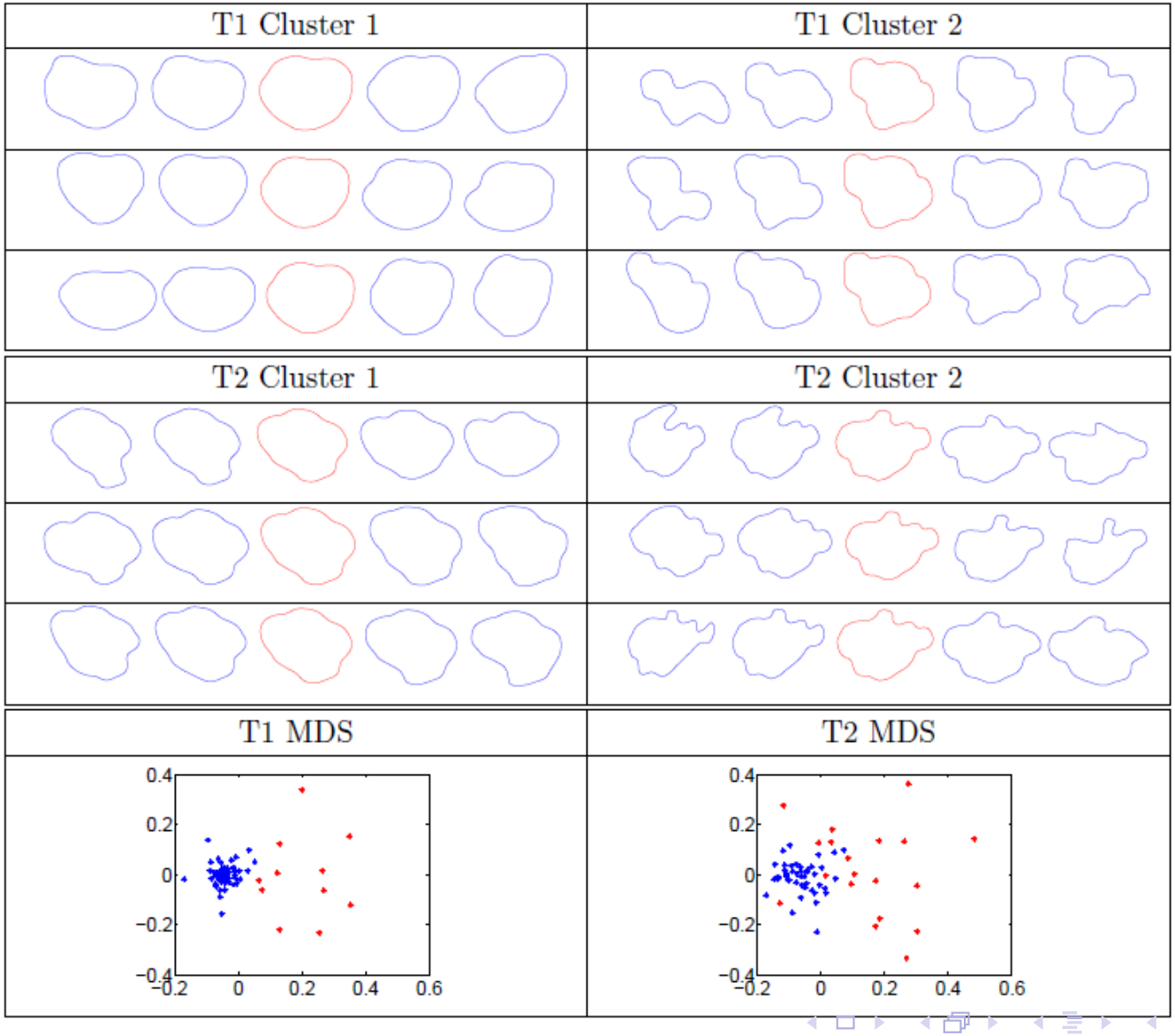
- Leave-one-out PCA-based reconstruction errors (measured via squared shape distance).

Elastic	Mean	Standard Deviation
T1	0.0097	0.0086
T2	0.0249	0.0199
Nonelastic		
T1	0.0138	0.0113
T2	0.0361	0.0293



CLUSTERING OF GBM TUMOR SHAPES

- Hierarchical clustering with complete linkage based on elastic shape distance.



CLUSTERING OF GBM TUMOR SHAPES

- Survival differences between clusters.

Survival (in months)	T1 Mean	T1 Median	T2 Mean	T2 Median
Cluster 1	18.8	14.4	18.2	14.2
Cluster 2	12.0	10.8	16.3	13.3
Difference	6.8	3.6	1.9	0.9

- Enrichment of tumor subtypes and genomic covariates in clusters.

1. Proneural subtype and PDGFRA mutation (in T2): PDGFRA plays an important role in cell proliferation and migration, and angiogenesis; this gene was found to be mutated in high amounts in the proneural subtype.

2. Classical and mesenchymal subtypes and EGFR mutation (in T2): EGFR mutation is a common molecular signature of GBM; it promotes proliferation of the tumor, which is associated with classical and mesenchymal subtypes.

SURVIVAL MODEL WITH SHAPE

- We represent tumor shapes via their PCA shape coefficients (separately for T1 and T2 tumors), and use them as tumor shape covariates in a survival model.
- We fit three proportional hazards (Cox) models:
  1. M1 with clinical covariates only,
  2. M2 with clinical and genomic covariates, and
  3. M3 with clinical, genomic and tumor shape covariates.
- For M3, due to a large number of tumor shape covariates, we fit the model with a lasso penalty and determine the value of the penalty parameter via leave-one-out cross-validation.
- Use concordance index to compare predictive ability of the three models.

Model	Predictors Significant at 0.05	C-index 1 (Harrell et al., 1982)	C-index 2 (Gömen and Heller, 2005)
$M1$ Clinical	`Age, KPS`	0.641	0.652
$M2$ Clinical+Genetic	`Age, KPS` `DDIT3, PIKC3A`	0.722	0.728
$M3$ Clinical+Genetic+Imaging	`Age, KPS, DDIT3` 11 PC shape coefs	0.859	0.841

Summary

- Statistical Models for Structured Object/Functional data
- Take structure into account for building probability models
- Computationally scalable to large (big) datasets
- Theoretical justifications for some of these approaches
- Generally applicability Multi-dimensional functions (images); multi-variate functional responses (integromics); Other settings (e.g. mobile activity data, climate data, EHR data) – in the works!

Acknowledgements

- Key Contributors:
  - Hojin Yang, Karthik Bharath, Sebastian Kurtek, Abhijoy Saha, Arvind Rao, Yang Ni, Min Jin Ha, Francesco Stingo, Jeff Morris

- Grants:
  - NIH R01 CA160736: Integrative Methods for High-dimensional Genomics Data
  - NIH R01CA194391: Graph-based Integrative Bayesian analysis of Genomics and Proteomics Data
  - NSF/NIGMS 1463233 : New Bayesian Nonparametric Paradigms of Personalized Medicine for Lung Cancer

- Main papers:
  - Quantlets (Yang et al, JASA, under revsion); Tree-structured data (Bharath et al, JASA, 2017); DEMARCATE (Saha et al, Neuroimage, 2016); Shape-based data (Bharath et al, JRSSC, 2018+); Bayesian Graphical Regression (Ni et al, JASA, 2018); PRECISE (Ha et al, Nature Scientific Reports, under revision)