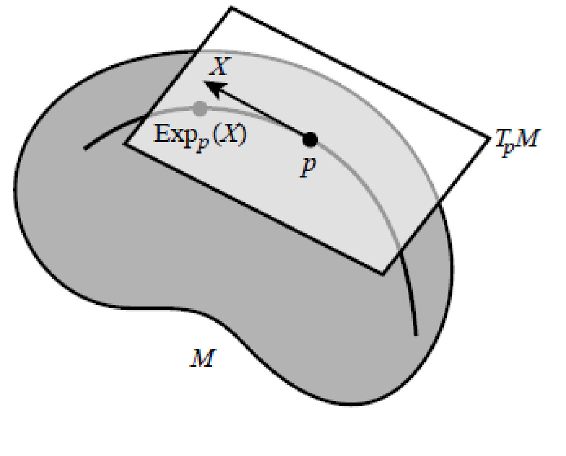
Statistics of Shape Methods

- Via Euclideanization and then via methods that assume Euclidean features
  - Euclideanization may be done in stages
    - E.g., poysphere to sphere, then PNS
- On Riemannian manifold
  - Lecture on that by Tom Fletcher on 3/7
  - Gaussian via Brownian motion [Sommer]
- On tangent planes
  - E.g., at mean
- The major reference:
  Marron & Dryden,
  Object Oriented Data Analysis



Euclidean Statistics Methods

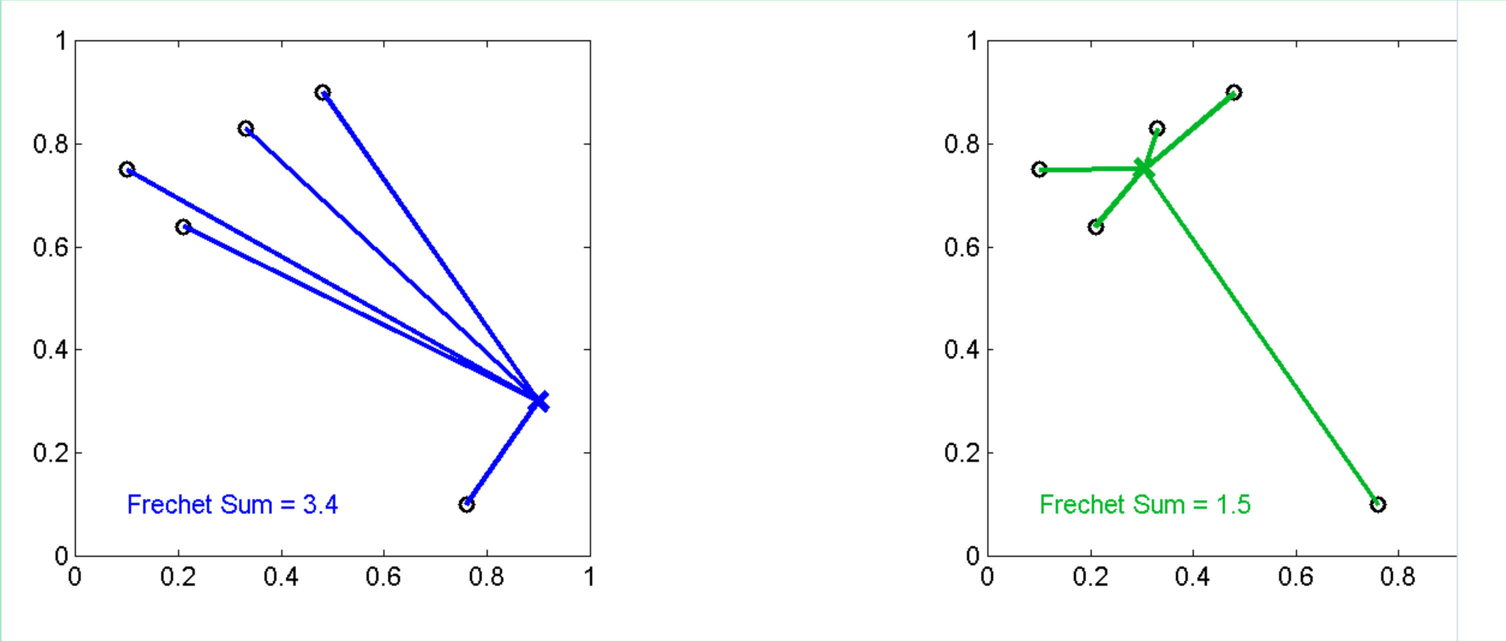
- PCA, and removal of noise
  - And other related producers of modes of variation
- Classification
  - Producing separation direction
    - Support Vector Machine and Distance-Weighted Discrimination
    - Kernels
    - Multi-entity analysis
  - Producing & using histograms along separation direction
- Hypothesis Testing
  - Permutation tests
  - DiProPerm
  - Corrections for multiple tests
- Segmentation
- Longitudinal methods

Means on Curved Manifolds

- Extrinsic
  - Do it in ambient space and then project onto the manifold
  - Fast, but often gives non-intuitive results
- Fréchet = barycenter
  - Minimizer of sum of squared distances from samples
  - According to an appropriate distance metric
  - Median: minimizes sum of distances from sample
    - Less effect of outliers
- Backwards
  - Requires a sequence of dimension reducing spaces
    - PCA
    - PNS
      - Possibly after mapping onto high-dimensional sphere

Means on Manifolds

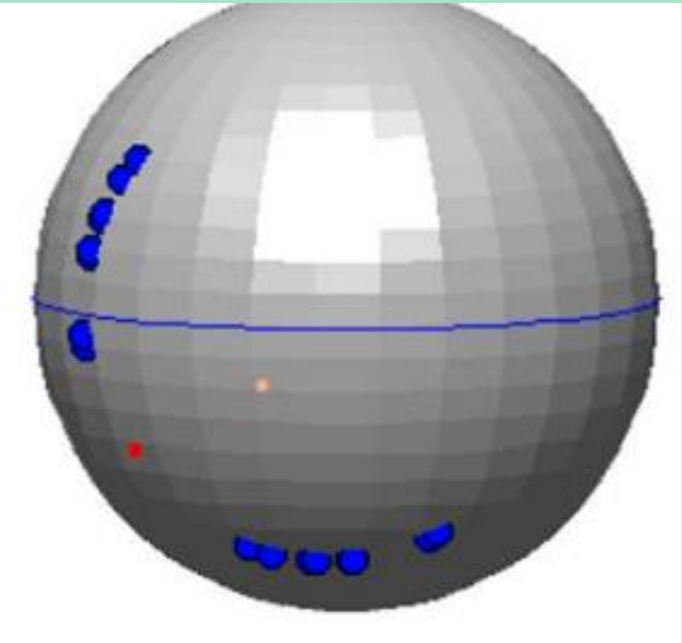
- Fréchet = barycenter
  – Minimizer of sum of squared distances from samples
  – According to an appropriate distance metric
  – Median: mimimizes sum of distances from sample
    - Less affect of outliers (see Marron & Dryden figure)



Means on Curved Manifolds

- Backwards mean from PNS

Backwards mean (light) is
better than Fréchet mean (red)



Nonlinear Effects of Curved Shape Spaces

- On means, on PCAish analysis, e.g., PNS
  - Effects depend on sign of curvature of the shape space
    - Differ on whether Euclidean distance is larger or smaller than the geodesic distances
- So the following do not act linearly
  - Including a new sample into a mean or PNS
  - Iteratively producing mean from means of successive subgroup means

Effects of Curved Shape Spaces

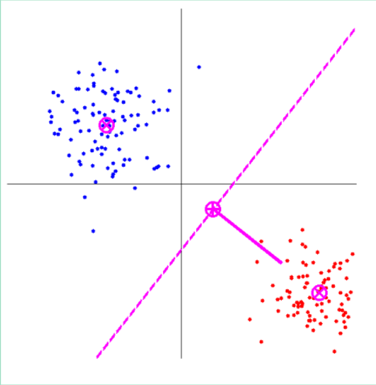
- On manifolds such as polyspheres the central limit theorem still holds: different approaches yield differences proportional to $1/\sqrt{N}$, with N = # of samples
  - And a limiting distribution that is a sort of Gaussian
  - For some kinds of shape such as trees with certain metrics central limit theorem does not hold, so convergence to mean is slower (“smeariness”), or even sticky
    - Stratified manifolds: have components of many dimensions
      - E.g., Trees with certain metrics
    - Spheres with data near both poles

Distributions About Means on Shapes

- Typically the shape features are
  - High dimension d, *but often not inherently*
  - Locally highly correlated
  - Low sample size n
  - So HDLSS for the noise but not for the signal
- Means for HDLSS act surprisingly for independent, equally distributed random variables
  - There are hardly any data samples near the mean
  - For equally distributed independent random variables with unit standard deviation, in the limit for large n, all the data lives near a sphere of radius $\sqrt{d}$ away from the mean
  - Can be a good model for noise
  - Ref: Marron & Dryden, section 14.2

Covariance, Modes of Variation, and Entropy on Curved Manifolds

- Want representation independent of boundary parameterization
- Often benefits from pre-centering
- Modes of variation: minimizer of distances to subspaces ordered by dimension
  - PCA, PNS (backwards)
- ICA: directions of maximal non-Gaussianity
    - Often captures causality well
  - Classification separation direction
    - Use DWD for robustness
    - Difference of means can be useful
- Entropy: Spread of distribution
  - So for correspondence, want low entropy
  - From principal variances λ_i
    - Problem from very small principal variances
  - For Gaussian: $(d/2)\,[1+ \ln(2\pi) + \text{avg} \ln \lambda]$



Modes of Variation Usage

- Dimensionality reduction
  - Coefficient of selected modes become features
- Segmentation by posterior probability optimization
  - Coefficients of shape eigenmodes are optimized over
    - $\arg\max_{\underline{\mathbf{z}}} (p(\underline{I} \mid \underline{\mathbf{z}})\, p(\underline{\mathbf{z}}))$ with $\underline{\mathbf{z}} = \mu_{\underline{\mathbf{z}}} + \Sigma_i\, c_i\, \sigma_i\, \mathbf{v}_i$
      with σ_i being covariance eigenvalues and
      $\mathbf{v}_i$ being covariance eigenmodes

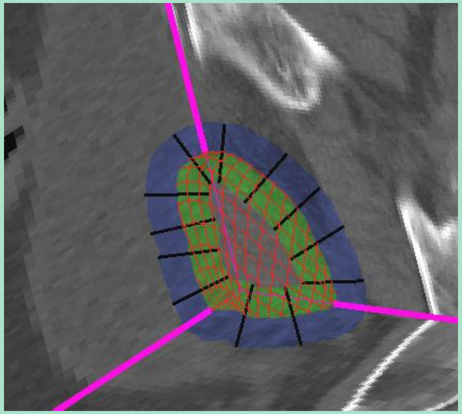
    $= \arg\min_{\underline{c}} (\Sigma_i\, d_i^2 + \Sigma_i\, c_i^2))$, where the representation $\underline{J}$ of $\underline{I}$ (which depends on $\underline{\mathbf{z}}$–based correspondence) has PCA results leading to $\underline{J} = \mu_{\underline{J}} + \Sigma_i\, d_i\, \rho_i\, \mathbf{w}_i$

Segmentation by Posterior Optimization

- Let $\underline{\mathbf{z}}$ be the geometric representation of one or more objects
- Let $\underline{J}$ be the representation of image intensity features
- Segmentation by posterior probability optimization is computing arg $\max_{\underline{\mathbf{z}}}$ p($\underline{\mathbf{z}}$ | $\underline{I}$)
  - $p(\underline{\mathbf{z}} \mid \underline{I}) = p(\underline{J} \mid \underline{\mathbf{z}})\ p(\underline{\mathbf{z}}) / p(\underline{J})$, so
    $\arg\max_{\underline{\mathbf{z}}} = p(\underline{J} \mid \underline{\mathbf{z}})\ p(\underline{\mathbf{z}}) = \arg\min_{\underline{\mathbf{z}}} \left(\log p(\underline{J} \mid \underline{\mathbf{z}}) + \log p(\underline{\mathbf{z}}) \right)$
  - = if both distributions are Gaussian,
    $\arg\min_{\underline{a}} \left(\Sigma_i \left({b^{lkhd}_i}^2/\lambda^{lkhd}_i\right) + \Sigma_j\left({a^{shape}_i}^2/\lambda^{shape}_i\right)\right)$, if
    shape and intensity patterns are independent, where
    - The b_i are the coefficients of principal components of the covariance of $\underline{J}$
    - The a_i are the coefficients of principal components of the covariance of $\underline{\mathbf{z}}$
    - The λ^{shape}_i are the principal variances of shape
    - λ^{lkhd}_i are the principal variances of $\underline{J}$, i.e., of the intensity features in shape relative coordinates

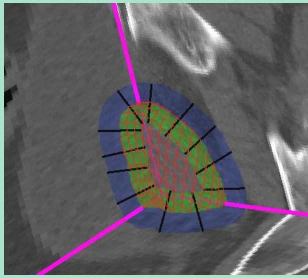
Segmentation by Posterior Optimization, 2

- The issues now are
  - What are the intensity features <u>J</u>?
  - How can we get correspondence of <u>J</u> over the training population and in the candidates in the population?
  - How can we get few eigenmodes of <u>**z**</u>, thus speeding the optimization?
- To get intensity features in correspondence their locations in the near exterior of the shape need to be in correspondence
  - The (θ, τ_1, τ_2) are appropriate for just that purpose
    - Though we might modify the shape correspondences to also reflect intensity pattern correspondences
    - τ_2 (the along-spoke radial distance) needs to be extended to the object's near exterior without crossing
    - So <u>J</u> comes from a tuple over a selected collection of (θ, τ_1, τ_2)



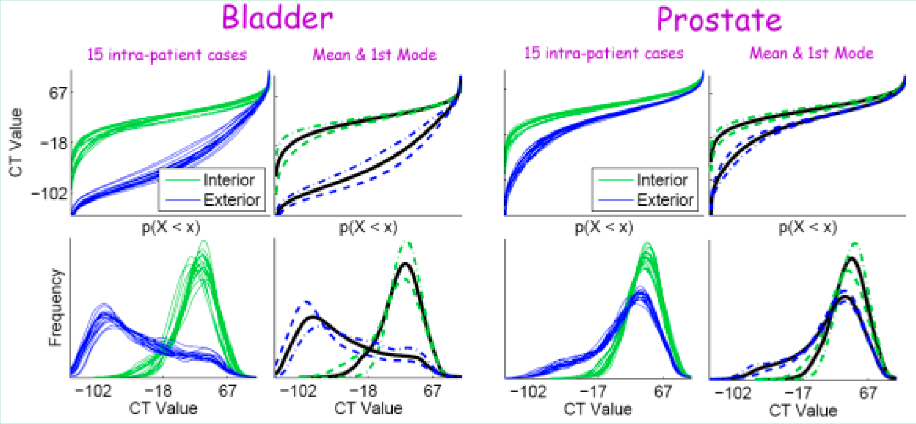
Segmentation by Posterior Optimization, 3

- What are the intensity features J?
  - Often histograms in corresponding regions are effective
    - Where the histograms are probability distributions



- Thus, the question is how to find metrics on the distances between probability distributions
  - Mallows distance = Wasserstein distance (Statistics terms) = Earthmover's distance (Computer Science term)
    - Map probability densities to quantiles; only need a few
    - The quantile at $P = P_k$ is the fraction of the density integral $< P_k$
  - Euclidean distances between quantiles = Earthmover's distance between the histograms
    - Rather few P_k are needed

Figure 4.4: QFs estimated from global image regions of the bladder and prostate from a single patient over 15 days. For each region, the mean QF and $\pm 2\sigma$ along the first principal is given, where the modes for region are computed independently. Histogram estimates of these QFs are also displayed.

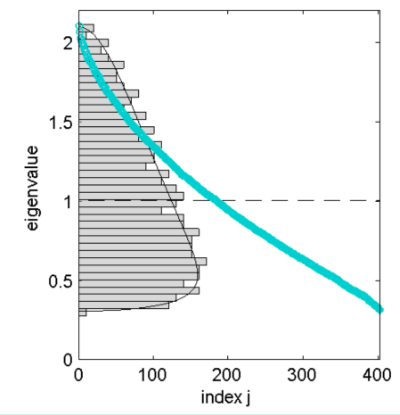


RE Broadhurst UNC dissertation

Segmentation by Posterior Optimization, 4

- How can we get few eigenmodes of $\underline{\mathbf{z}}$, thus speeding the optimization $\arg\min_{\underline{a}} \left(\Sigma_i \left({b^{lkhd}_i}^2/\lambda^{lkhd}_i\right) + \Sigma_j\left({a^{shape}_i}^2/\lambda^{shape}_i\right)\right)$?
  - This is a problem of dimensionality reduction
  - And it is a problem of getting accurate
  - eigenmodes and eigenvalues
    - For PCA, eigenmodes of a covariance matrix
- The deep problem is that noise pollutes the eigenvalues and eigenmodes and greatly increases the number of apparently significant eigenmodes.
  - See next slides
  - Scree plots show sorted eigenvalues
  - MP plots show density of eigenvalues

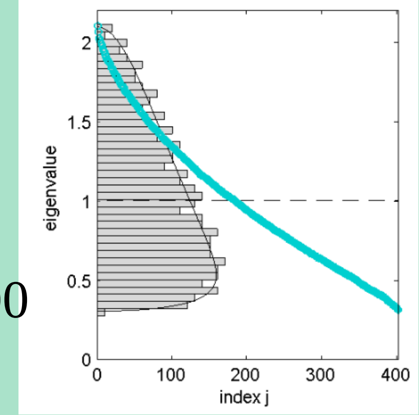
Scree plot vs. MP plots for IID indep noise



Getting the Right # of Eigenmodes

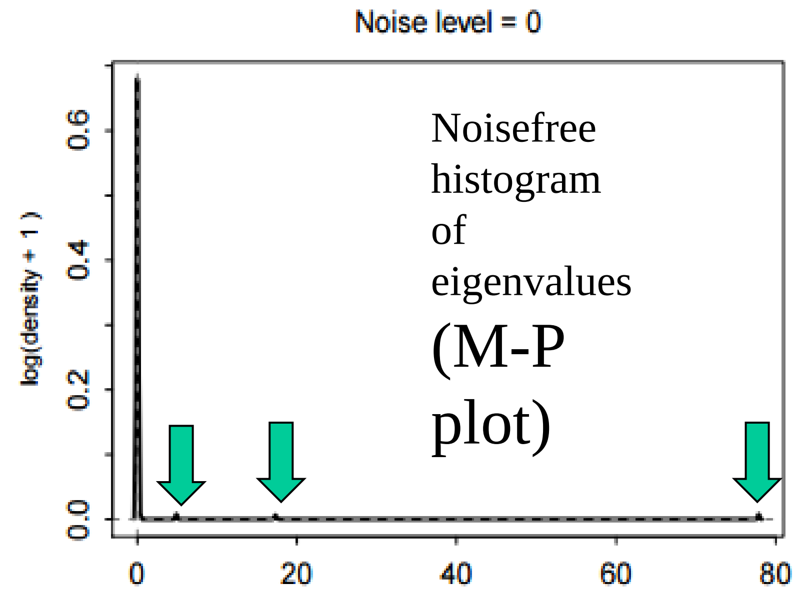
- How can we get few eigenmodes from Cov($\underline{\mathbf{z}}$), thus speeding the optimization?
- The deep problem is that noise pollutes the eigenmodes and greatly increases the number of apparently significant eigenmodes.
  - The covariance matrix is a sample from a distribution and has noise [Marron, ch 14]
  - The solution lies in Marchenko-Pastur analysis
    - Studies the probability distribution of eigenvalues (MP plots)
    - When n < d there will be a spike of height n-d at $\lambda = 0$
    - And the noise-based spread of that spike hides the correct eigenvalues

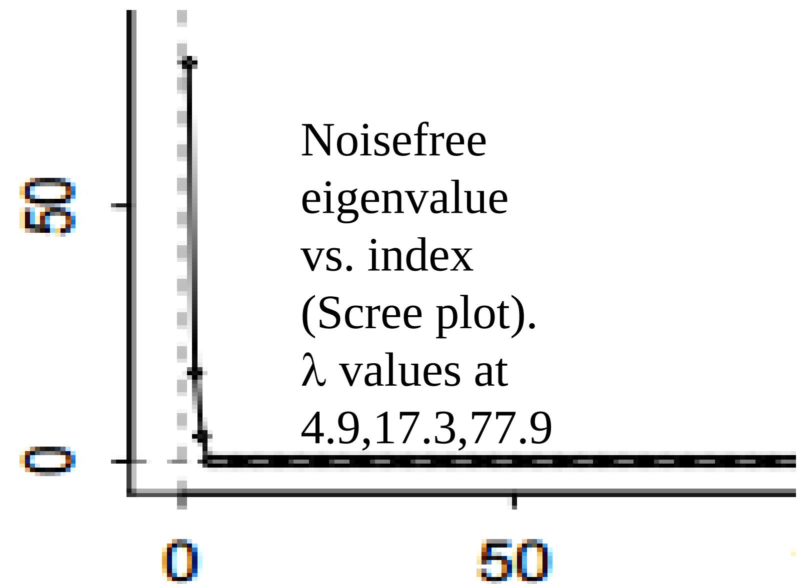
Scree plot vs. MP plots for IID indep noise; n=2000, d=400



Weakness in evaluation via principal variances (PCA eigenvalues) [Choi]

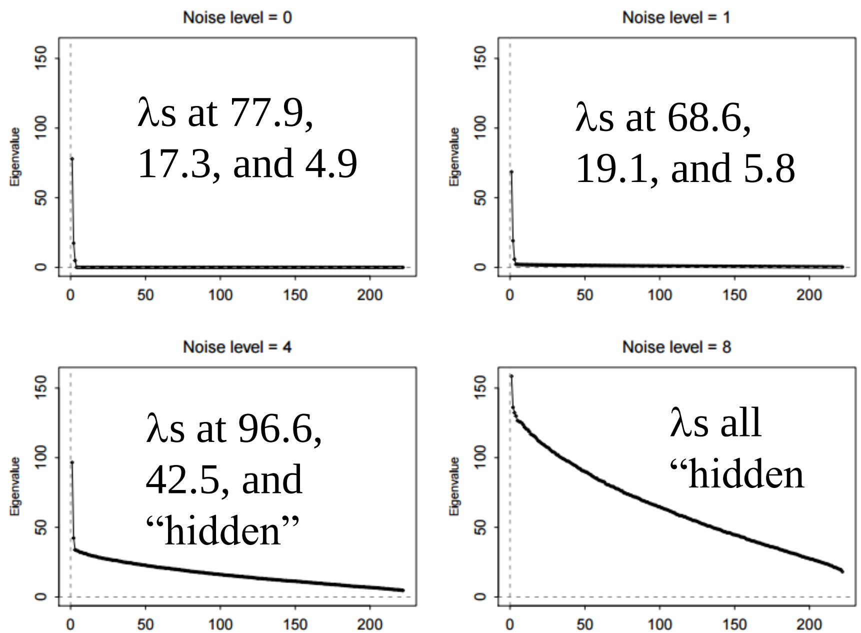
- There is noise in real data and noise due to data analysis
  - Affects the eigenmodes
  - Affects the eigenvalues
    (here actually in ratio 16:4:1)
    - Two ways to look at this
      - Scree plots: λ_i vs. i
      - Histogram of eigenvalues: Marchenko-Pastur plot
  - Misleads due to corruption and hiding of the eigenvalues
    - See next slide
- Choi worked on new methods based on Silverstein extensions to Marchenko-Pastur



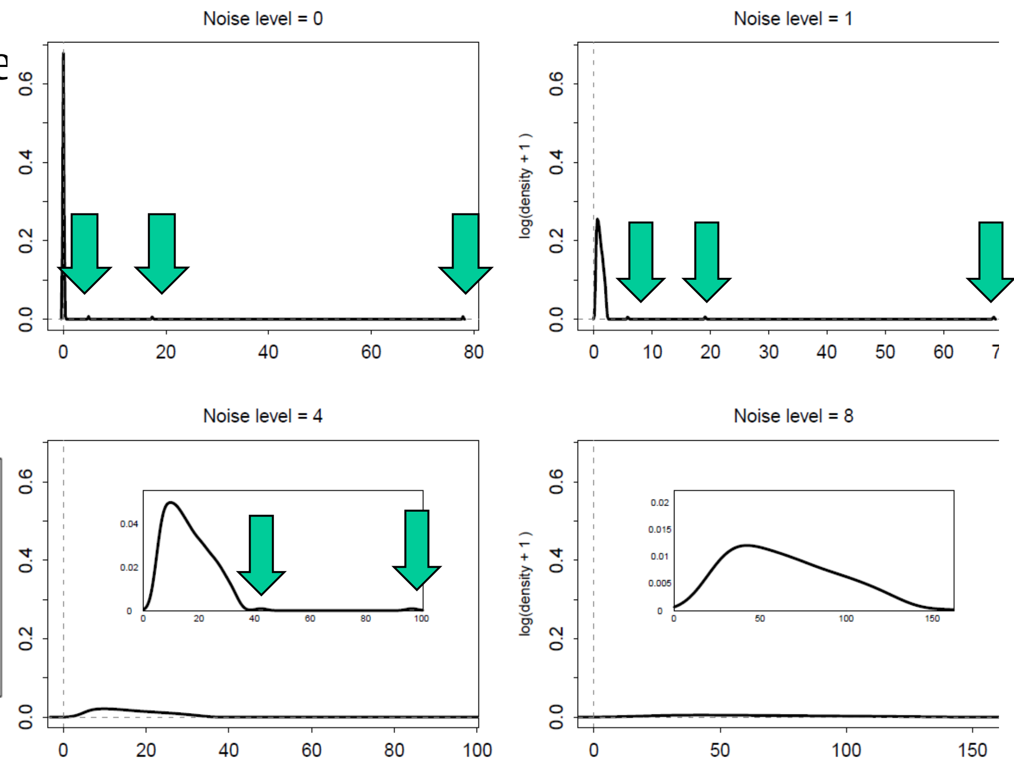


Effect of noise on principal variances

- Modifies the real eigenvalues
  - Apparent increase in total variance
  - Apparent change in fraction of total variance
- Changes the eigenvectors
- Masks the real eigenvectors



Eigenvalue vs. index (Scree plot), as noise level increases

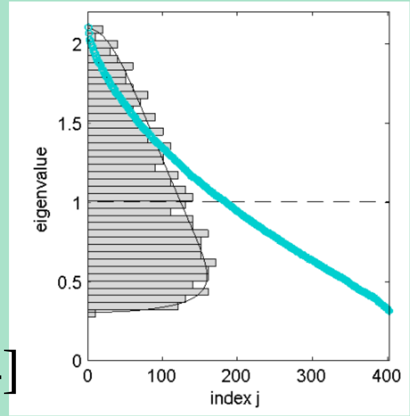


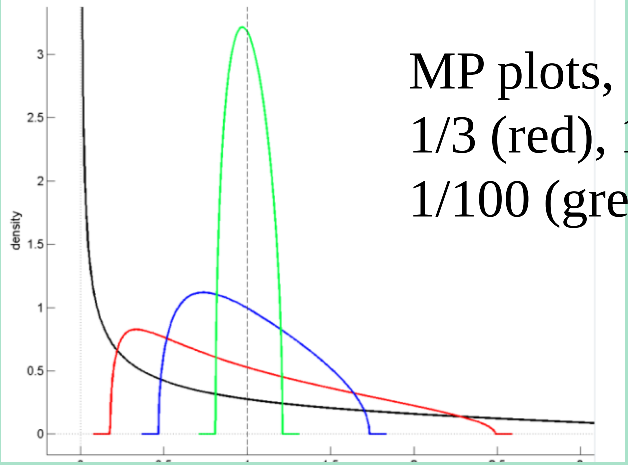
Histogram of eigenvalues (insets have larger ordinate scale) vs. as noise level increases

Estimation of # of modes by Marchenko-Pastur Analysis

- [Choi & Marron, see Marron & Dryden bibliog.]
  - "Estimation of the number of spikes using a generalized spike population model"
  - Can estimate noise from the Marchenko-Pastur histogram
  - When eigenvalues are high enough to be "pure noise"
  - Remaining problem, how to compute the eigenmodes as if unpolluted by noise

Scree plot vs. MP plots for IID indep noise [M&D, Ch 14]



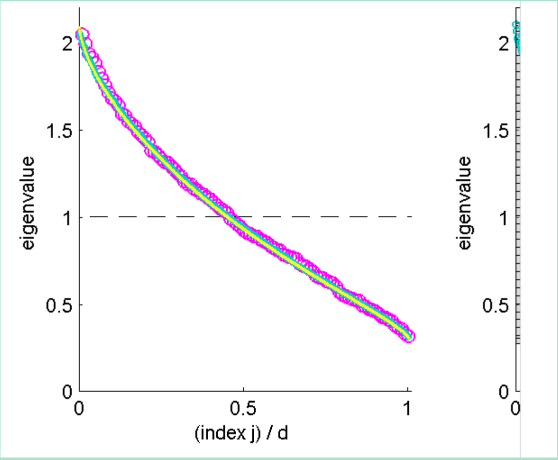


MP plots, d/n =1 (black), 1/3 (red), 1/10 (blue), 1/100 (green)

Applicability of Marchenko-Pastur Analysis

- HY Choi: transformations of PNS modes to allow MP analysis
- DIVAS [Prothero, …, Marron]
  - Study is on multiple entities (see later)
  - Independence across the features, not all that important
  - Independence among the instances of each feature!
  - Not sensitive to the particular probability distribution
  - Will see this later for multiple objects

Scree plots: d/n =1/5 for different noise distributions



Applicability of Marchenko-Pastur Analysis on PNS modes [Choi]

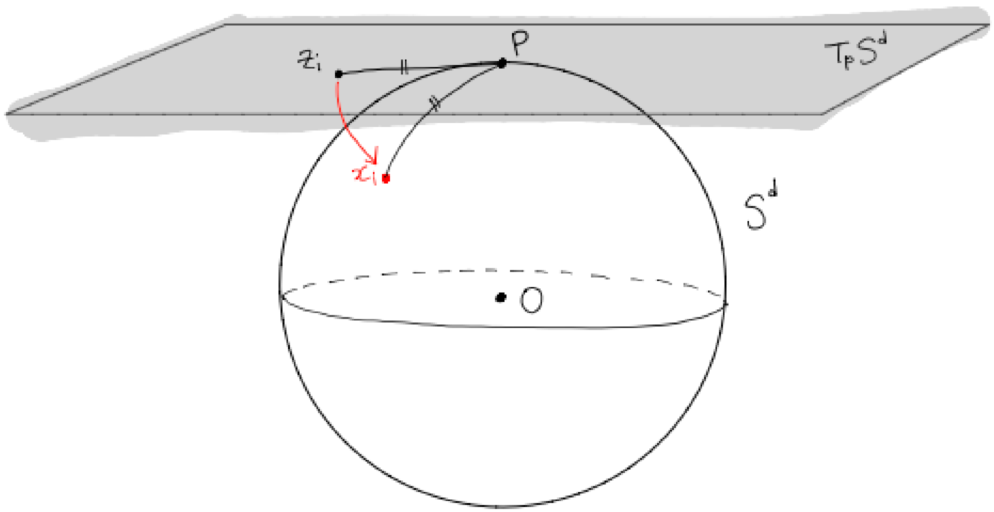
- Our “eigenanalysis” is via PNS, i.e., on sphere
- Hyo Young Choi studied this problem
  - By producing derived features from the sphere that do follow the MP distribution
    - But she only studied great subspheres, whereas the common analysis uses small subspheres, when hypothesis test supports that
  - Her pdf’s are on the course’s google drive
  - See next slides
- But the open problem remains: How to select the eigenmodesde by depolluting the MP-plot, given noise properties from high eigenvalues which show ~pure noise
  - Besides dealing with the actual subsphere approach

Applicability of Marchenko-Pastur Analysis on PNS modes [Choi], 2

Underlying noise distribution

In this work, the noise (x_i) are generated by a d-dimensional normal distribution with zero mean and variance $(\sigma^2/d)I_d$ and projected onto S^d by the exponential map. ($i = 1, \cdots, n$)

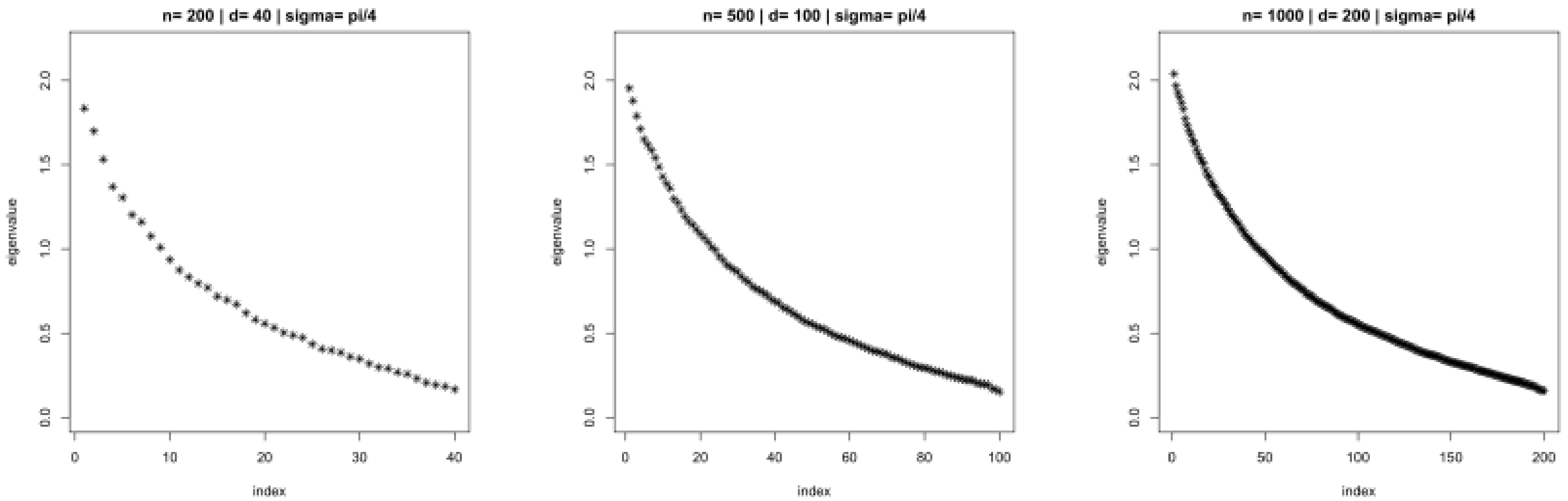
- WLOG, $\mathbf{z}_i \sim N_d(0, \frac{\sigma^2}{d}\mathbf{I}_d)$ on the tangent space at $p = e_{d+1} = (0, \cdots, 0, 1)$, denoted by T_pS^d.
- $x_i = \mathrm{Exp}_p(\mathbf{z}_i)$ where $\mathrm{Exp}_p(\mathbf{z}) = \left(\frac{\sin(\|\mathbf{z}\|)}{\|\mathbf{z}\|}\mathbf{z}^T, \cos(\|\mathbf{z}\|)\right)^T \in S^d \quad (\mathbf{z} \in T_pS^d)$



Applicability of Marchenko-Pastur Analysis on PNS modes [Choi], 3

PNS noise eigenvalues

Here, we fix the ratio $\frac{d}{n} = 5$ and let d and n grow.



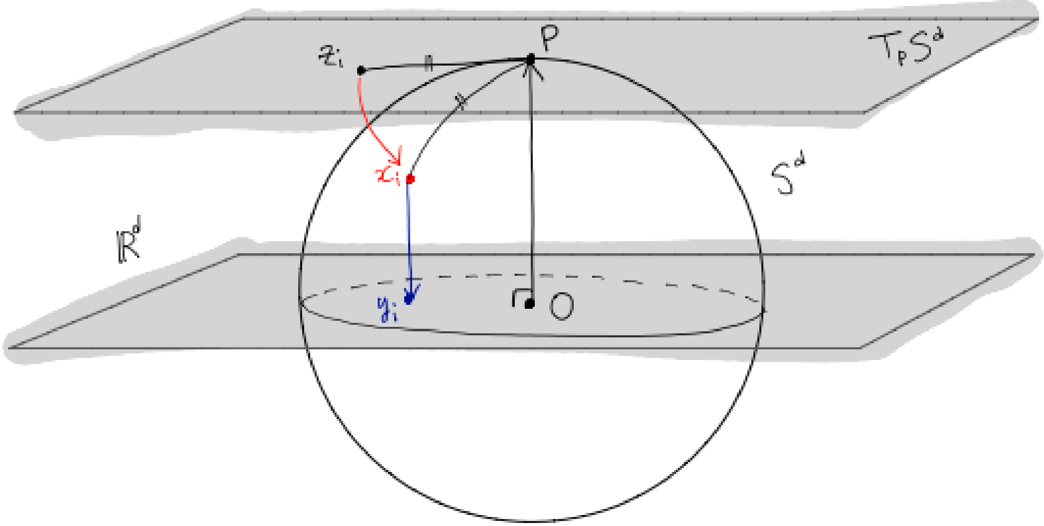
Applicability of Marchenko-Pastur Analysis on PNS modes [Choi], 4

We assume ...

- Only great spheres are selected. That is, $\hat{r}_k = \pi/2$ for $k = 1, \cdots, d-1$.
  - Then, PNS scores can be recursively expressed as

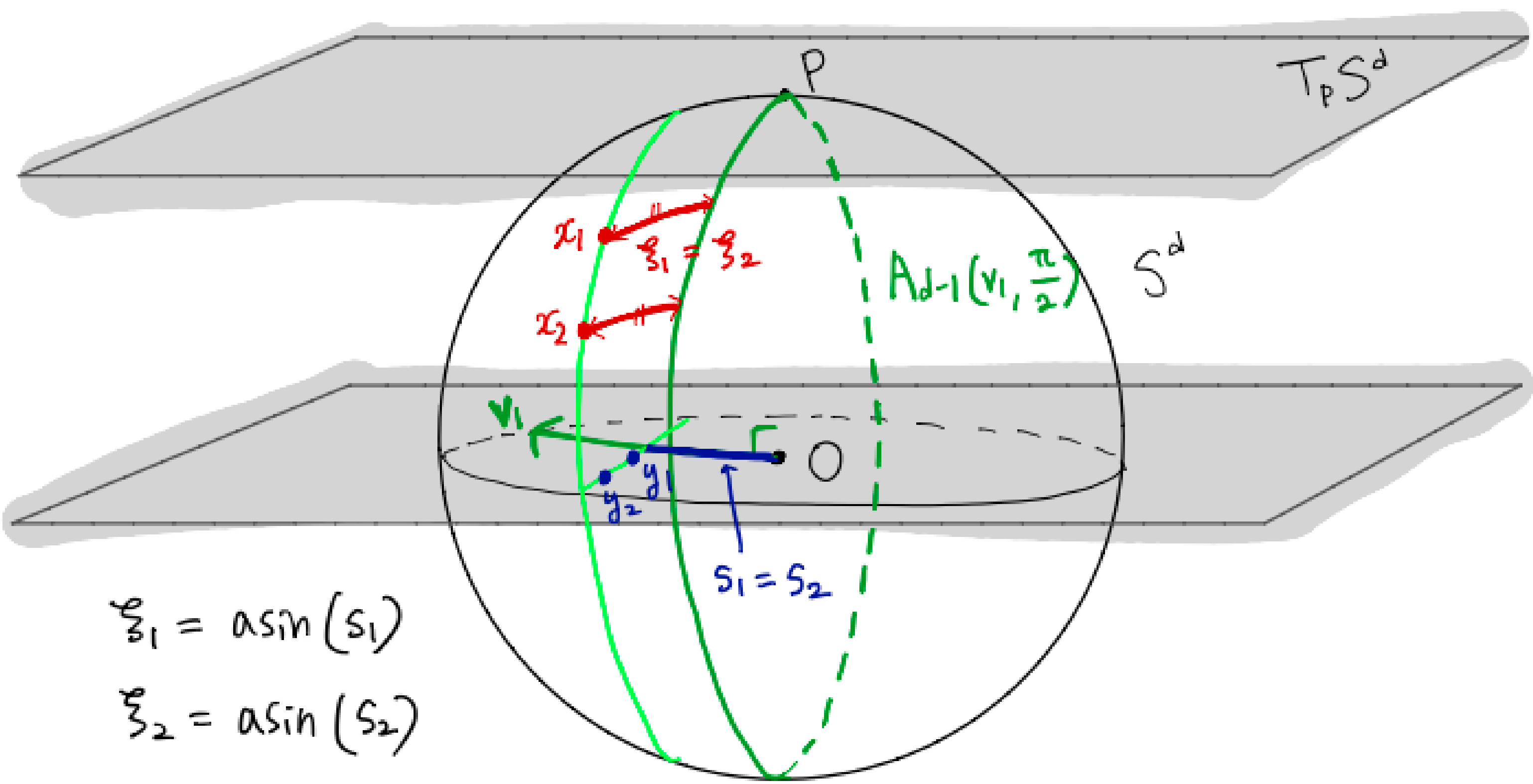
$$\xi_{i,k} = -\mathrm{asin}\Big(\frac{\hat{v}_k^T x_i}{\prod_{j=1}^{k-1}\cos(\xi_{i,j})}\Big) \qquad \text{for} \quad i = 1, \cdots, n.$$

- PNS mean is the true mean (the north pole).
  - This assumption is used to connect modes of variation in the spherical data $\{x_i\}_{1\le i\le n}$ to modes of variation in the orthogonally projected data $\{y_i\}_{1\le i\le n}$ onto $\mathbb{R}^d$.



Applicability of Marchenko-Pastur Analysis on PNS modes [Choi], 5

Orthogonal projection



Applicability of Marchenko-Pastur Analysis on PNS modes [Choi], 6

Orthogonal projection - special case

- Consider the PCA on y_i's with the uncentered sample covariance matrix, $\tilde{\mathbf{S}}_n = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T$.
- Denote the resulting eigenvectors, eigenvalues, and PC scores by $\tilde{v}_k$, $\tilde{\lambda}_k$, and $s_{i,k} = \tilde{v}_k^T y_i$, respectively. $(k = 1, \cdots, d,\ i = 1, \cdots, n)$
- Suppose that the PNS eigenvectors and PCA eigenvectors are identical, i.e. $\hat{v}_k = \tilde{v}_k$ for $k = 1, \cdots, d$.

Then,

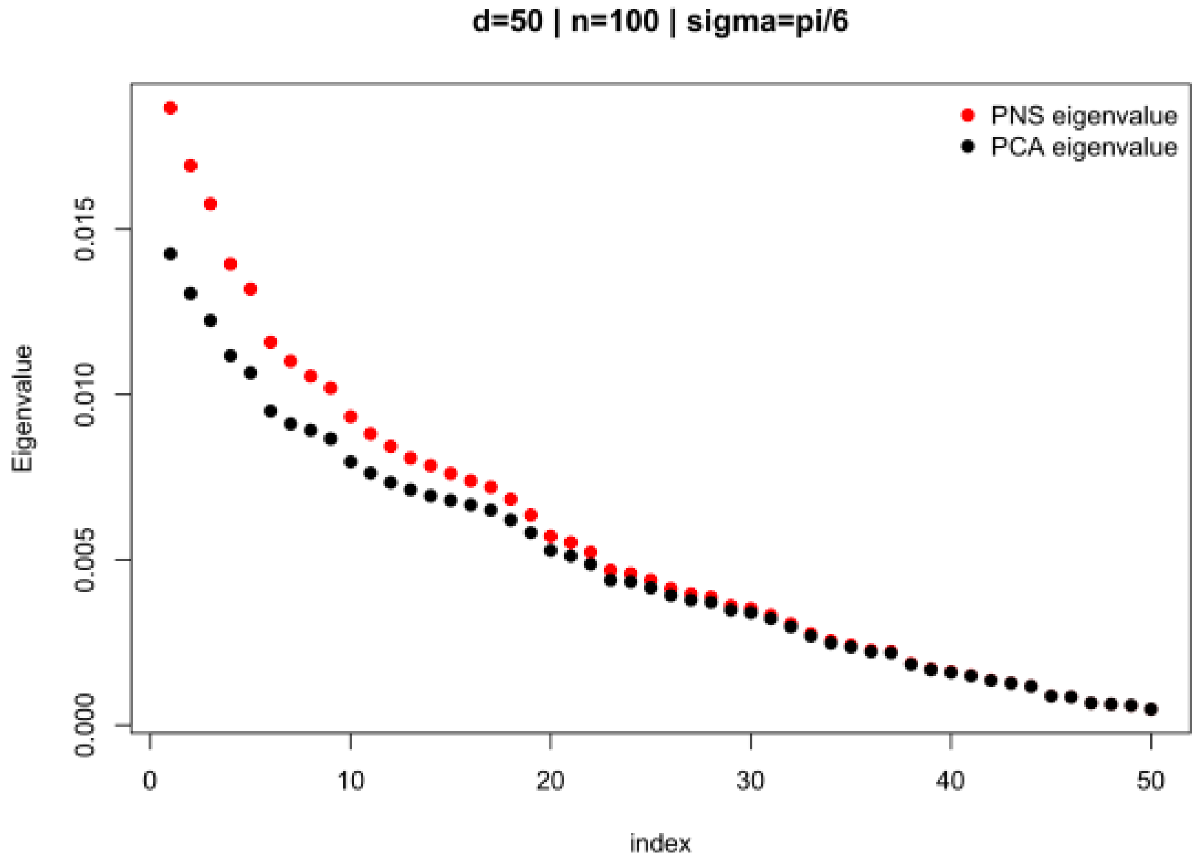
$$\xi_{i,k} = -\mathrm{asin}\left(\frac{s_{i,k}}{\sqrt{1 - \sum_{j=1}^{k-1} s_{i,j}^2}}\right)$$

and

$$\begin{aligned} \tilde{\lambda}_k &= \frac{1}{n}\sum_{i=1}^{n} s_{i,k}^2 \\ \tau_k &= \frac{1}{n}\sum_{i=1}^{n} \xi_{i,k}^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{asin}^2\left(\frac{s_{i,k}}{\sqrt{1 - \sum_{j=1}^{k-1} s_{i,j}^2}}\right). \end{aligned}$$

Applicability of Marchenko-Pastur Analysis on PNS modes [Choi], 7

PNS eigenvalues vs PCA eigenvalues from y_i's



Applicability of Marchenko-Pastur Analysis on PNS modes [Choi], 8

Preliminary results

Define

$$\beta_k = \frac{\tau_k}{\frac{1}{n}\sum_{i=1}^{n}\left(\prod_{j=1}^{k-1}\cos^{-2}(\xi_{i,j})\right)} \quad \text{for} \quad k = 1, \cdots, d-1.$$

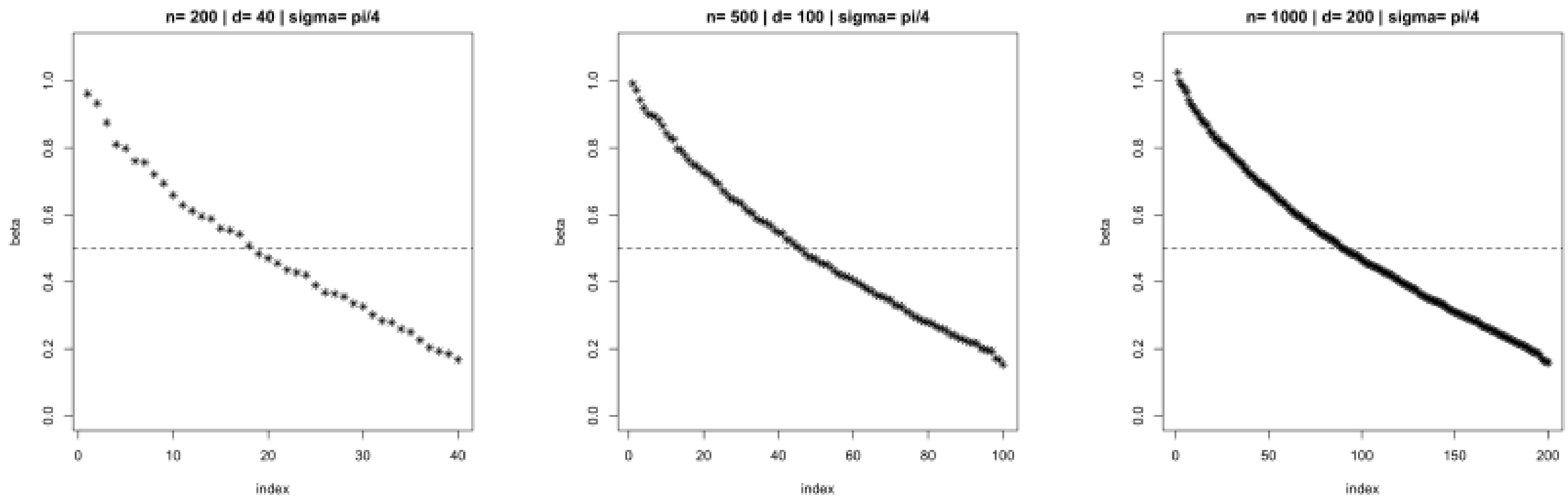
Then, the empirical distribution of $\{\beta_k\}$ converges to the M-P law with parameters y and $\phi = \sin(\sigma)$, i.e.,

$$\frac{1}{d}\sum_{j=1}^{d}\delta_{\{d\beta_k\}} \to F_{y,\phi^2}.$$

Applicability of Marchenko-Pastur Analysis on PNS modes [Choi], 9

Scree plot for beta's

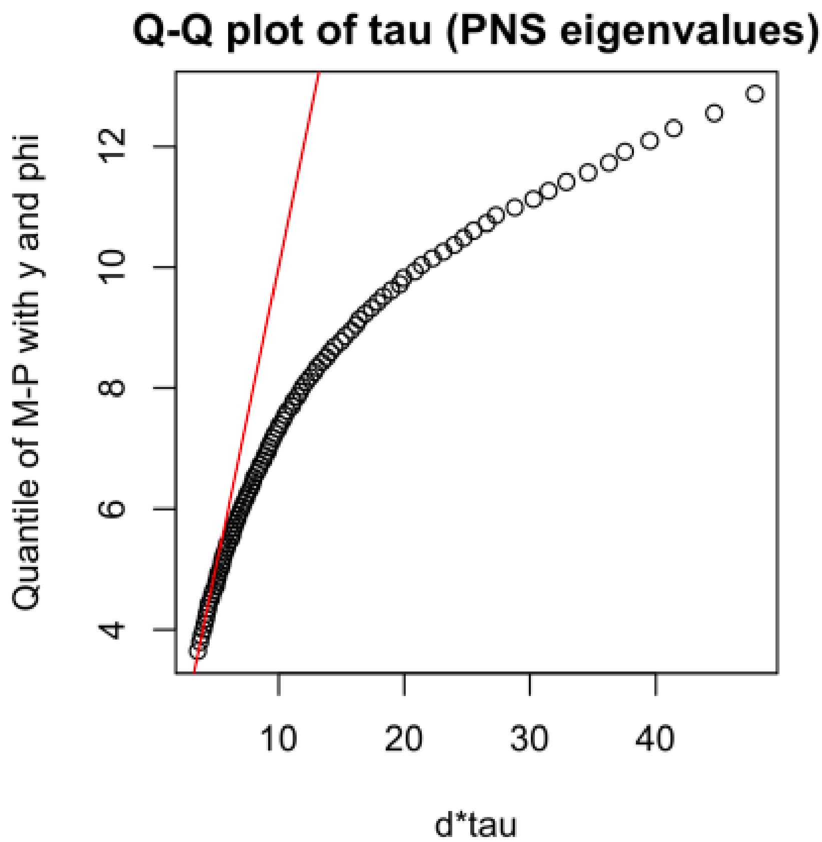
Here, we fix the ratio $\frac{d}{n} = 5$ and let d and n grow.

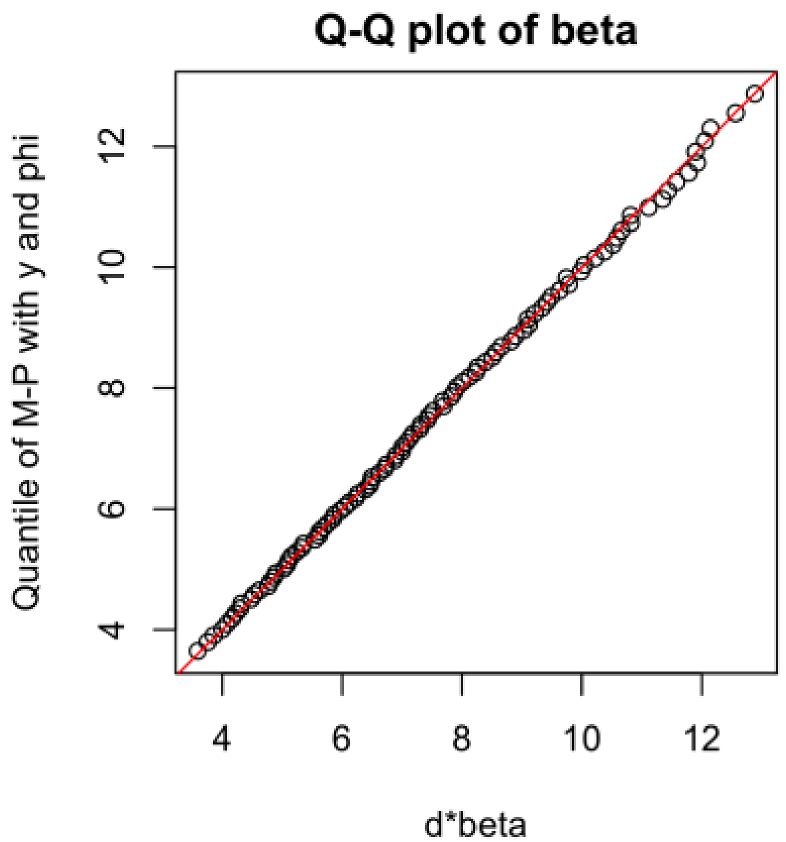


Applicability of Marchenko-Pastur Analysis on PNS modes [Choi], 10

Does it actually work?

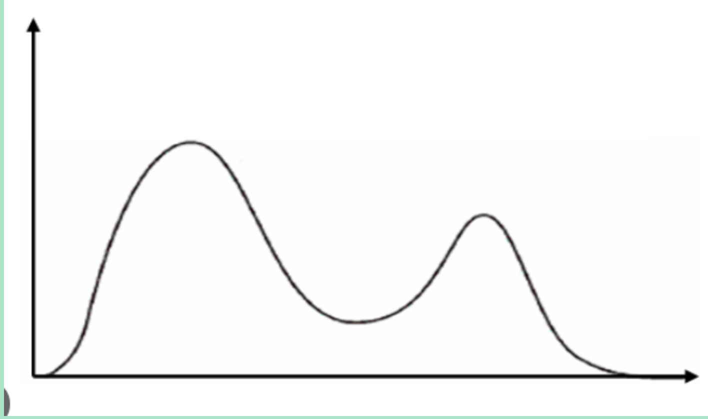
$n = 100$, $d = 1000$, $\sigma = \frac{\pi}{3}$





Correlation

- Covariance corrected for scale by dividing each feature by its standard deviation
- Each variable is normalized by its standard deviation
- Careful: zero correlation does not mean statistical independence
- Careful about its meaning in multimodal distributions
  - Non-independence may come in which mode the random variable falls



Gaussians

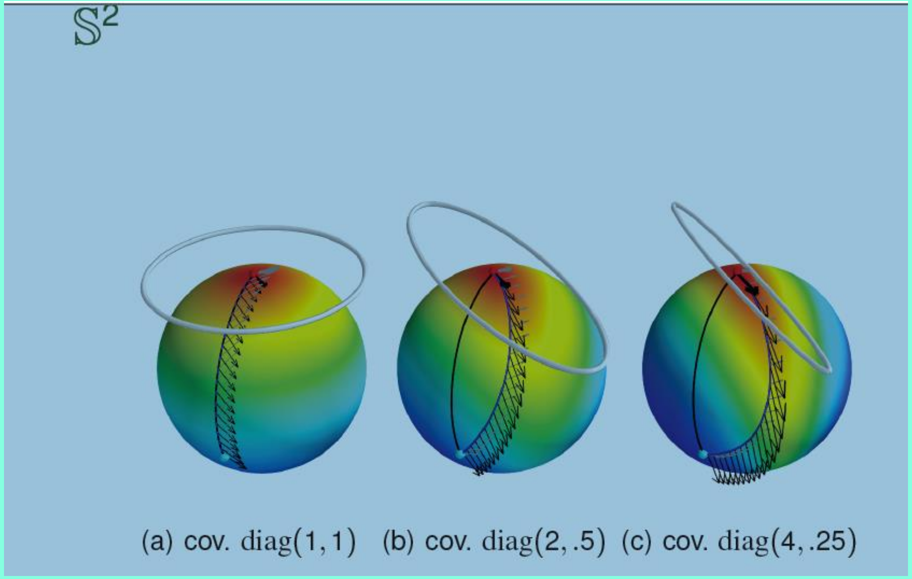
- Importance
- Relation to PCA
- Production
  - Analytic form in Euclidean space
    - Given Principal frame R and eigenvalues Λ
  - Diffusion: $\partial f(\underline{x},t)/\partial t = \nabla^2 f$ with $f(\underline{x},0)=\delta(\underline{x})$ with $t=\sigma^2/2$
  - Brownian motion
- On curved surfaces, esp. spheres
  - Wrapped Gaussian
  - von Mises distribution
    - ~ wrapped Gaussian on sphere
    - With an analytic form not needing sums over wraps
    - Most commonly used due to its rather simple form
  - Brownian motion (random walks)
    - Moving frames [Sommer]

von Mises Distribution

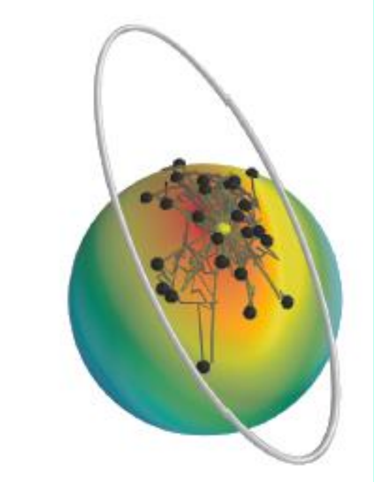
- See Wikipedia
- On spheres of arbitrary dimension
  - ~ wrapped Gaussian
- Maximum entropy distribution for data on circle (thought of as complex) for which the first real and imaginary circular moments (coefficients of Fourier terms of base frequencies) are specified
- Formula as function of latitude θ from north pole
  - $p(\theta \mid \mu, \kappa) = \exp(\kappa \cos(\theta - \mu))\ (/(2\pi I_0(\kappa))$
  - $\kappa \sim 1/\sigma^2$
  - $I_0(\kappa)$ is modified Bessel function of the first kind of order 0

Gaussian Probability distributions on curved manifolds [Sommer]

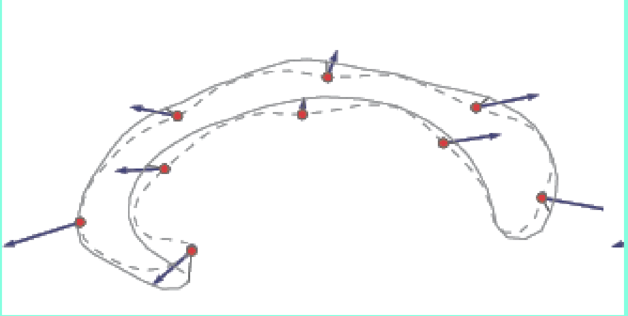
- Augment the manifold with a local rolling tangent frame; probabilistically analyze on that
  - Then project back down to the manifold
- Yields probability densities on paths
  - That is on shape changes! (how to analyze p(shape change) has been an open question

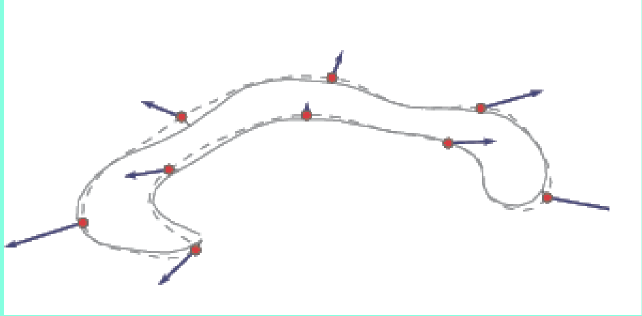


Results of [Sommer] Path Probabilities on a Toy Problem



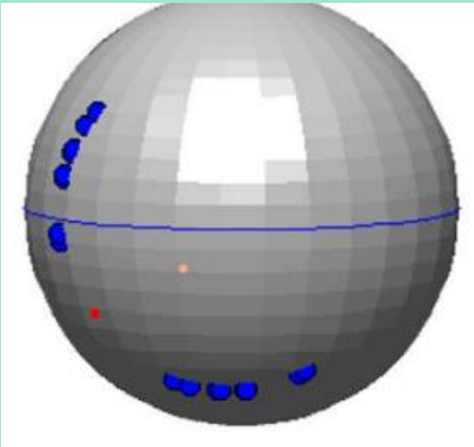
- Corpora callosa by 9-pt PDM
- 10 training cases
- Mean and covariance computed together from these paths weighted by their probabilities (ML)
- Vectors are momenta from most probable path from training case (dashed) to mean (solid) curves





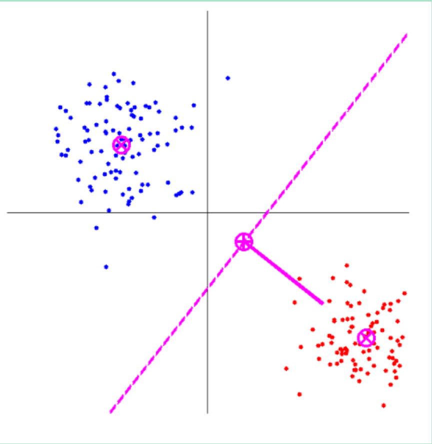
Hypothesis Testing

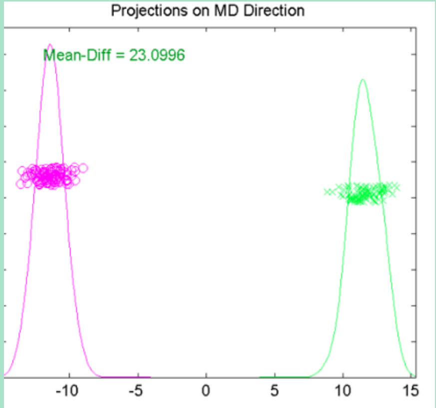
- To test differences between classes
- To evaluate statistical methods
  - For example, classification methods
- To make choices in statistical methods
  - For example, between great sphere subdimensional space and small sphere subdimensional space at a dimension-lowering step of PNS
    - Small subsphere is best when it is clear that the data falls into that form
    - Great subsphere is best when the data is in the form of a cluster



Permutation Tests for Hypothesis Testing

- Consider a classification procedure producing a separation direction in feature space
  - For now a vector **v** in Euclidean space
    - But can be a geodesic on a curved space
  - With ability to project cases (feature tuples) on the separation geodesic
    - Along geodesics to the separation geodesic
  - For each class produces a histogram
  - The question is how statistically significant the difference between the histograms is

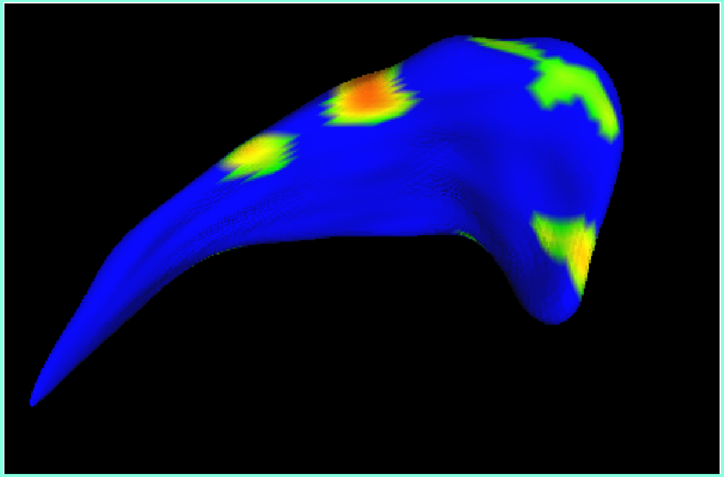




- Done by permutation tests
  - By geometric object property (GOP)
  - Correction for multiple tests (GOPs)
  - Evaluating a classification
    - DiProPerm [Marron & Dryden, sec 13.1]

Objectives for Hypothesis Testing on (Geometric) Features z

- **Characterization of differences** between classes
  - By <u>hypothesis testing</u> on null hypothesis: **no difference**
  - What kind of (geometric) differences?
  - Where are there (geometric) differences?



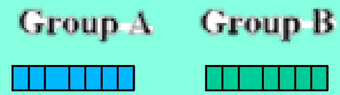
Hypothesis Testing

- **Null hypothesis:** two classes are the same in regard to features
- **Reject null hypothesis** if observed interclass differences are appropriately unlikely to happen by chance under null hypothesis
  - Choose the P_0 value such that if the observed interclass deviation or more happens at random under the null hypothesis with probability $P < P_0$, we will reject the null hypothesis

Univariate Hypothesis Testing

(to determine significant differences)

- Test is by **DOUBLE NEGATIVE**:
  Reject (null) hypothesis *that* the two classes are *not different*
  - I.e., reject hypothesis that two classes are the same, i.e., observed differences come from random sampling
- Typically tests on magnitude of differences between class means: m_A and m_B
  - $T_0 = | m_A - m_B | / (s_A^2/n_A + s_B^2/n_B)^{½}$
  - Test with significance threshold y: If T´ s.t. $1\text{-}p_{T'}(T \leq T') = \gamma$ and $T_0 \geq T'$, then reject
    - "t-test", depends on Gaussian assumption within each class
  - Other measures of difference are possible, esp. for parameter conglomerates



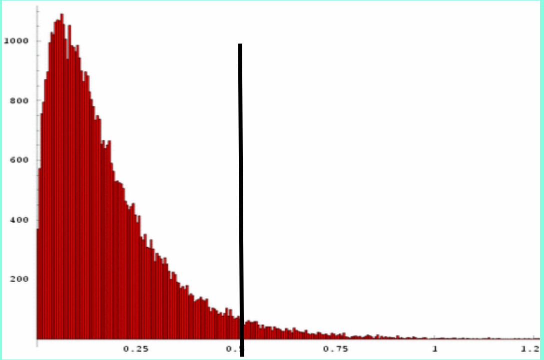
$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

"A t-test is a statistical test in which the test statistic follows a student t-distribution if the null hypothesis is supported" [Wikipedia]

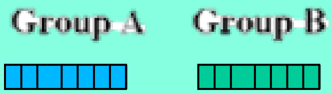
Univariate Hypothesis Testing by Permutation Test

- Still test on magnitude of differences between class means: m_A and m_B
  - $T_0 = | m_A - m_B | / (s_A^2/n_A + s_B^2/n_B)^{½}$
  - But create distribution of T under null hypothesis empirically
    - With that distribution see percentile of T_0 in that distribution: p-value =

      #Perms larger / #Perms total

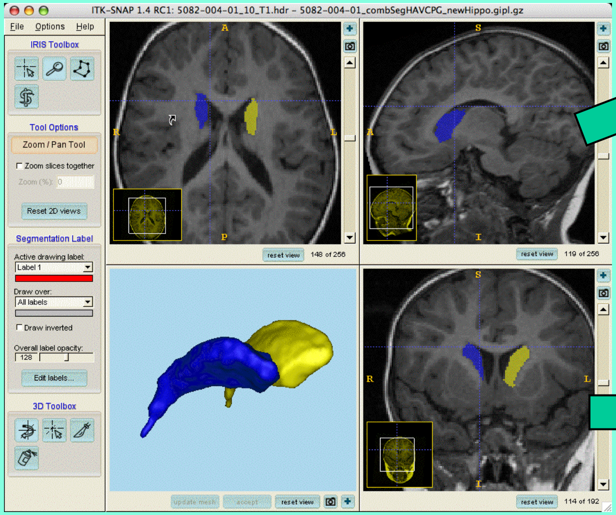


- Under null hypothesis Group A and Group B are not different, so any permutation into a new pair A, B is equivalent.
  - Each pair produces a T (normalized mean diff.)
  - Over all permutations produces the empirical T distribution



Concept: Shape Analysis

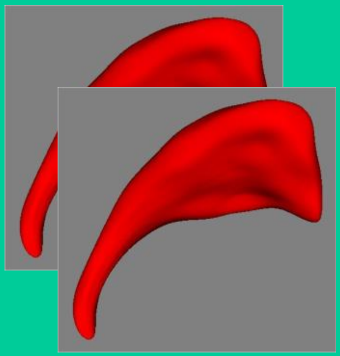
- Traditional analysis: e.g., regional volume
- Other possibility: Local analysis



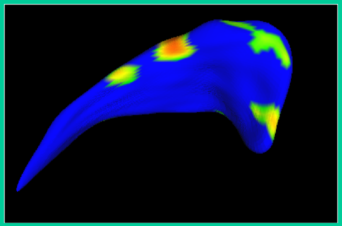
Binary Segmentation

Volumetric analysis: Size, Growth

Shape Representation



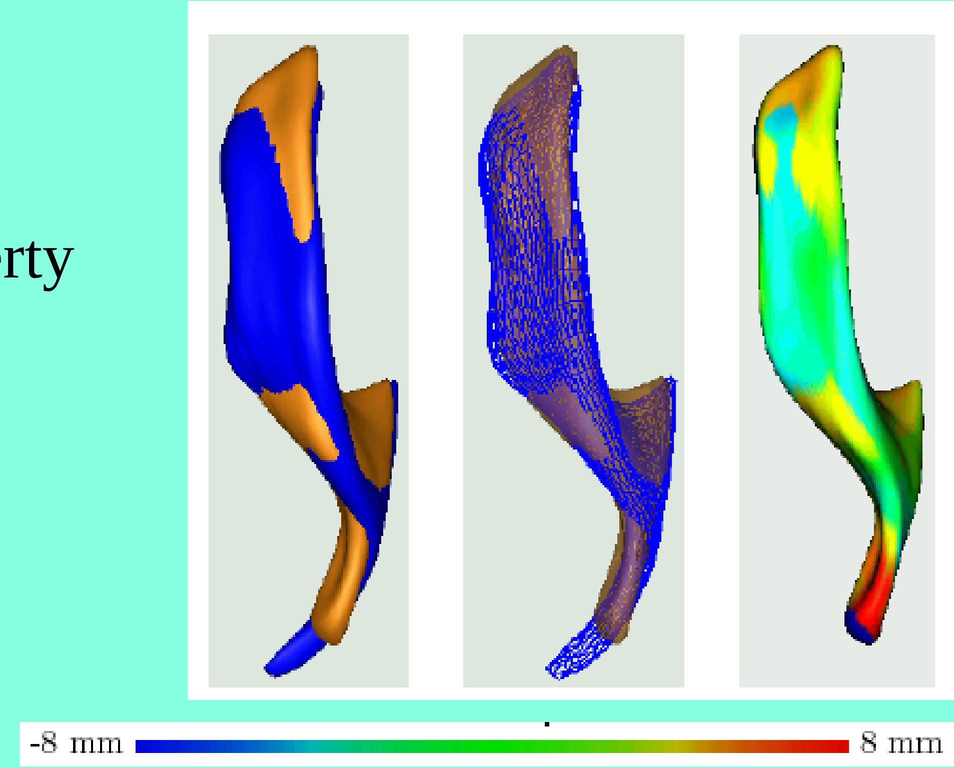
Local Statistical analysis



Slide: M. Styner

Statistical Analysis of Shape

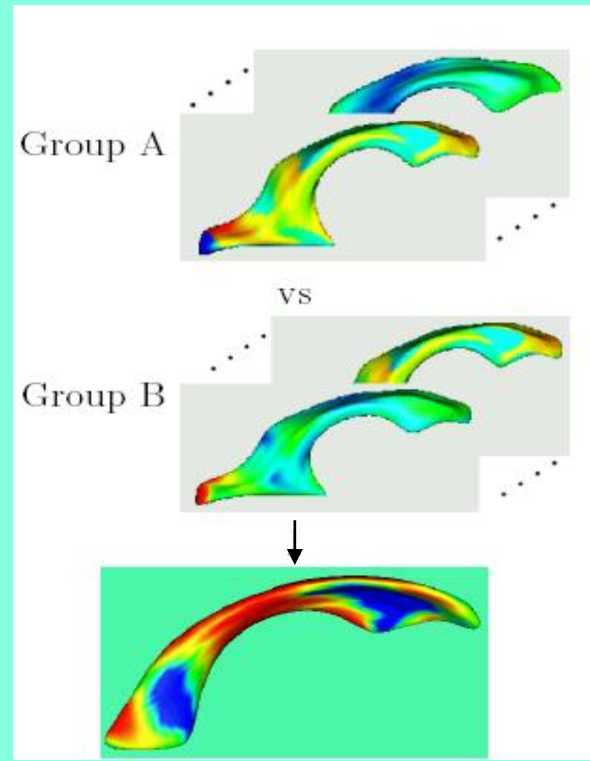
- "Locations" with
  - Correspondence
  - Pose normalized or pose-free
- Analyze Geometric Object Property
  - GOP per location
  - Univariate, e.g.,
    - Log Thickness
  - Multivariate, e.g.,
    - Point locations (x,y,z)
    - Deformation displacement vector
      - Log length, PNS pair
        for direction
    - S-rep GOPs, e.g., fitted frame rotations
- Multivariate + Global shape features, e.g., volume



Slide: M. Styner

Hypothesis Testing for Multiple Features

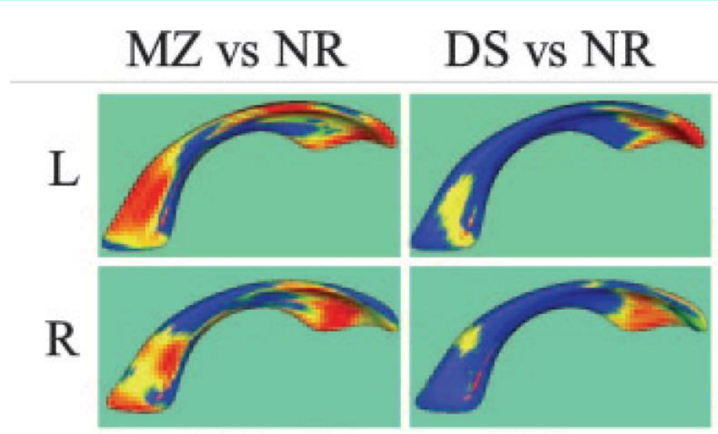
- Hypothesis test for each location/feature set:
  - 2 groups: Is the mean of group A different from the mean of group B?
    - E.g., Schizophrenia group vs Control group
  - **P-value of group mean difference**
  - Significance map
  - Threshold α = 5%, 1%, 0.1%
  - Or z value (difference from mean/std. deviation)
- Parametric distribution, e.g. Gaussian, or
- Non-parametric: model free
  - **P or z value** directly from observed distribution
  - Distribution estimation via permutation tests
  - **But need correction for multiple comparisons**
    - Otherwise, test on some variables will be randomly significant



Slide: M. Styner

Many features, many, too many ...

- Many local features computed independently
  - 1000 - 5000 features
- Even if all features would be pure noise, many locations would be computed as significantly different
- Overly optimistic ⇒ Raw p-values
  - High sensitivity, low specificity
- But overly pessimistic ⇒ Correction based on no inter-GOP correlation
  - That is, FWER or FDR correction
- Multiple comparison problem
  - P-value correction
  - General Linear Mixed Modeling
    - Model covariance structure
  - Random Field Approaches (Worsley)
    - Many assumptions, smoothing



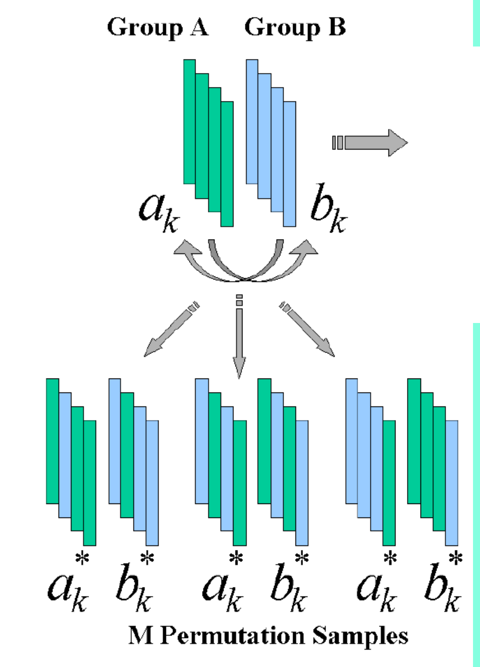
Hypothesis Testing with Locality

- **Goal:** given training data sets of objects $\underline{\mathbf{z}}^k$ in two classes, determine whether there are significant differences between the classes and if so, where
  - Where (locality): positions or other parameters
  - Training data: $\{\underline{\mathbf{z}}^k \mid 1 \leq k \leq n_{A;}\ n_A+1 \leq k \leq n_A + n_B\}$

- **Method:** Hypothesis test with initial Geometric Object Property(ies) (GOPs) at each location
  - A GOP may be a tuple, e.g., object normal direction
  - With corrections for multiple comparisons, which will lead to a different threshold for each location × GOP
    - Commensuration by turning T values into p values
    - Making p values for each feature std Gaussian
    - Then decorrelation via PCA

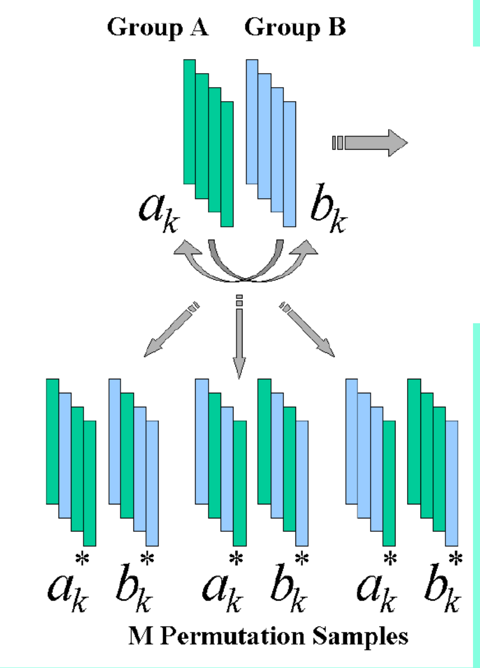
Permutation Hypothesis Tests on >1 Variable

- Estimate variable's distribution
  - Permute group labels
    - N_{a}, N_{b} in Group A and B
    - Create M permutations
    - Compute means μ_{ja}, μ_{jb} within perms, j = GOP index, and thus $|\mu_{ja} - \mu_{jb}|$ within perms
      - Use backwards means for non-Euclidean GOPs
    - Not ready to histogram because
      - GOPs are incommensurate
      - Some GOPs are non-Euclidean
      - Each GOP has its own distribution
      - GOPs are correlated



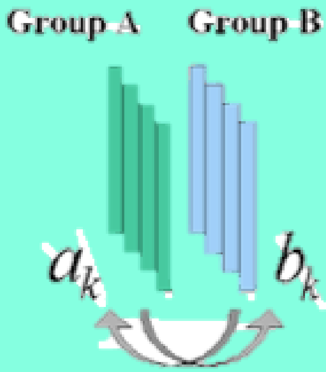
Permutation Tests on >1 Variable
μ-diff fixup

- Commensurate by transforming each mean difference into a probability via its histogram
- Make distributions same and joint distribution Gaussian by turning each distribution into a standard Gaussian
    - Two cumulative histogram transformations: quantiles have uniform probability
        - Cumulative dist. of v'ble → uniform
        - Quantile f'n of normal → st'd Gaussian
- What's left is handling covariance of transformed variables



Hypothesis Testing for Shape- or Appearance-based Diagnosis

- **Multiple tests w/ high correlation** (locations, local shape features)
  - Some features live on spheres – need backward means, others are Euclidean
  - Correlation, esp. between nearby location
  - Allows much less pessimistic multiple test correction
  - Markov approaches are possible
- **General approach**, using permutation tests:
  - Analyze each feature separately to map it onto a standard normal Euclidean variable
    - PNS allows this for variables on spheres
  - Compute covariance matrix of mapped variables
  - Permutation test overall or on variables modified to have no correlation with the others

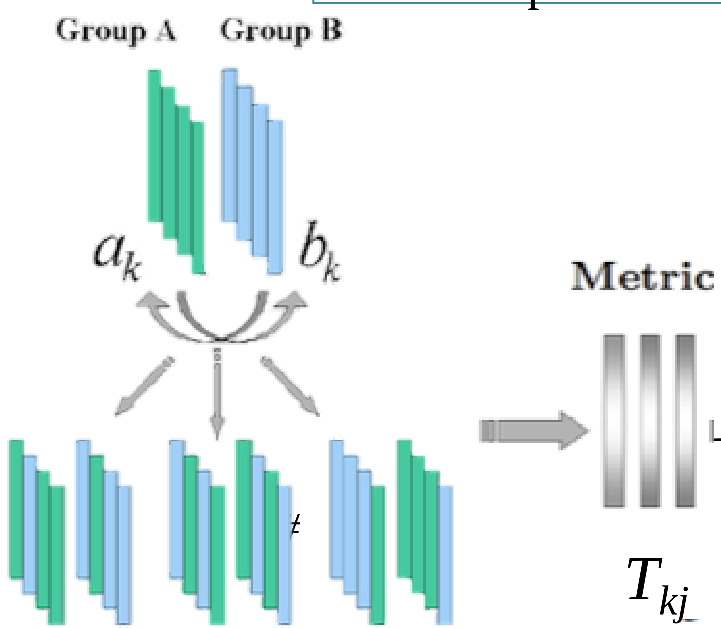


Steps in non-linear global permutation test

Global test: on
$M_k = U_k^T \Sigma_U^{-1} U_k$
cumulative histogram
Local test: on marginal from nonisotropic Gaussan w/ Σ_U

Form matrix U from $\mathbf{U}_k$ tuples
Form covariance matrix: Σ_U

Map uniform distribution to normal distribution
Φ is a normal distribution c.d.f.
$U_{kj} = \Phi^{-1}(p_{kj} - 1/2N)$ (in Std. normal dis)



Compute cumulative hist C_j of $T_{\_j}$ over all permut'ns k

For each permutation and feature, get p-value
$p_{kj} = C_j^{-1}(T_{kj})$
p_{kj} *is in [0,1]*

create N permutations

For each located-feature j, permutation k

$T_{kj} = d(\hat{\mu}_{akj}, \hat{\mu}_{bkj})$ *(CPNS d, backwards mean)*, also T_{0j}

Decorrelating the Standard Normal Variables

- Note that all hyperplanar (including dimension 1) cross-sections through the mean of a Gaussian are Gaussian
  - Also, the principal dimension-1 cross-sections yield uncorrelated variables
- Σ_U is made up or correlated standard 1-dimensional Gaussians
  - So the cross-sections are not principal (nor orthogonal)
- Use PCA on Σ_U to produce new, uncorrelated variables formed as the eigenvector directions
  - Multiple test correction assuming non-correlation is applicable
  - For shapes (U formed from GOPs) there will be few such variables (ones with low eigenvalues can be cut out)

P-value correction via FDR or FWER

- False Discovery Rate (FDR)
  - More relaxed assumptions
    - More power than Bonferroni, higher specificity than uncorrected
  - Used in fMRI, VBM and Deformation field analysis
- FDR: Proportion of false positive tests among those test for which H_0 is rejected
  - Bounds expected rate among those tests that show significance only.
- FWER correction: Rate of false positives among all tests, whether or not H_0 is rejected

1. Controlling False Discovery Rate: A practical and powerful approach to multiple testing, Y Benjamini, Y Hochberg, J.R. Stat Soc Ser B 57 1995
2. Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate, CR Genovese, NA Lazar, T Nichols, NeuroImage 15 2002

Slide: M. Styner

FDR: Definition

Classifications of Voxels in V Simultaneous Tests

	Declared active (discovered)	Declared inactive (not discovered)	
Truly active	V_{aa}	V_{ai}	T_a
Truly inactive	V_{ia}	V_{ii}	T_i
	D_a	D_i	V

- V: number of tests
- Active = rejected H_0, significant test

$$\mathrm{FDR} = \frac{V_{ia}}{V_{ia} + V_{aa}} = \frac{V_{ia}}{D_a}$$

inactive declared active

all active

Slide: M. Styner

FDR

$$\text{FDR} = \frac{V_{ia}}{V_{ia} + V_{aa}} = \frac{V_{ia}}{D_a}$$

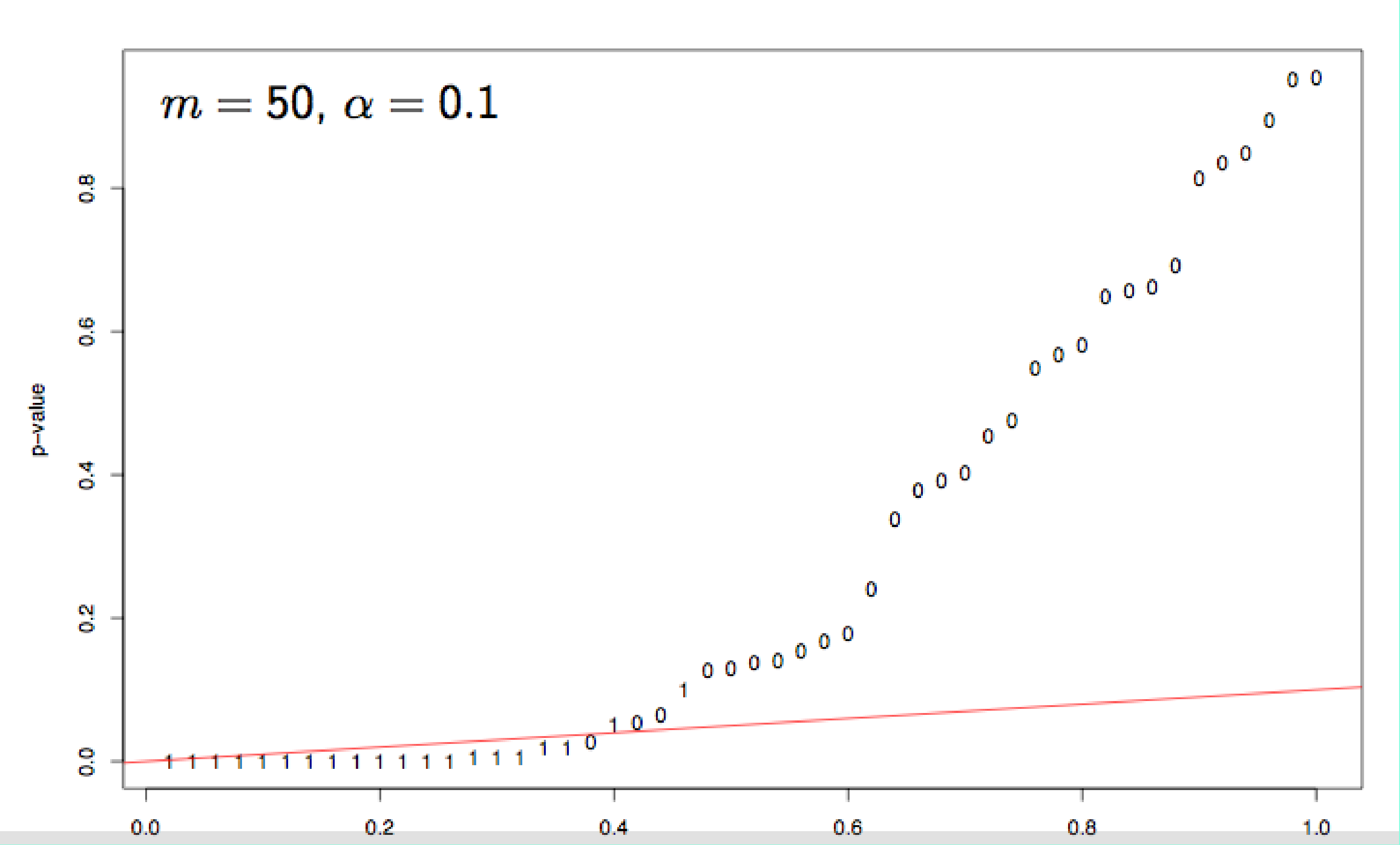
- Of course, V_{ia} is unknown, only D_a, D_i, V
- Controlling FDR: ensure that on average the FDR is not bigger than predefined proportion q.

$$\mathsf{E}(\text{FDR}) \le \frac{T_i}{V} q \le q,$$

- Benjamini and Hochberg:
- Many tests, $T_i/V \approx 1$
- Commonly: $q = 0.05$
  - Higher values are reasonable in many problems

Slide: M. Styner

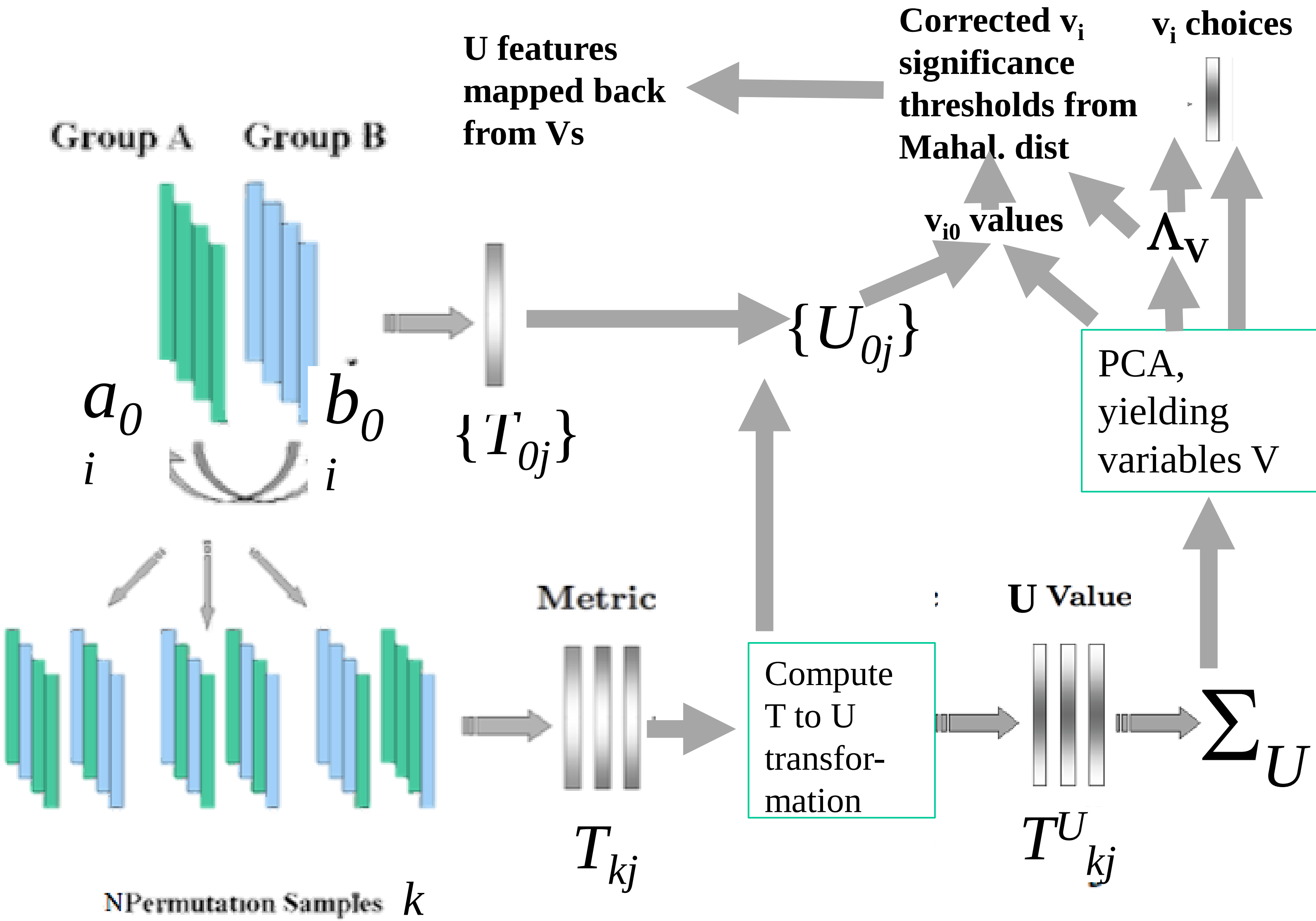
FDR interpretation



All p-values below line through origin with slope q are different/active

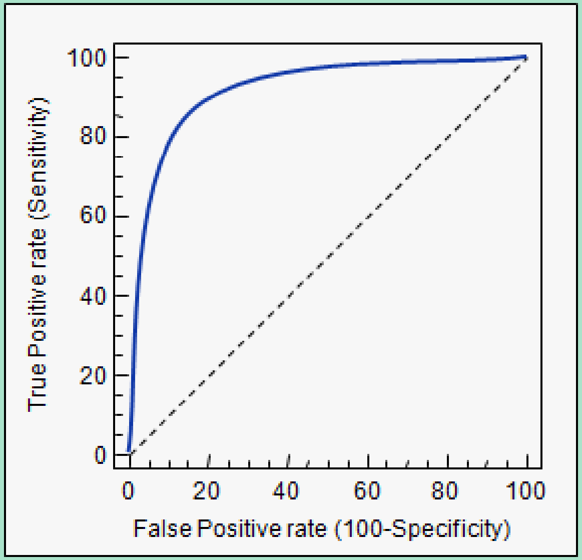
Slide: M. Styner

Corrected Multivariate Hypothesis Testing: by Permutation Tests [Pizer]



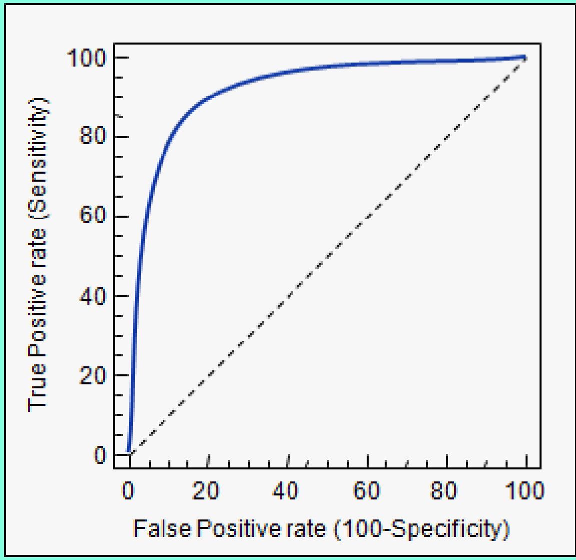
AUC as a Way to Compare Classification Methods

- ROC: Plot True Positive Rate (TPR) vs. False Positive Rate (FPR)
- Area under ROC (AUC) is error rate in a 2-alternative forced choice experiment
  - 0.5 is pure guessing, 1.0 is perfect
- How to produce an AUC?
  - By a random holdouts approach



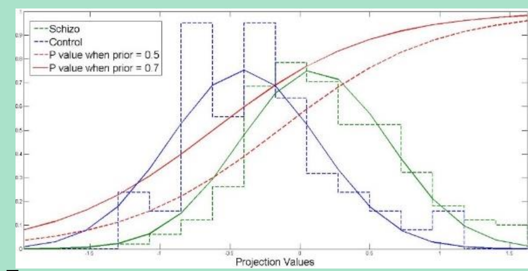
ROC Analysis

- An important means of evaluating any medical procedure involving a detection of a signal
  - Signal can be classification of voxel as the object to be segmented
- True positive rate (*sensitivity*) plotted against false positive rate (1-*specificity*)
  - TPR: P(correct decision | signal present)
  - FPR: P(wrong decision | signal absent); a measure of conservatism
  - Common measure: area under ROC (AUC)
    - Equivalent to 2-alternative forced-choice fraction correct



AUC by Random Holdouts

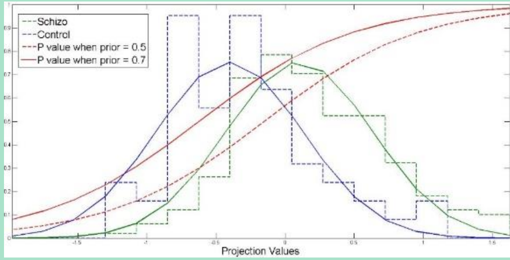
- Assume data has N_i in class i
- Pick a subdivision fraction:
  f = # training cases / N = # all cases, e.g., 80%
  - # test cases is complement
  - In each holdout # of test cases is (1-f) N_1 in class 1 and (1-f) N_2 in class 2, with the cases chosen at random
- Run the experiment many times
  - Each test case yields a d value, leading to a class choice, given a threshold
  - Yield TPR and FPR for various thresholds



- Two options for combining trials
  - Average AUCs per random set
  - Combine FPR,TPR data over sets, to yield a single AUC

Combining random holdouts results into an AUC

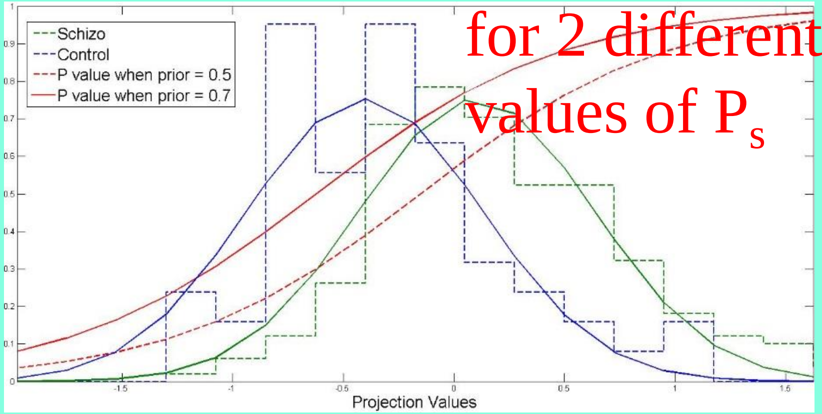
- Each test case yields a d value along separation direction
- But d values are not commensurate across random holdouts
  - They have different separation directions
- So turn d values into p values via Bayesian analysis



  - Then the set of p values for each class can be coalesced to produce 2 histograms, which can yield an ROC and thus an AUC

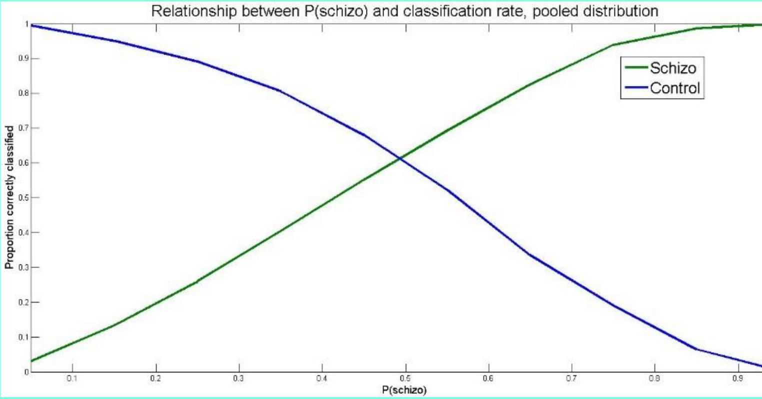
Going from histograms on d to P(schizo | D) function

P(schizo|D) for 2 different values of P_s



d

- Bayesian formulation
  - Two histograms are Gaussians with common variance from histo's P(d | schizo) and P(d | control)
    - $P(\,schizo|\,d) = \dfrac{p_s\,p(d|schizo)}{p_s\,p(d|schizo) + \,(1-p_s)\,p(d|control)}$
  - There is a parameter, P_s, the prior probability of being schizo
    - Each value of P_s yields a different P(schizo | d) function
    - Applied to test data, each value of P_s yields a different true positive rate and true negative rate
    - These rate curves yield an ROC

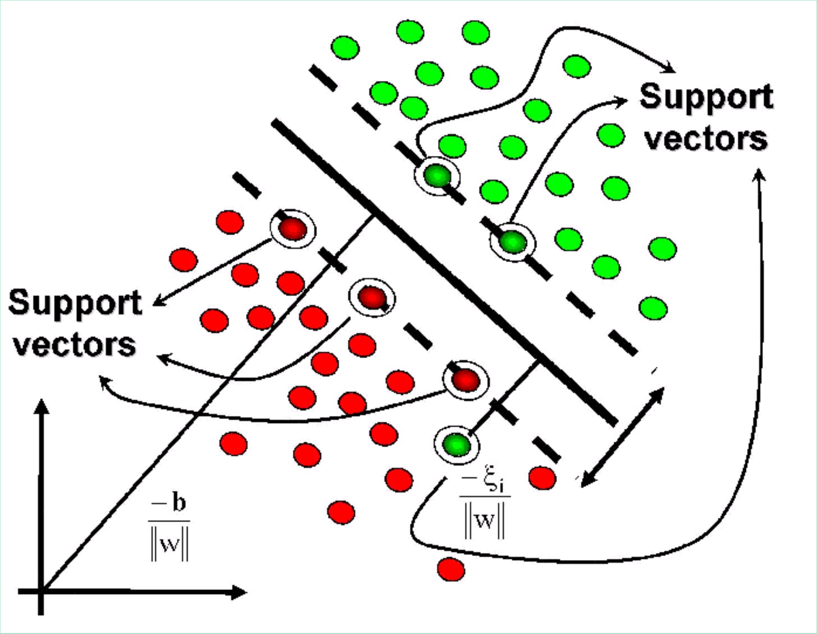


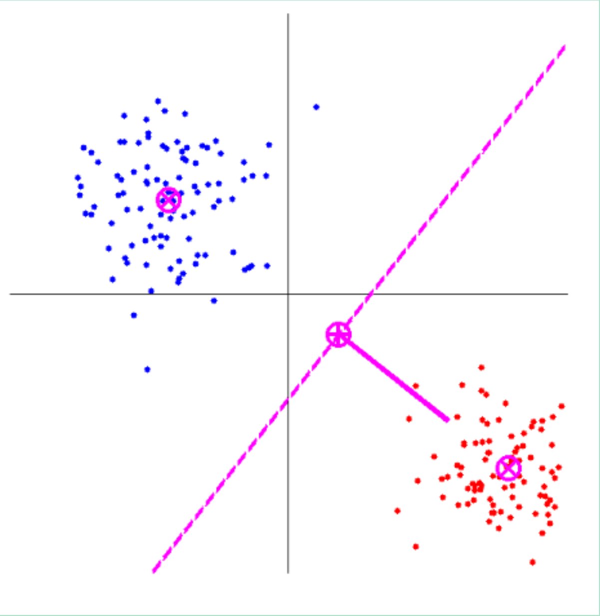
Significance between AUC or ROC Differences

- No test exists because handling the correlations between holdouts has not been solved
- A statistics Professor has recommended developing a new measure of classification success designed to allow significance tests
- This is an open research problem

Shape Classification by Separating Directions

- Vector between means of classes: Δμ
- Distance Weighted Discrimination (DWD)
  - vs. SVM, vs. Δμ
- Separator can be geodesic when on curved manifold
- Kernels to allow non-hyperplanar separators





Support Vector Machine

- Objective function to optimize over separating planes
  - Term for size of gap between classes
  - Term for misclassified cases
- Ultimately depends on the few cases nearest the gap plane ("support vectors")

Data:

$$\{(\mathbf{x}_i, y_i)\} \quad \mathbf{x_i} \in \mathbb{R}^p \quad y_i \in \{-1, 1\}$$

Optimization problem:

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x_i} - \mathbf{b}) \geq \mathbf{1}$$

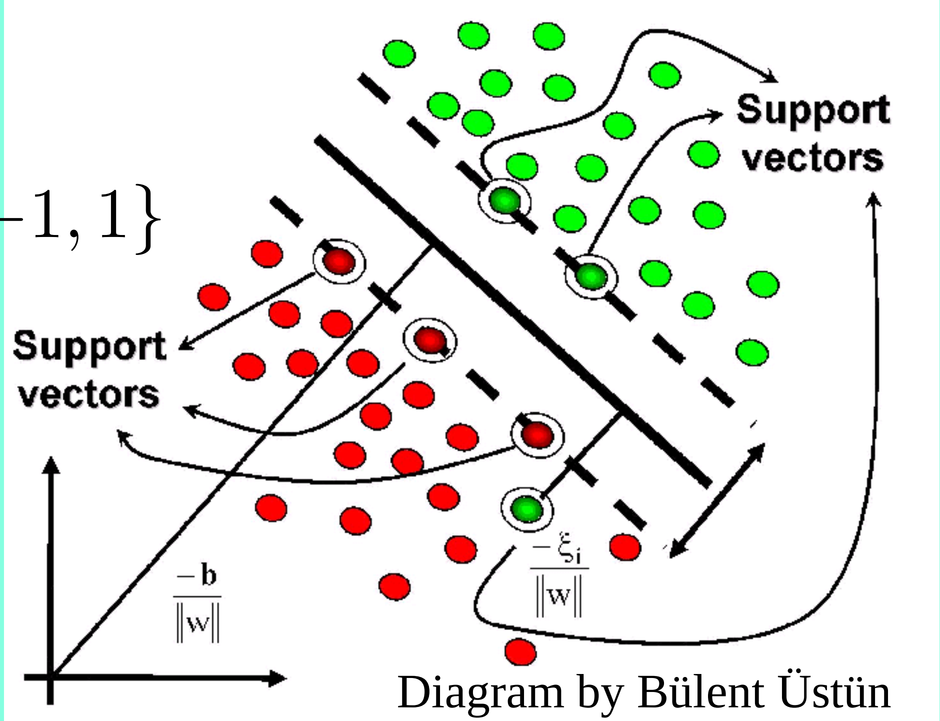
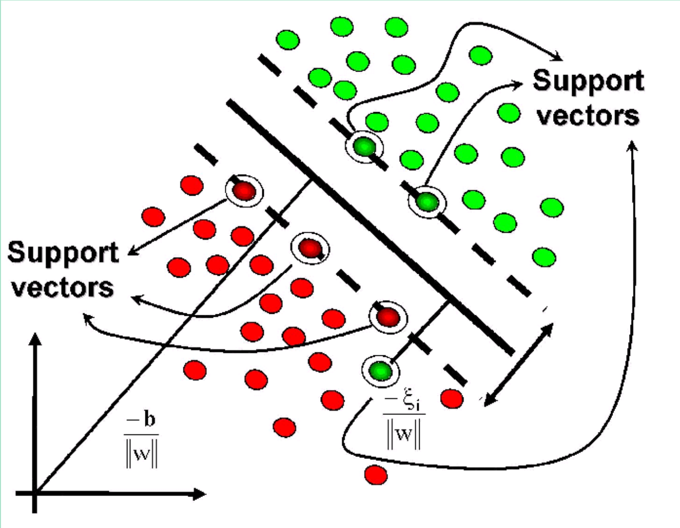


Diagram by Bülent Üstün

Shape Classification by Optimal Divider

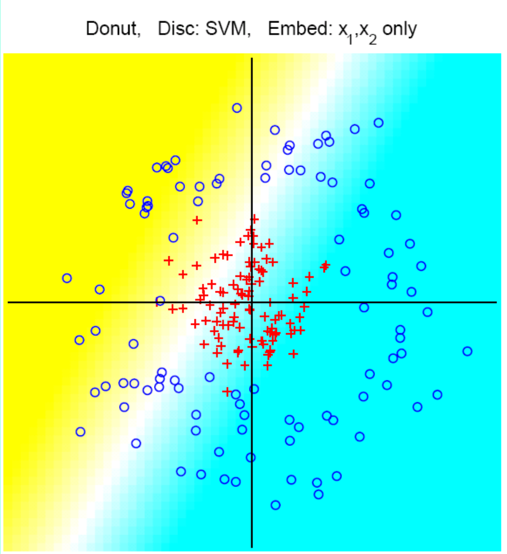
- Support Vector Machine (SVM)
  - Maximize gap between classes + misassignment penalty
    - Uses only training points at boundary of gap

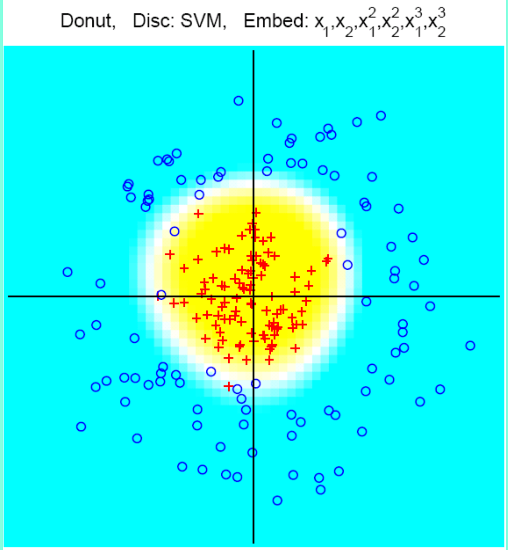


- Distance Weighted Discrimination (DWD)
  - Minimize $\Sigma_i\, 1/r_i$ + mis-assignment penalty; r_i = distance to divider (geodesic distance)
    - Uses all training points, weighted higher, the closer they are to divider
    - Thus more robust and more accurate
    - Or use power of r; if ∞, it is SVM

Separating Hyperplanes

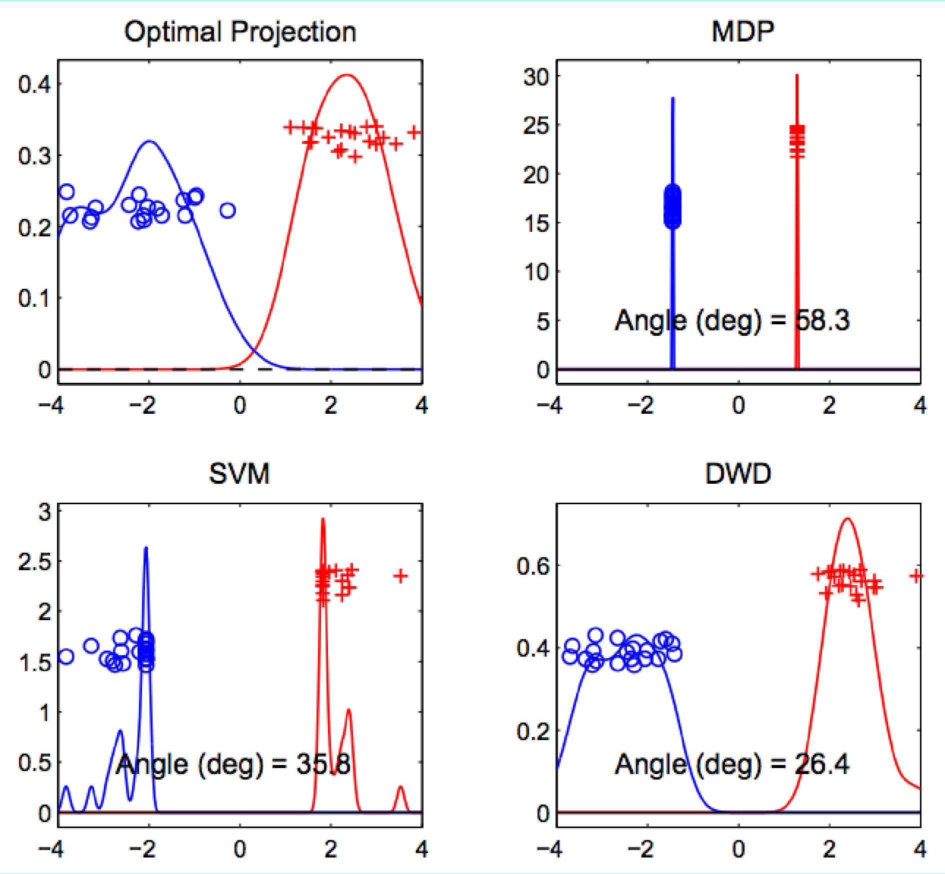
- Normal to (n-1)-dimensional hyperplane gives separating dimension
  - A geodesic
- Plane (geodesic) gives threshold along normal for decision
  - Additional features as functions of other features to get nonlinear decision boundaries (kernel methods)
- Support Vector Machine (SVM) or Distance Weighted Discrimination (DWD)



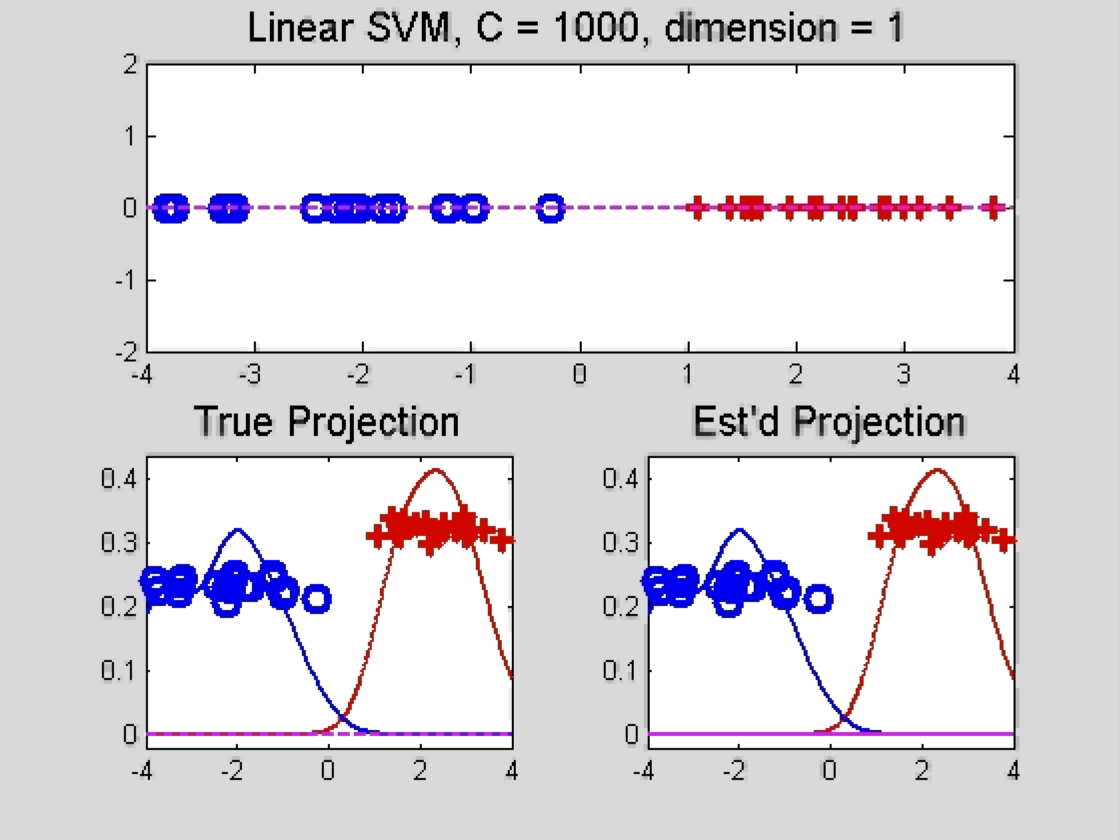


Distance Weighted Discrimination (DWD) [Marron]

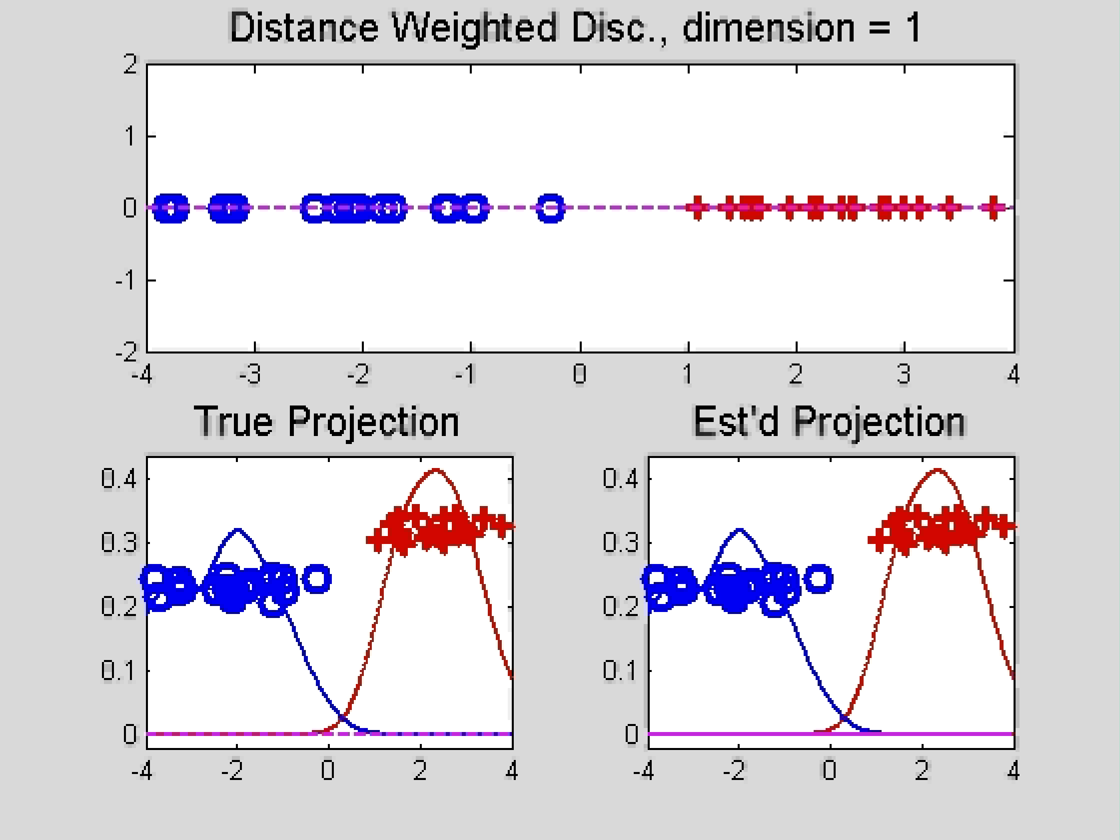
- Objective function to optimize
  - Term for distances of cases to hyperplane (sum of reciprocal distances)
    - Can be geodesics on curved manifold
    - Separating direction can be geodesic on curved manifold
  - Term for misclassified cases with distance weighting
- More robust than SVM when high dimension, low sample size



SVM vs. dimension on toy data



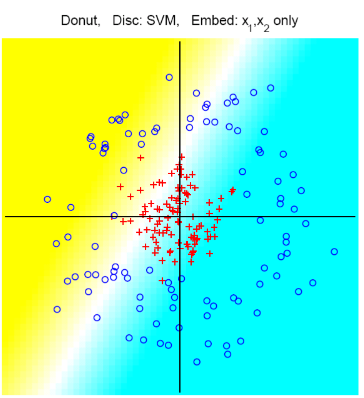
DWD vs. dimension on toy data

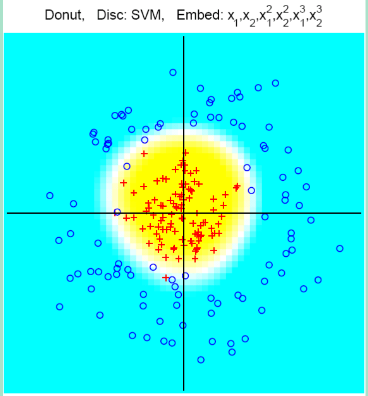


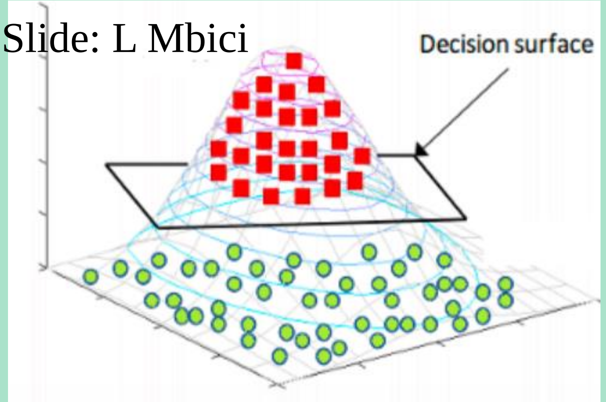
Separation using Kernels

- Works for DWD and for SVM
- Approach
  – Augment features by combinations of them
    - Embeds feature space into a higher dimension feature space
    - Changes what inner product means
    - Common features: radial basis functions
      – A la TPS; or isotropic Gaussians

  – Hyperplane divider in augmented feature space mapped on the original feature space can separate by non-hyperplanes
    - For example, a circle or an ellipse for a 2D feature space







Slide: L Mbici

Shape Statistics Effectiveness Measures

- Generalizability, Specificity
- Cross-validation
- AUC (and ROC) for classification
- DiProPerm
- Applications
  - Objectives and methods
  - Measures of success

Generalizability and Specificity of Model for Probability of Shape [Davies 2003]

- Generalizability
  - Measures closeness of new instances of the object to the probability distribution estimated from the training cases.
  - Calculated by computing a shape space, spanned by the eigenmodes, on all but one of the training cases and computing the distance between the last shape and its projection onto this shape space, and doing this for each left out shape.
- Specificity
  - Measures how well the estimated probability distribution represents only valid instances of the object.
  - Calculated as average distance between random samples in the computed shape space with their nearest members of the data.

Generalizability and Specificity of S-rep-implied Boundary PDM for Probability of Shape [Tu 2015]

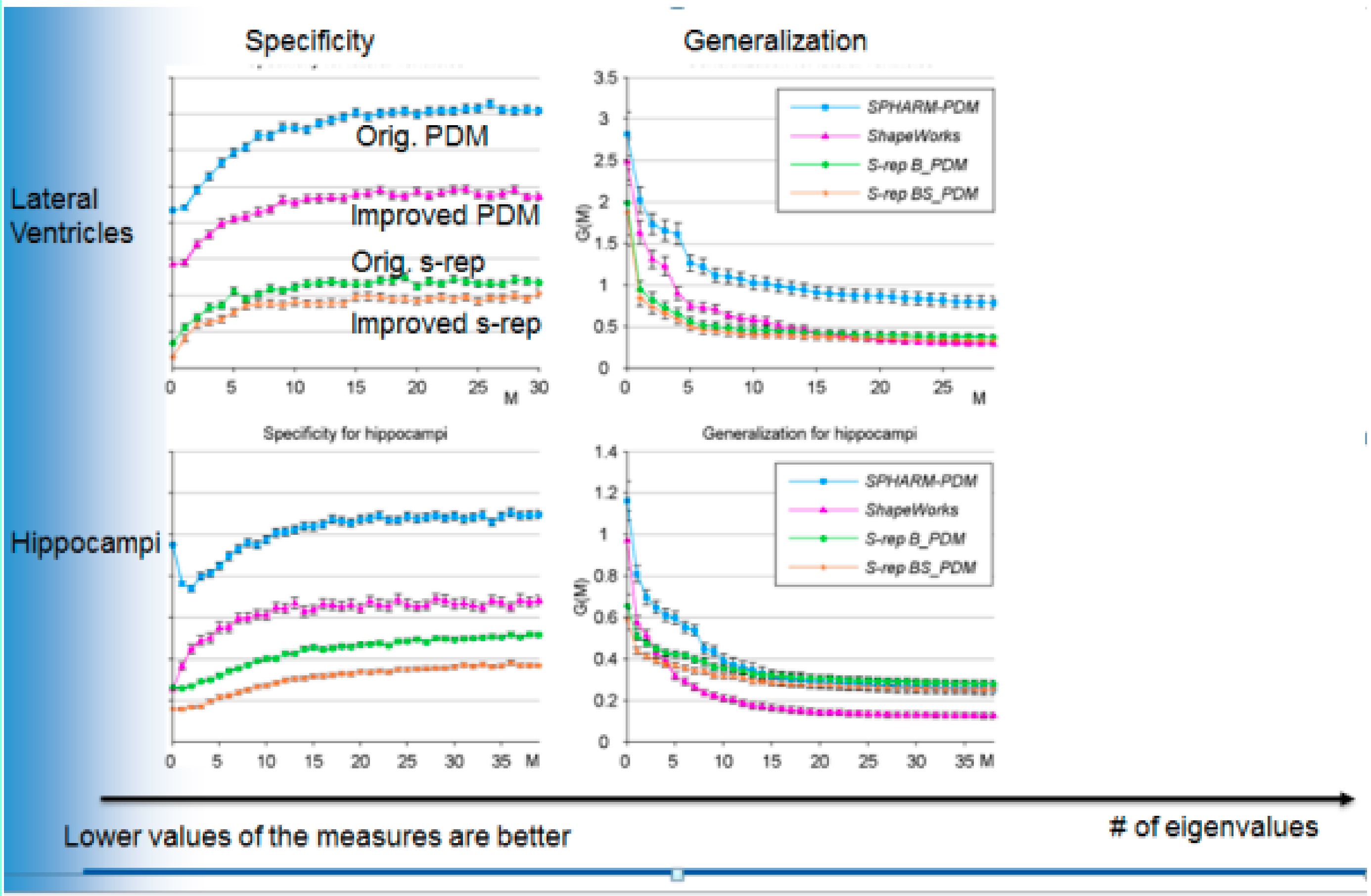
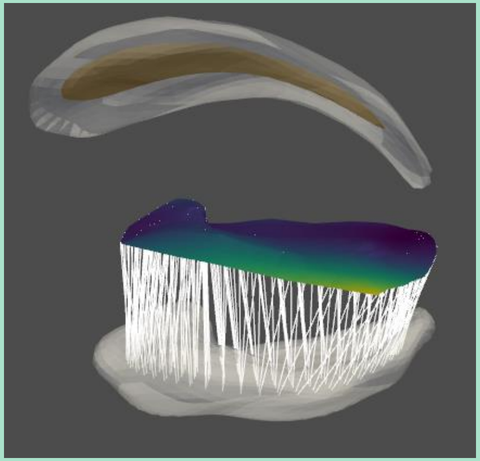
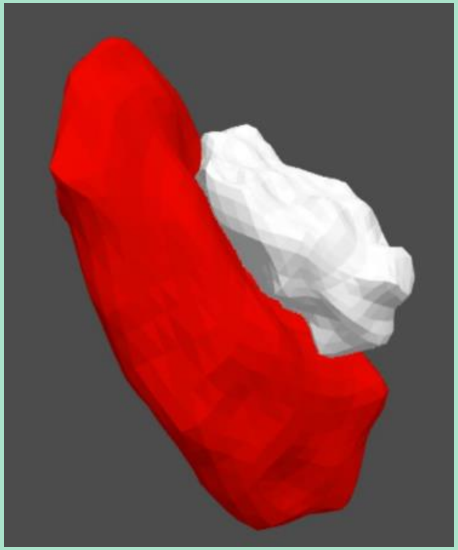


Figure 13. Generalization and specificity improvements due to s-rep-based and PDM-based correspondence optimization

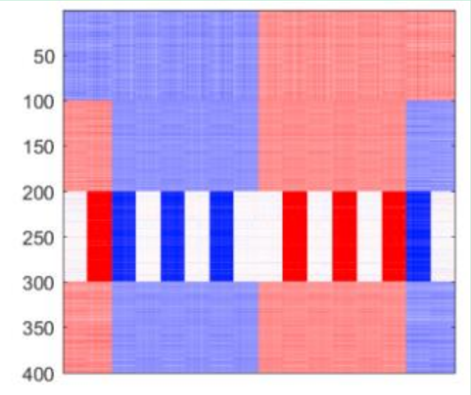
Multi-entity Shape Analysis

- AJIVE has become DIVAS
  - [Prothero, …, Marron, March 2023]
    - On my Google drive for course
  - Want to separately capture joint variations among entities, individual variations of each entity, noise variation
  - For objects, entities can be individual object geometric features and inter-object (e.g., links) geometric features
    - Note a disease may jointly affect many objects
    - Focusing classification on joint variation can be more robust due to removal of noise





AJIVE [Feng 2018]



- Assume K blocks (here objects, K=2); k=1,…,K
  - With d_k Euclidean features, n training cases; $d_k >> n$ in our situation
    - Different blocks may have different features
  - For each k restrict space to $m_k < n-1$ dim space: delete dimensions with 0 eigenvalues and those designated as noise
    - So data matrix is K blocks X_k, each $R^{m_k \times n}$
- Want decomposition of each X_k into features in Joint matrix J, in individual matrices I_k, and into error matrix E_k
  - E_k from low eigenvalued modes of PCA of X_k

AJIVE, 2

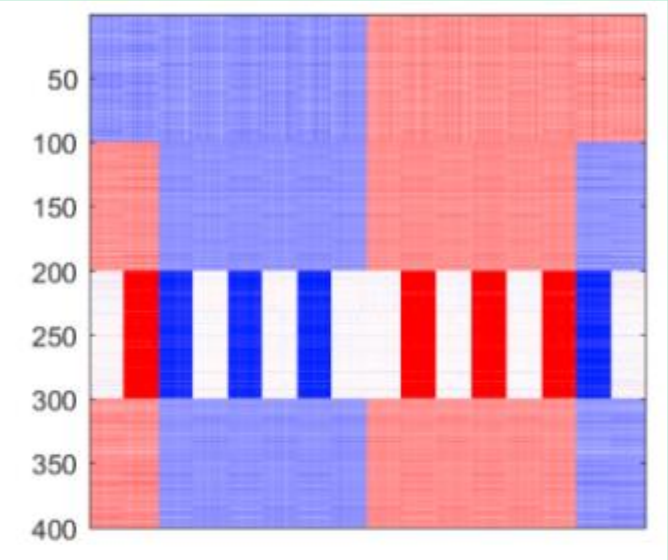


Fig. from Prothero et al.

- Want decomposition of $R^{n-1 \times n}$ matrices X_k into Joint matrix J, individual matrices I_k, and error matrix E_k
  - n is number of training cases
  - J made of joint features orthogonal to each other, and to space of each I_k
  - I_k features orthogonal to each other but not necessarily to those of I_k for a different k

AJIVE, 3

- Features forming J come from a subspace of principal features with small angles across the X_k spaces
- Linear algebra is in data item space rather than feature space
- The hard part is choosing the rank (dimension spanned by the feature tuples) of J

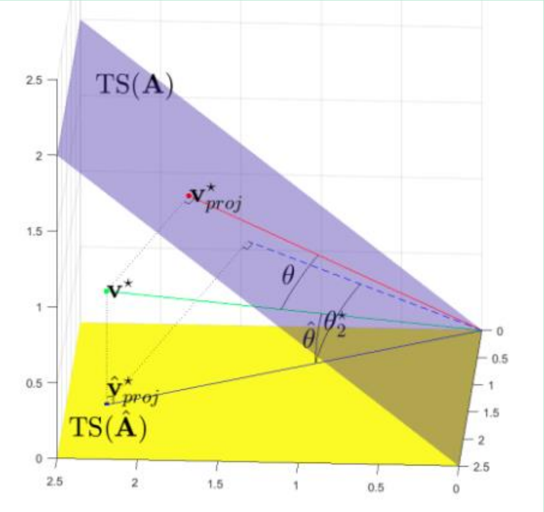


Fig. from Prothero et al.

AJIVE, 4: Algorithm [Z Liu]

- Let $U_k \Lambda_k V_k^T$ be result from SVD of X_k
  - And removal of rows of U_k and V_k^T according to eigenvalues Λ_k
  - Yields updated X_k having lowered noise
- Form basis vectors of J by Principal Angle Analysis (PAA) on the the V_k^T pair (tuple if K>2)
  - Small angles indicate joint correlation
  - If rows of V_1^T are $\mathbf{v}_i$ and rows of V_2^T are $\mathbf{q}_i$,

$$\theta_j = \min_{v_j, q_j} \; cos^{-1}\left(\frac{< v_j, q_j >}{||v_j|| \cdot ||q_j||}\right),$$

$$v_j \perp v_i, \; q_j \perp q_i, \; i \in [1, \; j-1]$$

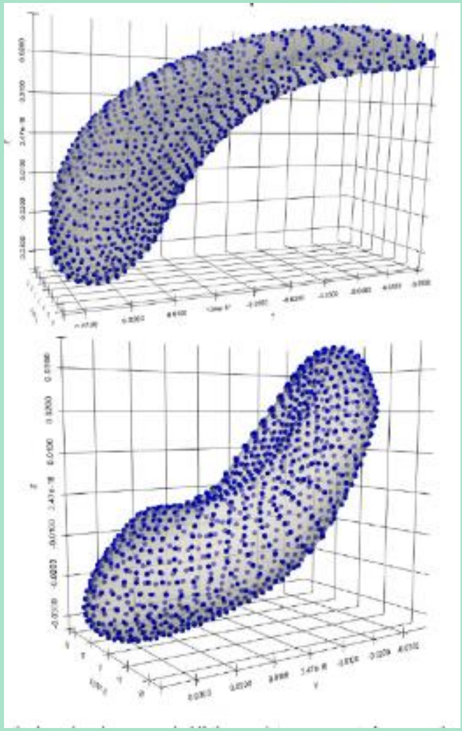
for very small θ_j

AJIVE to DIVAS

- AJIVE has become DIVAS
  - [Prothero, …, Marron, March 2023]
  - Make rank choices by hypothesis tests
  - Allows joint matrices of subsets of blocks
  - Uses Marchenko-Pastur analysis

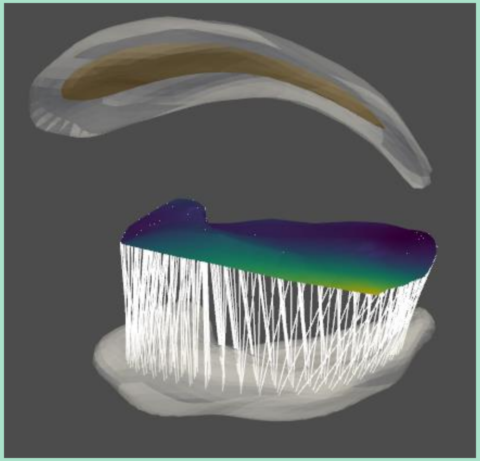
NeuJIVE [Z Liu, on MIDAG website]

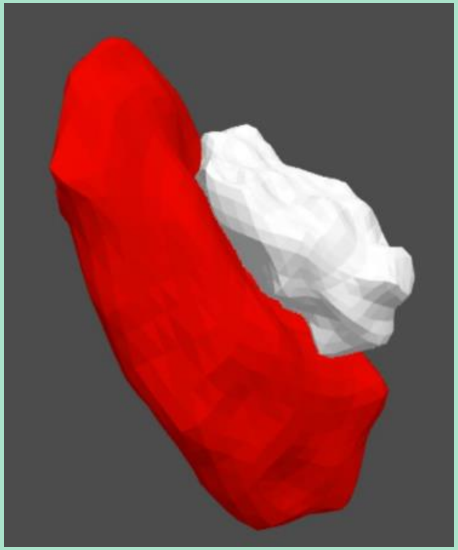
- AJIVE on features derived from PNS
- The principle is that joint variation can show disease more robustly by removing individual and noise variations
- Liu’s results on aligned PDM’s from spoke ends of block 1: hippocampus and block 2: caudate confirm that



Multi-object Classification [Z Liu, on MIDAG website]

- AJIVE on features derived from PNS
  - And using affine frames
- Principle: joint variation can show disease more robustly by removing individual and noise variations
- Liu’s results confirm that
  - For hippocampus and links to linking surface between hippocampus and caudate
    - But not caudate as it is not a good classifier by itself
  - And better classification than PDM’s of the two objects





Longitudinal Analysis of Shape

- How does shape change over time?
- Longitudinal data set:
  - Set of homologous objects (e.g., anatomical structures), each object being observed repeatedly at several time points; for now non-cyclic, e.g., aging
  - Shape varies across individuals
  - For each individual shape temporal paths differ
    - Temporal sampling may differ among individuals
    - Initial times and rates of changes of shape longitudinal changes may differ among individuals
- Two approaches
  - Use object features $\underline{\mathbf{z}}$ on curved manifold and study $\underline{\mathbf{z}}(t)$
    - See Fletcher lectures in this course and his papers
    - Schiratti, …, Durrleman, for mixed effects models on manifolds: "Learning spatiotemporal trajectories from manifold-valued long'l data"
  - Study deformations in time: $\phi(\underline{x},t)$. See upcoming slides

Longitudinal Analysis of Shape: Analysis Objectives

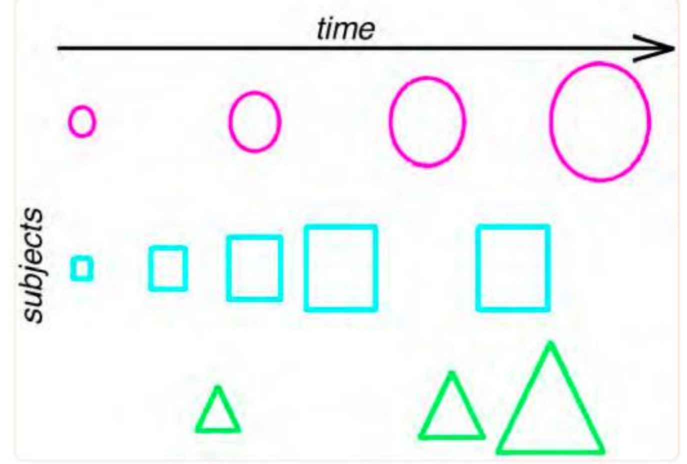
- Qualitative and quantitative assessment of change trajectories
- Detection of common growth patterns shared in a population
- Characterization of their appearances in different subjects.

Longitudinal Analysis of Shape via Space Deformations

- References
  - Durrleman, S., Pennec, X., Trouvé, A., Braga, J., Gerig, G., Ayache, N.: Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data, IJCV **2013**
    - I quote and just previously have quoted figures and other material from this paper

- Given a longitudinal shape data set,
  - Estimate a mean growth scenario M(t) representative of the population, and the variations of this scenario both in terms of shape changes and in terms of change in growth speed
  - Characterize the typical variations in shape and in growth speed within the studied population by
    - deriving intrinsic statistics in the space of spatiotemporal deformations
  - Can be used to detect systematic developmental delays across subjects

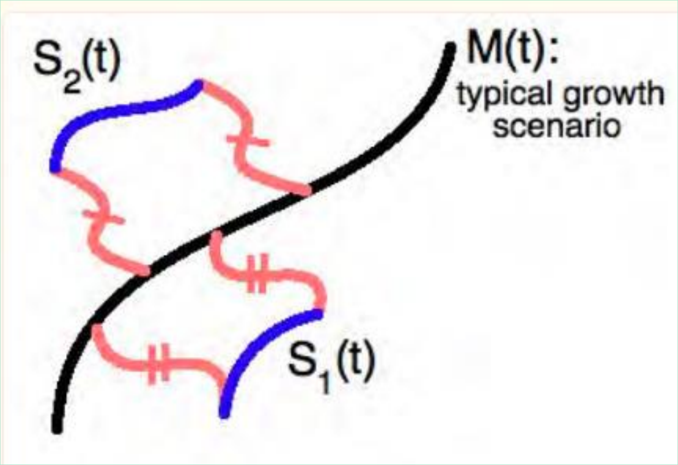
Longitudinal Analysis of Shape via Space Deformations, 2

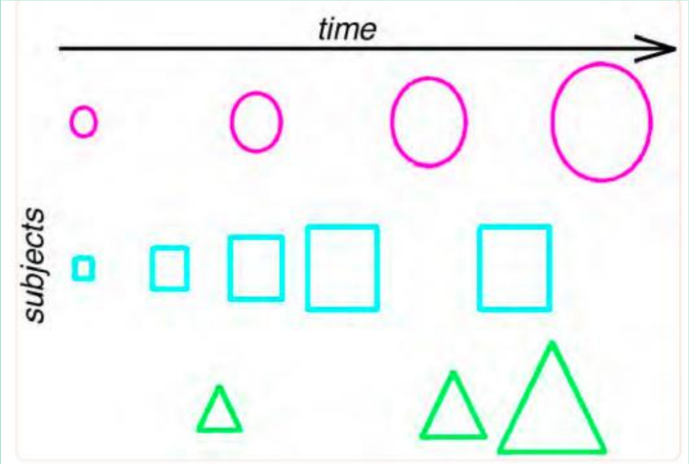


- Data
- Issues
  - Need to do subject by subject deformations across time:
  - Need to do subject by subject temporal deformations from atlas
    - Time passes on differently for different subjects
  - Need to do inter-subject deformations from atlas
    - Thus need an atlas, M(t), typically a kind of mean
  - Want to include covariates such as age (especially) and gender in the analysis
  - Would like to do machine learning on this
    - M Ren et al. (NYU), Local Spatiotemporal Representation Learning for Longitudinally-consistent Neuroimage Analysis, NeurIPS 2022

Longitudinal Analysis of Shape via Space Deformations, 3

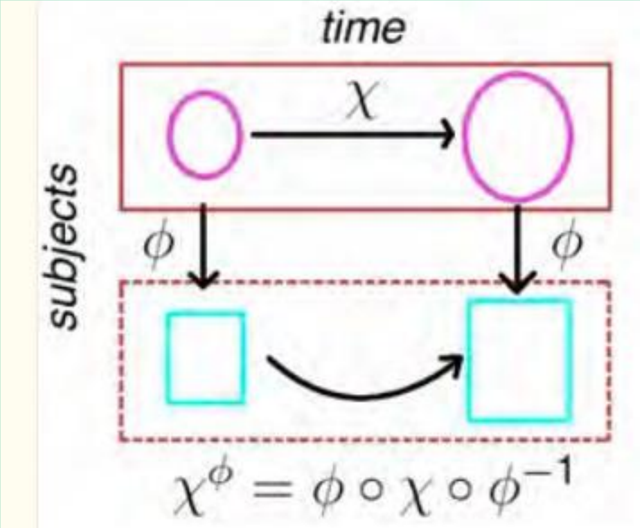
- Issues
  - Need to do subject by subject deformations across time: $S_i(t)$
  - Need to do subject by subject temporal deformations from atlas
    - Time passes on differently for different subjects
  - Need to do inter-subject deformations from atlas
    - Thus need an atlas, M(t), typically a kind of mean

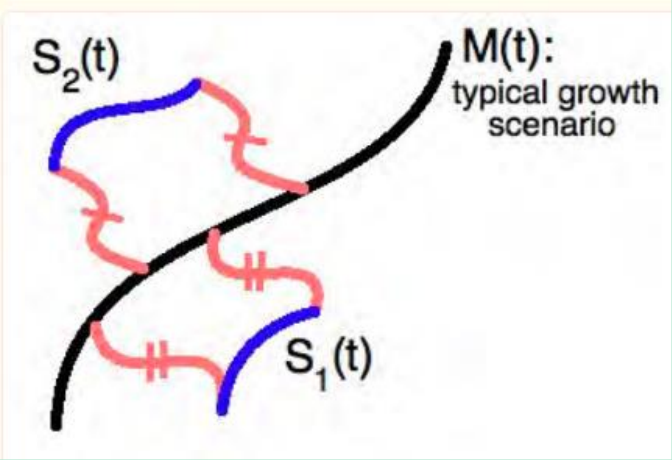




Subject-specific Longitudinal Analysis of Shape via Space Deformations

- Let $\chi(t)$ be deformation of shape across time of atlas
  - Let $\chi_j(t)$ be deformation of shape across time of subject j
  - Let $\phi_{jt}(\underline{x})$ be deformation of shape at time t from the atlas of subject j





- Via LDDMM approach minimize

$$E(\chi) = \sum_{t_i} \mathrm{d}(\chi_{t_i}(M_0), S_i)^2 + \gamma^\chi \mathrm{Reg}(\chi)$$

Subject-specific Longitudinal Analysis of Shape via Space Deformations

- Let $\chi(t)$ be deformation of shape across time of atlas
  - Let $\chi_j(t)$ be deformation of shape across time of subject j
- But time runs differently for each subject
  - Let $\psi(t)$ deform time: $\Phi(x, y, z, t) = (\phi(x, y, z, \psi(t)), \psi(t))$

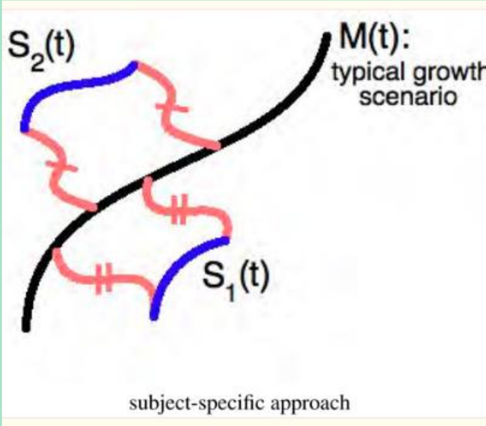
  where the geometrical part has the form

$$\phi(x, y, z, t) = \chi_t \circ \phi_0 \circ \chi_t^{-1}(x, y, z)$$

- Via LDDMM approach minimize

$$E(\phi, \psi) = \sum_{t_j} \mathrm{d}(\phi(S(\psi(t_j))), U_{t_j})^2 + \gamma^{\phi}\mathrm{Reg}(\phi) + \gamma^{\psi}\mathrm{Reg}(\psi)$$

  - where U_{tj} is the shape of object j at time t



Atlas Formation in Longitudinal Shape Analysis via Space Deformations,

- Find $\underline{M}_0(t)$ that is miminizer integrated over time series for subjects j of energy between $\underline{S}_j(t)$ and $\underline{M}_0(t)$
  - A sort of Fréchet mean
  - The energy should be

$$E\left((\psi^s)_{s=1,\ldots,N_{\text{subj}}}, (\phi^s)_{s=1,\ldots,N_{\text{subj}}}, \chi, M_0 \right) =$$

$$\sum_{s=1}^{N_{\text{subj}}} \left\{ \sum_{t_j^s} \mathrm{d}(\phi^s(\chi_{\psi^s(t_j^s)} M_0), S^s(t_j^s))^2 + \gamma^\phi \mathrm{Reg}(\phi^s) + \gamma^\psi \mathrm{Reg}(\psi^s) \right.$$

$$\left. + \gamma^\chi \mathrm{Reg}(\chi) \right\}$$

Longitudinal Shape Analysis Including Age Covariate [S Hong]

- S Hong, J Fishbaugh, J Wolf, M Styner, G Gerig, the IBIS Network. Hierarchical Multi-Geodesic Model for Longitudinal Analysis of Temporal Trajectories of Anatomic Shape and Covariates. Springer Nature, 2019
  –

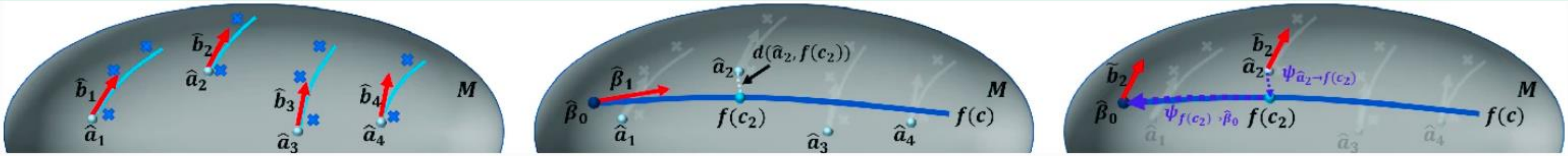


Illustration of the proposed method. (a) Subject-wise geodesic trajectory estimation. (b) The example of an intercept model $f(c)$ with a single covariate c. (c) Stop-over parallel transport ϕ from $\hat{a}_2$ to $\hat{\beta}_0$.

Longitudinal Shape Analysis Including Age Covariate [S Hong]

- Generalized linear model but on manifold
  - Regression: see Fletcher discussion of this
  - Ordinary generalized linear model, but on curved manifold
    - Subject-wise trajectory: $(\hat{a}_i, \hat{b}_i) = \underset{a_i, b_i}{\operatorname{argmin}} \sum_{j=1}^{N_{obs}^i} d^2(y_{ij}, Exp(a_i, b_i t_{ij})),$ $Y_i = Exp(\hat{a}_i, \hat{b}_i t)$
  - Handling covariate, e.g., age
    - It is associated with each shape data item
    - Yields Intercept and slope each linear over sample values

$$Y = Exp(Exp(f(\boldsymbol{\eta}), g(\boldsymbol{\theta})t), \epsilon).$$

    - Needs to be carried to subject average geodesic: done by parallel transport to and along trajectory (see next slide)

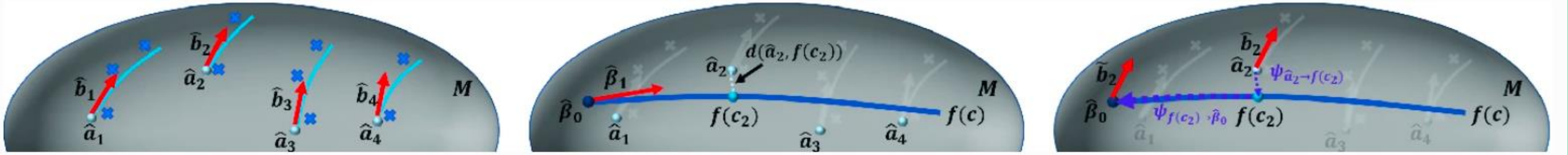


Illustration of the proposed method. (a) Subject-wise geodesic trajectory estimation. (b) The example of an intercept model $f(c)$ with a single covariate c. (c) Stop-over parallel transport ϕ from $\hat{a}_2$ to $\hat{\beta}_0$.

Longitudinal Shape Analysis Including Age Covariate [S Hong], 2

- Generalized linear model but on manifold
- Handling covariate, e.g., age
  - It is associated with each shape data item
  - "Linearly" implies geodesic intercept η and slope θ over sample values

$$Y = Exp(Exp(f(\boldsymbol{\eta}), g(\boldsymbol{\theta})t), \epsilon).$$

$$\beta_k \in T_{\beta_0}M \text{ [5]},$$

$$f(\boldsymbol{\eta}) = Exp(\beta_0, \beta_1\eta_1 + \ldots + \beta_{N_\eta}\eta_{N_\eta}).$$

$$g_0(\boldsymbol{\theta}) = \gamma_0 + \gamma_1\theta_1 + \ldots + \gamma_{N_\theta}\theta_{N_\theta},$$

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{N_\eta}) = \underset{\beta_0,\ldots,\beta_{N_\eta}}{\operatorname{argmin}} \sum_{i=1}^{N_s} d^2(f(\boldsymbol{\eta}^i), \hat{a}_i).$$

  - Needs to be carried to subject average geodesic: done by parallel transport of tangent vectors to and along trajectory
    - Angle and scale of vector kept constant along geodesic

$$\tilde{b}_i = \phi(\hat{b}_i) = \psi_{\hat{f}(\boldsymbol{\eta}^i)\to\hat{\beta}_0}(\psi_{\hat{a}_i\to\hat{f}(\boldsymbol{\eta}^i)}(\hat{b}_i)),$$

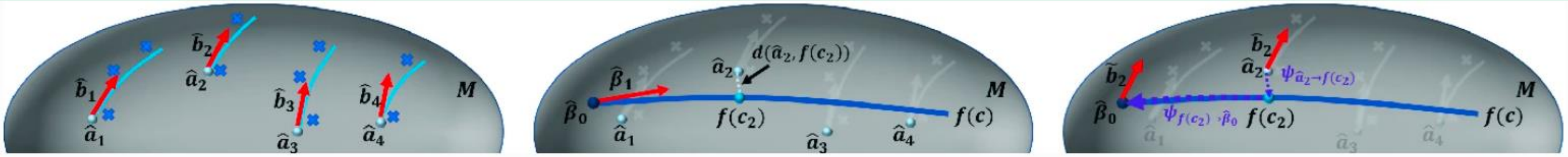


Illustration of the proposed method. (a) Subject-wise geodesic trajectory estimation. (b) The example of an intercept model $f(c)$ with a single covariate c. (c) Stop-over parallel transport ϕ from $\hat{a}_2$ to $\hat{\beta}_0$.