Bayesian ambient space inference for object data

Ian Dryden

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

CIRM, Luminy, September 1st, 2017.
Joint work with: Huiling Le, Kwang-Rae Kim, Wen Cheng, Xianzheng Huang, David Hitchcock.

Outline

Outline

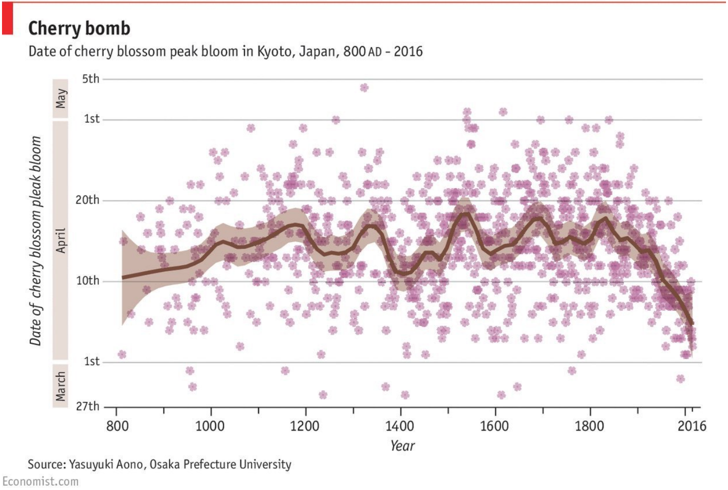
A New Era

"What steam was to the 18th century, electricity to the 19th, and hydrocarbons to the 20th, data will be to the 21st century. That's why I call data a new natural resource."

Ginni Rometty, Chairman, President and CEO of IBM

New?

Me: ”What’s new about data?”



Traditional types of data

What types of data are there?

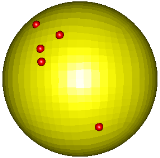
- Counts, e.g. $\{0, 1, 2, \ldots\}$

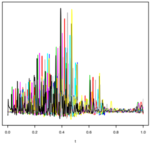
- Measurements, e.g. 27.52

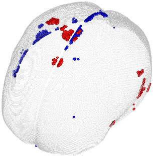
- Many measurements (Vectors), e.g. $(3.2, 1.2, 54.3, 2.1)$

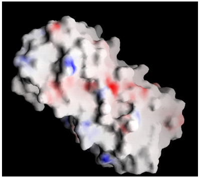
Object Data

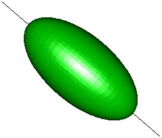
- Circular and spherical data
- Functions
- Dynamical systems
- Shapes and manifold data
- Images
- Trees

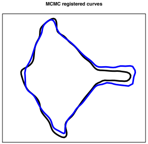


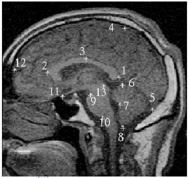












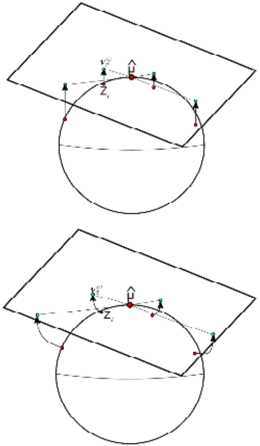
Left: FLAT manifold Right NON-FLAT manifold

- Shortest distance between two points may not be a straight line.
- Need to adapt conventional FLAT space data analysis for analysis on manifolds

Shape analysis

KEY ASPECTS:

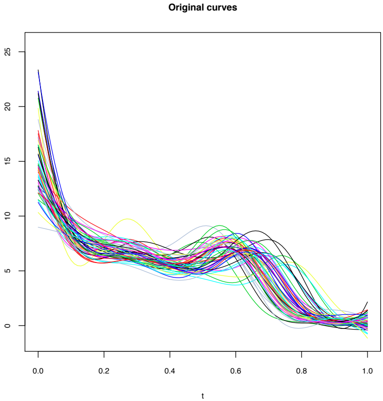
- SHAPE: remove REGISTRATION information (e.g. Rotation, Translation, Scale - D.G. Kendall)
- Shape data usually lie on a non-flat manifold
- Approximation using a flat tangent space
- Carry out PCA and further statistical inference



Outline

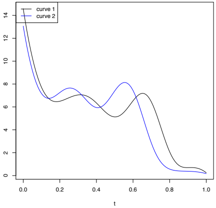
Functional Data Analysis

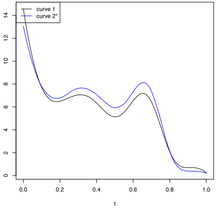
Example 1: Berkeley Girls growth-rate data (54 curves - age 1-18)



Functional Data Analysis

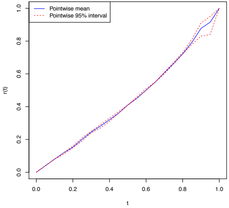
Two curves (before and after registration)





Warp

Time-warp (in Diff[0, 1]) - posterior mean and 95% credibility interval



Ambient versus Quotient Spaces

- The **Ambient Space** M , contains standardized functions.
- Usually a simple metric space, e.g. $\mathbb{R}^p$, $\mathbb{L}^2$, S^{p-1}.

Ambient versus Quotient Spaces

- The **Ambient Space** M , contains standardized functions.
- Usually a simple metric space, e.g. $\mathbb{R}^p$, $\mathbb{L}^2$, S^{p-1}.
- The **Quotient Space** $Q = M/G$, contains the equivalence classes - where the group G of transformations has been removed.
- Q is usually non-Euclidean: the geometry can be complicated.

Ambient versus Quotient Spaces

- The **Ambient Space** M , contains standardized functions.
- Usually a simple metric space, e.g. $\mathbb{R}^p$, $\mathbb{L}^2$, S^{p-1}.
- The **Quotient Space** $Q = M/G$, contains the equivalence classes - where the group G of transformations has been removed.
- Q is usually non-Euclidean: the geometry can be complicated.
- In which space should we work M or Q?

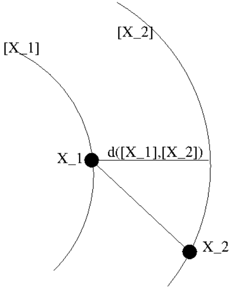
Comparing two objects

- Data in ambient space: X_1 and X_2
- Distance in quotient space:

$$d([X_1],[X_2]) = \inf_{\gamma \in G} d(X_1, X_2 \circ \gamma)$$

where γ is an isometric transformation, e.g. a time warp: $\mathrm{Diff}[0,1]$

- Invariance property $d([X_1 \circ \gamma_0],[X_2 \circ \gamma_0]) = d([X_1],[X_2])$.

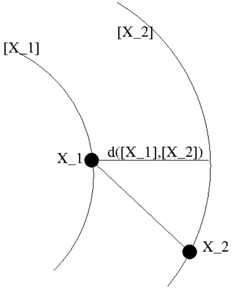


Comparing two objects

- Data in ambient space: X_1 and X_2
- Distance in quotient space:

$$d([X_1], [X_2]) = \inf_{\gamma \in G} d(X_1, X_2 \circ \gamma)$$

  where γ is an isometric transformation, e.g. a time warp: $\mathrm{Diff}[0, 1]$
- Invariance property $d([X_1 \circ \gamma_0], [X_2 \circ \gamma_0]) = d([X_1], [X_2])$.
- Geometry/models are usually simpler in the ambient space



Square Root Velocity Function (SRVF)

- Let f be a real valued differentiable curve function $f(t) : [0,1] \to \mathbb{R}^m$.
- The SRVF is defined as $q : [0,1] \to \mathbb{R}^m$, where

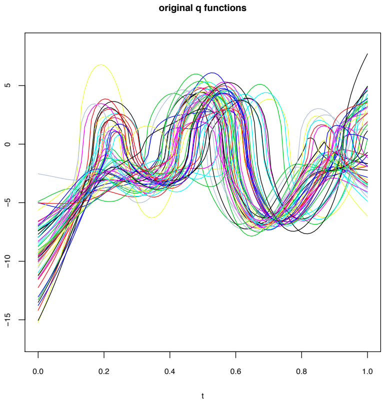
$$q(t) = \frac{\dot{f}(t)}{\sqrt{||\dot{f}(t)||}}$$

and $||f||$ denotes the standard $\mathbb{L}^2-$norm (Srivastava et al., 2011; cf. Younes, 1998).
- Why use the SRVF? The Fisher-Rao (Elastic) metric is reduced to a standard $\mathbb{L}^2$ metric under SRVF representation.

$$d_{FR}(f_1, f_2) = \|q_1 - q_2\|.$$

Ambient space curves

Berkeley Girls q-functions of growth-rates - need to align them using a time warps



Likelihood Model for q function

- A Gaussian process for $q_1(t) - q_2^*(t)$, where

  $q_2^*(t) = \sqrt{\dot{\gamma}(t)}q_2(\gamma(t))$, given a fixed $\gamma(t)$.

- Let $q_1([t])$ and $q_2^*([t])$ denote the finite M points of $q_1(t)$ and $q_2^*(t)$

- The joint distribution is multivariate normal,

$$q_1([t]) - q_2^*([t]) \sim N(0_M, \Sigma_{M\times M})$$

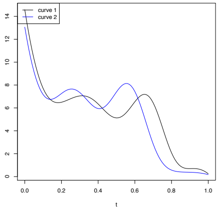
given a fixed $\gamma(t)$.

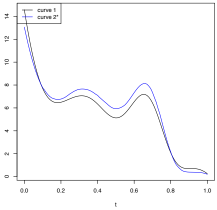
Prior Model for the Warp

- Discretize the time warp
- Let $\gamma([t])$ denote $\{\gamma([t_i]), i = 0, 1, 2, \ldots, M\}$, the collection of discretized points on the warping function.
- Define $p_i = \gamma([t_i]) - \gamma([t_{i-1}])$.
- Note $\Sigma_{i=1}^{M} p_i = 1$ and $0 < p_i < 1$.
- Denote $P_M = (p_1, p_2, \ldots, p_M)$ and treat P_M as a random vector, a Dirichlet prior is assigned to $\{P_M|\gamma([t])\}$, i.e. $\pi(P_M) \sim Dirichlet(a)$.
- Large a encourages unit slope $\dot{\gamma}(t) = 1$.

Functional Data Analysis

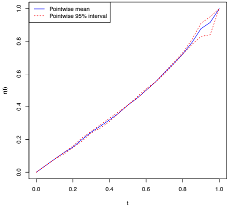
Two curves (before and after registration)





Warp

Time-warp (in $\mathrm{Diff}[0,1]$) posterior mean and 95% credibility interval ($a = 1$ here)



Multiple curves

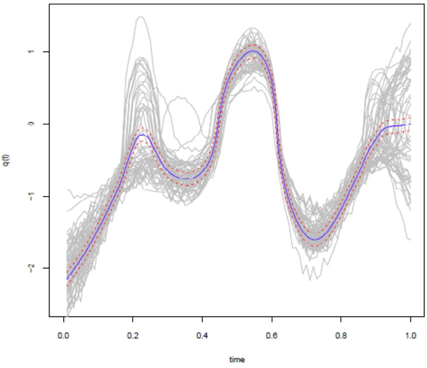
- q-functions: $q_i \sim N(\mu, \kappa^{-1} I), i = 1, \ldots, n$.
- Ambient space mean $\mu = E[X]$ (Gaussian prior).
- warps $\gamma_i(t), i = 1, \ldots, n$ independent Dirichlet prior
- $\kappa \sim \Gamma(\alpha, \beta)$ independent prior
- Simulate from the posterior distribution

$$(\mu, \kappa, \gamma_1, \ldots, \gamma_n) | q_1, \ldots, q_n.$$

using Markov chain Monte Carlo simulation.

Bayesian analysis - posterior mean

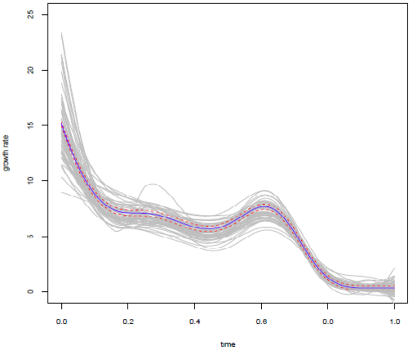
Berkeley Girls growth-rates - q-functions ($a = 50$)



Bayesian analysis

Berkeley Girls growth-rates - icons ($a = 50$)

$$f(t) = \int_0^t q(s)|q(s)|ds$$



Ambient space asymptotic normality and consistency

Subject to the conditions of the Bernstein-von Mises theorem (van der Vaart (1998, p141), we have

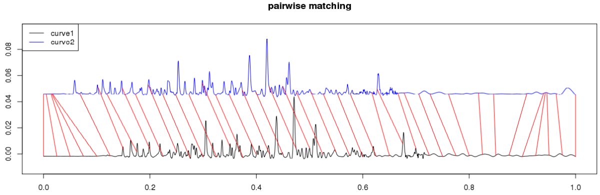
$$\sqrt{n}(\hat{\mu}([t]) - \mu([t])) \to N(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{\mu}^{-1} \dot{\ell}_{\mu([t])}(X_i) \ , \ I_{\mu([t])}^{-1})$$

in total variation norm as $n \to \infty$, where $\dot{\ell}_{\mu([t])}(X_i)$ is the derivative of the log-likelihood corresponding to observation i. We can state that $\hat{\mu} \to \mu$ in probability as $n \to \infty$, and hence the ambient space mean is consistent. (cf. Allassonnière et al., 2007, 2010)

Quotient space registration

- Let f_1, f_2 be two functions with SRVFs q_1, q_2.
- Warp f_2 to f_1 to minimize the Fisher-Rao distance using the optimal warp

$$\hat{\gamma} = \inf_{\gamma \in \Gamma} ||q_1 - \sqrt{\dot{\gamma}}(q_2 \circ \gamma)||^2.$$



- The solution can be obtained by Dynamic Programming (Srivastava et al., 2011).

Which is better?

- Ambient space model (e.g. Gaussian with mean μ, variance $\sigma^2 I$) easier to understand and interpret. (+)

Which is better?

- Ambient space model (e.g. Gaussian with mean μ, variance $\sigma^2 I$) easier to understand and interpret. (+)
- Ambient space marginal likelihood very complicated in general. (-)

Which is better?

- Ambient space model (e.g. Gaussian with mean μ, variance $\sigma^2 I$) easier to understand and interpret. (+)
- Ambient space marginal likelihood very complicated in general. (-)
- Ambient space posterior mode/MLE consistent for mean μ in ambient space. (++) (Allassonnière et al., 2007).

Which is better?

- Ambient space model (e.g. Gaussian with mean μ, variance $\sigma^2 I$) easier to understand and interpret. (+)
- Ambient space marginal likelihood very complicated in general. (-)
- Ambient space posterior mode/MLE consistent for mean μ in ambient space. (++) (Allassonnière et al., 2007).
- Quotient space (least squares) estimator biased in general for μ (-) (Miolane et al., 2017)

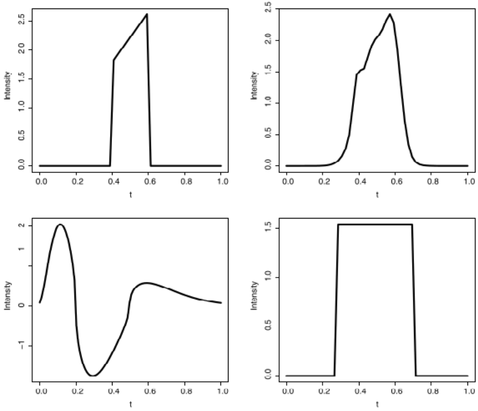
Which is better?

- Ambient space model (e.g. Gaussian with mean μ, variance $\sigma^2 I$) easier to understand and interpret. (+)
- Ambient space marginal likelihood very complicated in general. (-)
- Ambient space posterior mode/MLE consistent for mean μ in ambient space. (++) (Allassonnière et al., 2007).
- Quotient space (least squares) estimator biased in general for μ (-) (Miolane et al., 2017)
- Quotient space estimator consistent for population Fréchet mean (+) (Bhattacharya and Patrangenaru, 2003)

Which is better?

- Ambient space model (e.g. Gaussian with mean μ, variance $\sigma^2 I$) easier to understand and interpret. (+)
- Ambient space marginal likelihood very complicated in general. (-)
- Ambient space posterior mode/MLE consistent for mean μ in ambient space. (++) (Allassonnière et al., 2007).
- Quotient space (least squares) estimator biased in general for μ (-) (Miolane et al., 2017)
- Quotient space estimator consistent for population Fréchet mean (+) (Bhattacharya and Patrangenaru, 2003)
- Quotient space inference - faster (dynamic programming for warping) and relatively easy (+).

Simulation study - sample size n, noise σ



Add iid $N(0, \sigma^2)$ noise and apply a Dirichlet(1) warp to each individual in the sample of size n.

Results - log mean square FR distance versus $\log n$

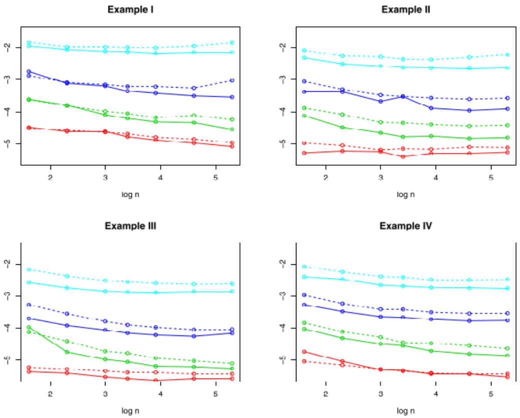
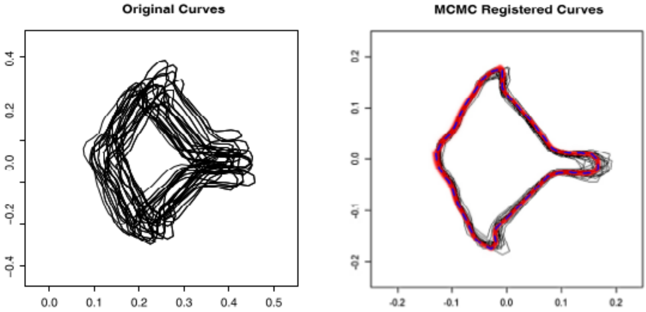


Figure 3: The logarithm of the mean square Fisher–Rao distance to the true mean μ_A versus logarithm of sample size n. The full line is the ambient space estimator and the dotted line is the quotient space estimator. The colors are red ($\sigma = 0.1$), green ($\sigma = 0.3$), blue ($\sigma = 0.5$) and cyan ($\sigma = 1$).

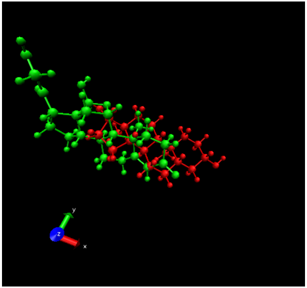
Higher dimensions



Outline

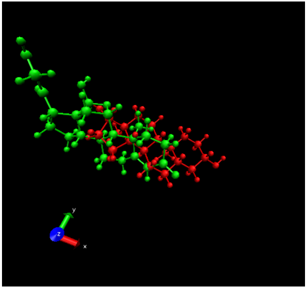
Example 2: Bayesian molecule matching

- Common task in cheminformatics and bioinformatics - the alignment and comparison of two or more molecules.



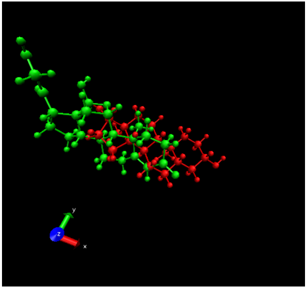
Example 2: Bayesian molecule matching

- Common task in cheminformatics and bioinformatics - the alignment and comparison of two or more molecules.
- Geometric similarity ('steric' properties) is a key property.



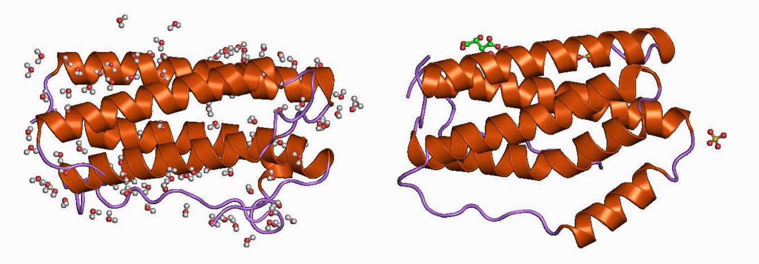
Example 2: Bayesian molecule matching

- Common task in cheminformatics and bioinformatics - the alignment and comparison of two or more molecules.
- Geometric similarity ('steric' properties) is a key property.
- Aligning molecules is vital but extremely difficult



Molecule matching

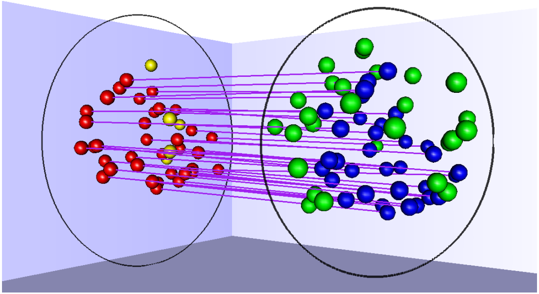
- When comparing molecules we are interested in similar parts of molecules rather than the whole match. Matching is sensitive to a prior parameter governing extent of overlap.



proteins: 1bgc [Granulocyte colony-stimulating factor], 1il6 [Interleukin-6] [wikipedia]

Match matrix and registration

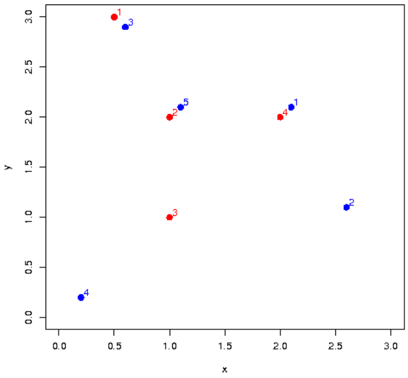
We need to estimate a match matrix M with 1 in position (j,k) if molecule 1 atom j matches to molecule 2 atom k, otherwise zeros; a rotation matrix Γ; and a translation vector γ.



Model: Given M, molecule 2 is a Gaussian perturbation of the matching atoms in molecule 1, independent with common variance $\sigma^2 = 1/\kappa$.

Example Match Matrix

Molecule 1: $n_1 = 4$ points (red) and Molecule 2: $n_2 = 5$ points (blue).



Matching points:
$1 \to 3$; $2 \to 5$;
$3 \to$ no match; $4 \to 1$
Match matrix

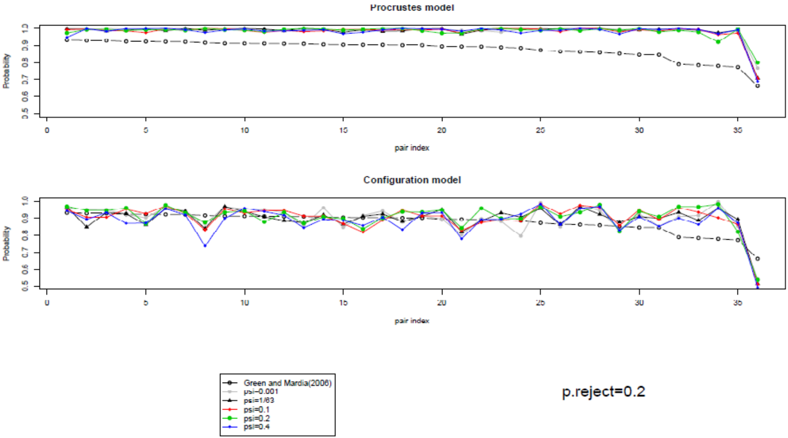
$$M = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Here $p = 3$ matching points.

Which approach is better - ambient versus quotient?

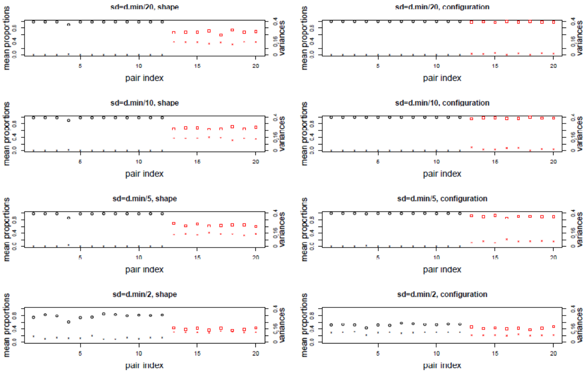
- Bayesian inference using Markov chain Monte Carlo (MCMC) simulation.
- Kenobi and ILD (2012) compare an ambient space model of Green and Mardia (2006) with a quotient space model (ILD et al., 2007; Schmidler, 2007).
- In a range of settings: performance similar but not the same. Invesigated protein matching and simulation studies.
- Ambient space MCMC algorithms less ‘sticky’
- Quotient space MCMC algorithms gave higher posterior probabilities of true matches.
- But Ambient space MCMC algorithms gave lower posterior probabilities of false matches.

Quotient above, ambient below



p.reject=0.2

Simulations: quotient left, ambient right



Estimated probability of correct match (black) and unmatch (red). Mean and variance from 100 simulations, of length 100,000 after burn-in.

Reason for general similarity of approaches?

- Marginal posterior density (Ambient Space inference).

$$\pi_A(\Theta|X) = \int_\gamma \pi(\Theta, \gamma|X) d\gamma. \tag{1}$$

- Quotient space posterior density

$$\pi_Q(\Theta|X) \propto \sup_\gamma \pi(\Theta, \gamma|X). \tag{2}$$

- We can consider (2) to be an approximation to the marginal density (1) where the integral is approximated using Laplace's method.

Laplace's method

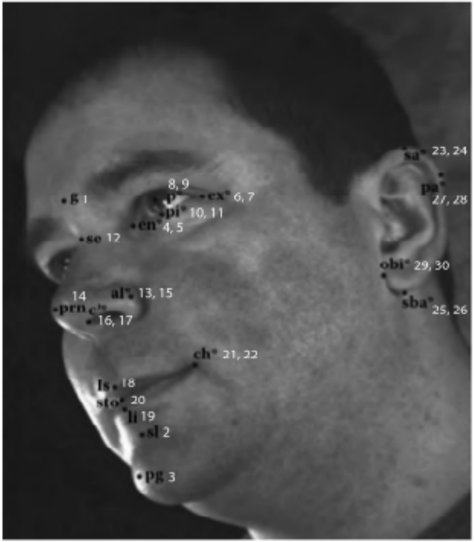
- Laplace's method:

$$\int b(t)\exp\{-Mr(t)\}dt \approx b(\hat{t})\left(\frac{2\pi}{M}\right)^{p/2}|\Sigma_{\hat{t}}|^{1/2}\exp\{-Mr(\hat{t})\}.$$

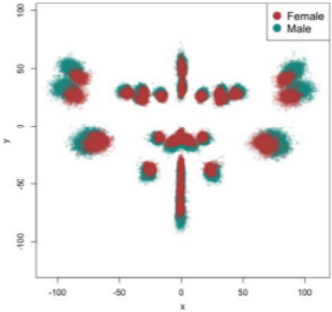
  where the gradient of $r(t)$ is zero at $\hat{t}$, and $\Sigma_{\hat{t}}$ is the inverse of the Hessian matrix at $\hat{t}$ (postive definite).
- The approximation is exact when $(\gamma|\Theta)$ is multivariate Gaussian.

Outline

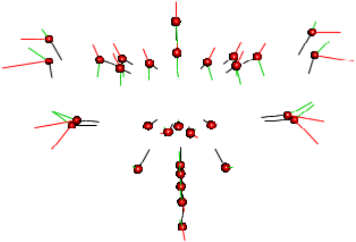
Example 3: 3D face landmarks

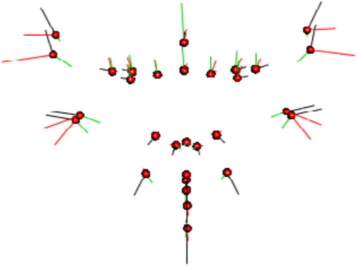


Procrustes (least squares) registered data



Principal components analysis


(b) Female


(c) Male

Ambient space regression model

$$\begin{aligned} Y_i &= \mu(x_i)\Gamma_i + \varepsilon_i, \\ &= \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)\Gamma_i + \varepsilon_i \end{aligned}$$

where β_0 lower triangular (for identifiability) $\Gamma_i \in SO(m)$ (rotations),

$$\text{vec}(\varepsilon_i) \overset{i.i.d.}{\sim} N_{km}\left(\text{vec}(\mathbf{0}), \sigma^2 I_m \otimes I_k\right).$$

Likelihood and prior

$$f(Y_1,\ldots,Y_n, \mid \beta,\Gamma_1,\ldots,\Gamma_n,\sigma^2) =$$
$$\frac{1}{(2\pi)^{nkm/2}(\sigma^2)^{nkm/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \mathrm{tr}\left[(Y_i - X_i\beta\Gamma_i)^\top (Y_i - X_i\beta\Gamma_i)\right]\right).$$

$$\kappa = 1/\sigma^2 \sim \text{Gamma}(a,b) \ ;$$
$$\Gamma_i \sim \text{matrix Fisher}(F_0), \quad i = 1,\ldots,n \ ;$$
$$p(\beta \mid \Gamma_1,\ldots,\Gamma_n,\kappa) \propto 1,$$

Posterior

The joint posterior for $(\beta, \Gamma_1, \ldots, \Gamma_n, \kappa)$ is given by

$$\begin{aligned} & p(\beta, \Gamma_1, \ldots, \Gamma_n, \kappa \mid Y_1, \ldots, Y_n) \\ \propto \quad & \exp\left(\sum_{i=1}^{n} \mathrm{tr}(F_0^\top \Gamma_i)\right) \left[\prod_{i=1}^{n} \sin\theta_{i2}\right] \kappa^{a+nkm/2-1} \exp\left(-\frac{\kappa}{b}\right) \\ & \times \exp\left(-\frac{1}{2}\kappa \sum_{i=1}^{n} \mathrm{tr}\left[(Y_i - X_i\beta\Gamma_i)^\top (Y_i - X_i\beta\Gamma_i)\right]\right). \end{aligned}$$

Regression models

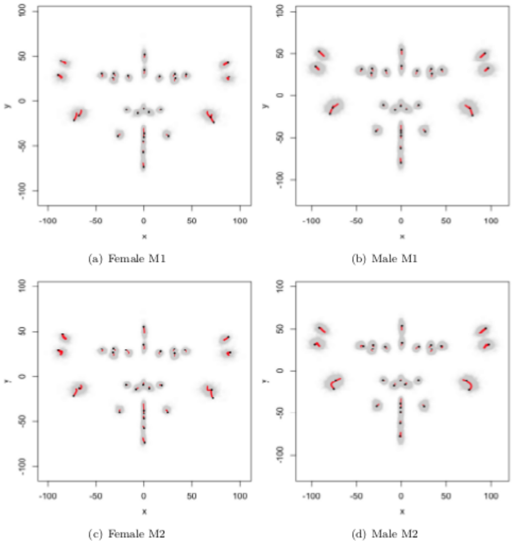
$$\begin{aligned}
\text{M1}: \quad Y_i^H &= \{\beta_0 + \beta_1 \mathtt{age}_i\}\,\Gamma_i + \varepsilon_i, \\
\text{M2}: \quad Y_i^H &= \left\{\beta_0 + \beta_1 \mathtt{age}_i + \beta_2 \mathtt{age}_i^2\right\}\Gamma_i + \varepsilon_i \\
\text{M3}: \quad Y_i^H &= \left\{\beta_0 + \beta_1 \mathtt{age}_i + \beta_2 \mathtt{age}_i^3\right\}\Gamma_i + \varepsilon_i,
\end{aligned}$$

where $Y_i^H = HY_i$. Then we define the predicted model as pre-multiplying each $\widehat{Y}_i$ by C, for example for M1,

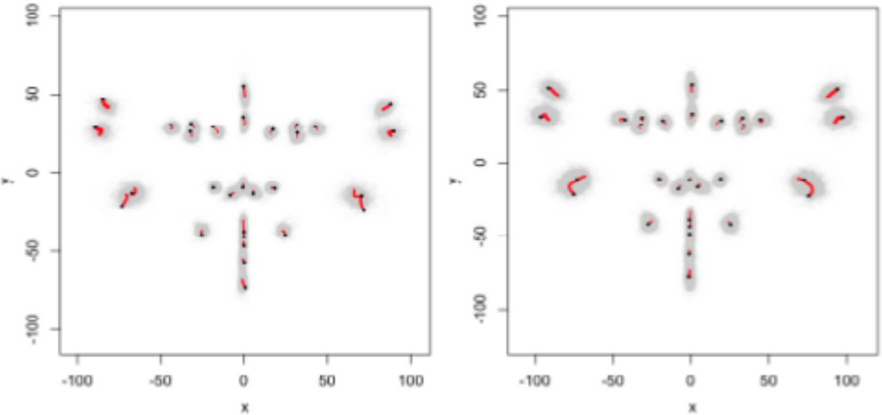
$$C\widehat{Y}_i = \left\{H^\top \widehat{\beta}_0 + H^\top \widehat{\beta}_1 \mathtt{age}_i\right\}\widehat{\Gamma}_i,$$

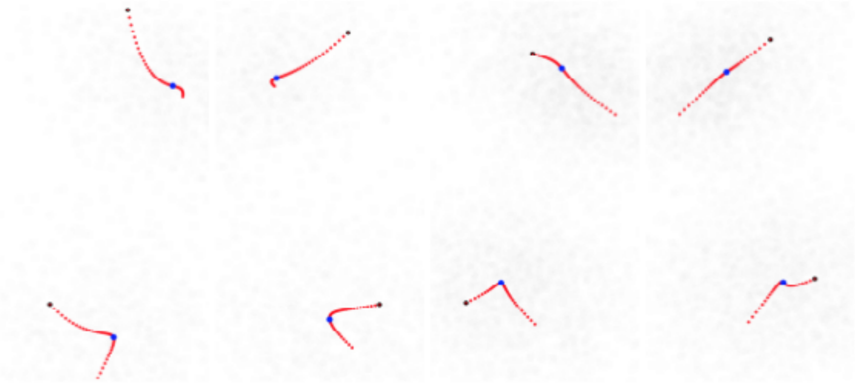
where $C = I_k - \frac{1}{k}1_k 1_k^\top$, I_k is the $k \times k$ identity matrix, 1_k is the column vector of k ones, and $\widehat{\beta}_j = \frac{1}{\mathcal{N}_\mathcal{B}} \sum_{t \in \mathcal{B}} \beta_j^{(t)}$ is the arithmetic mean of MCMC sample (100k) for β_j after burn-in (100k).

Predicted faces using M1 and M2

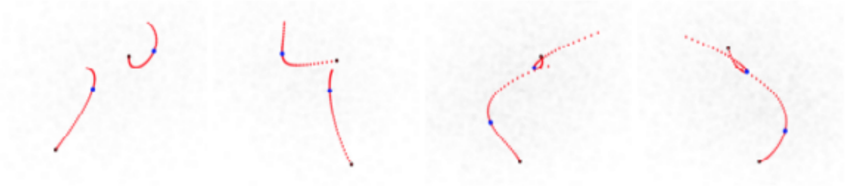


M2 chosen using AIC





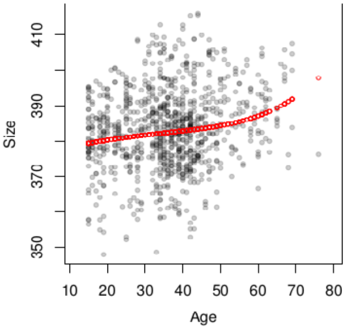
(a) Female: ear, top left. (b) Female: ear, top right. (c) Male: ear, top left. (d) Male: ear, top right.



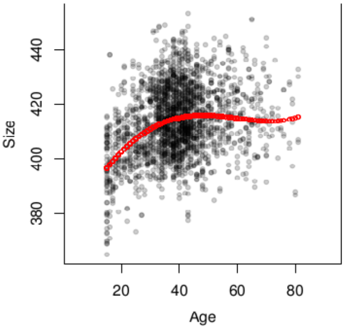
(e) Female: ear, bottom left. (f) Female: ear, bottom right. (g) Male: ear, bottom left. (h) Male: ear, bottom right.

(i) Female: lips. (j) Male: lips.

Figure 4: Ears and lip (M2).



(a) Female



(b) Male.

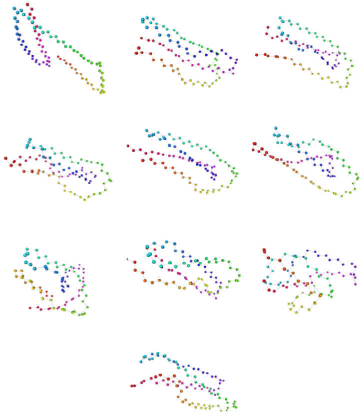
Outline

Many other applications. Function and Shape: Arteries

Mean differences

Shape and Time: Enzymes

Enzyme data $k = 88$ landmarks in 3D, time series $n = 4216$.
Some snapshots at 10 equally spaced time points.



Four PNS clusters

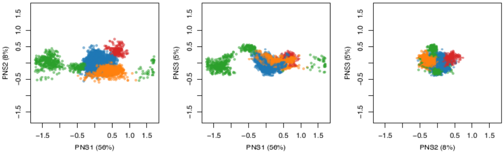


Figure 41: PNS plot with clustering color scheme.

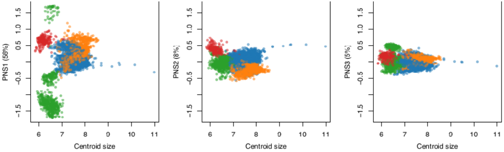
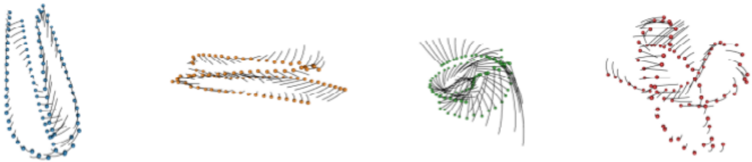


Figure 42: Centroid size vs. PNS plot.

Enzyme clusters

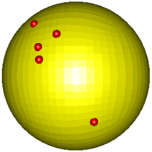
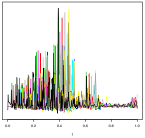


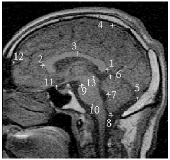
(a) Cluster 1. (b) Cluster 2. (c) Cluster 3. (d) Cluster 4.

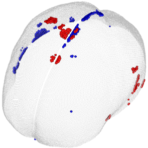
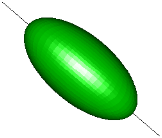
Clustering using Princpal Nested Spheres: difficult but useful.

Needed: more on methods, models and uncertainty

- Circular and spherical data
- Functions
- Dynamical systems
- Shapes and manifold data
- Images
- Trees




SELECTED REFERENCES

- Allassonnière, S., Amit, Y. and Trouvé, A. (2007). Towards a coherent statistical framework for dense deformable template estimation, *JRSS B*, **69**, 3–29.
- Cheng, W., Dryden, I. L., Hitchcock, D. B., and Le, H. (2014). Analysis of proteomics data: Bayesian alignment of functions. *Electronic Journal of Statistics*, **8**, 1734–1741.
- Cheng, W., Dryden, I. L., and Huang, X. (2016). Bayesian registration of functions and curves. *Bayesian Analysis*, **11**, 447–475.
- Dryden, I.L. and Mardia, K.V. (2016). Statistical shape analysis, with applications in R. 2nd Edition. Wiley, Chichester.

REFERENCES (CONT.)

- Green, P.J. and Mardia, K.V. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics, *Biometrika*, **93**, 235-254,
- Jung, S., Dryden, I.L. and Marron, J.S. (2012). Analysis of principal nested spheres. *Biometrika*, 99, 551-568.
- Kenobi, K. and Dryden, I.L. (2012). Bayesian matching of unlabelled point sets using Procrustes and configuration models. *Bayesian Analysis*. **7**, 547-566.
- Miolane, N., Holmes, S., Pennec, X. (2017). Template shape estimation in Computational Anatomy: Correcting an asymptotic bias. SIAM Journal of Imaging Science.
- Srivastava, A., Klassen, E., Joshi, S.H., and Jermyn, I.H. (2011). Shape Analysis of Elastic Curves in Euclidean Spaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (7), 1415-1428.

Thanks!

EPSRC grant and Royal Society Wolfson Research Merit Award.

EPSRC Engineering and Physical Sciences Research Council THE ROYAL SOCIETY

http://www.maths.nottingham.ac.uk/∼ild