STATISTICS
TEXAS A&M UNIVERSITY

Background on Functional Data and Functional PCA

November 19, 2015

Functional Data

Objects can often be represented as functions:

- Variable Stars: magnitude as a function of time
- Galaxies: flux as a function of wavelength
- Images: flux as a function of x, y location

Storing Functional Data

- Let x_i be the i^{th} functional observation.
- $x_i : \mathbb{R} \to \mathbb{R}$
- Record each function at a set of points $\lambda_1, \ldots, \lambda_p$.
- $X \in \mathbb{R}^{n \times p}$ with $X_{ij} = x_i(\lambda_j)$

Example: Synthetic galaxy photometry.

- λ_j are the effective restframe wavelengths.
- $x_i(\lambda_j)$ is flux of galaxy i in filter λ_j.

Multivariate Data

- Collect p quantities on n objects
- $X \in \mathbb{R}^{n \times p}$ is data.
  - X_{ij} is value of quantity j for object i.

Example: Features extracted from variable stars

- $X_{i1} =$ period of star i
- $X_{i2} =$ amplitude of star i
- . . .

Structure in Functional Data

Some Methodology only Appropriate for functional data:

- Derivatives: Supposing λ_j are ordered, could analyze derivatives:

$$\frac{d}{d\lambda}x_i(\lambda)\Big|_{\lambda_j} \approx \frac{x_i(\lambda_{j+1}) - x_i(\lambda_j)}{\lambda_{j+1} - \lambda_j}$$

  New data matrix X' where

$$X'_{ij} = \frac{d}{d\lambda}x_i(\lambda)\Big|_{\lambda_j}$$

- Registration: x_i may be "out of alignment" with other observations

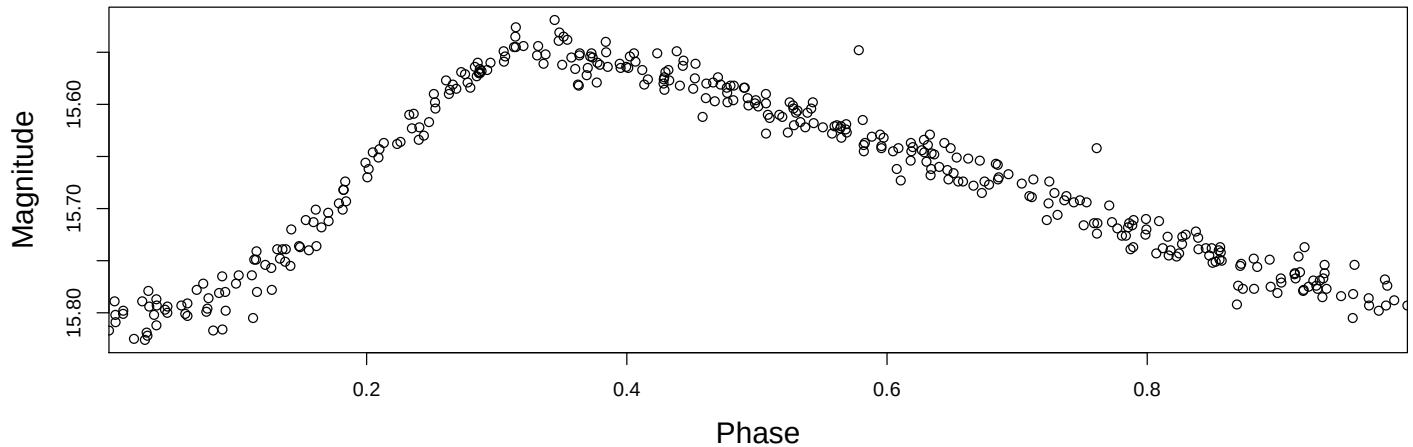
$$x_i(\lambda_j) \to x_i(\lambda_j + \lambda)$$

Quality of Functional Data

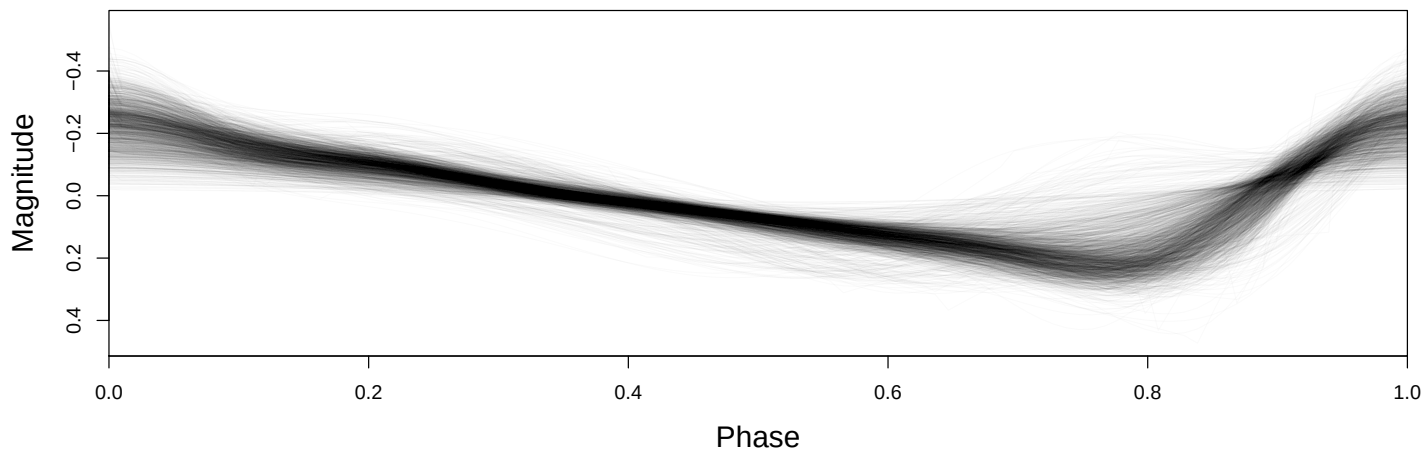
1. Well sampled
   - Infer (by smoothing) a continuous function from raw observations.
   - Produce X matrix at any grid of $\lambda_1, \ldots, \lambda_p$.
2. Sparsely sampled
   - Not possible to smooth data prior to applying statistical methodology.

Example of Well Sampled Data

Single object is densely sampled so can infer continuous function.



Large Set of Functions (smoothed and resampled):



Could represent as $X \in \mathbb{R}^{n \times p}$, $n = \#$ l.c.'s, $p = \#$ sampled phases.

Example of Sparsely Sampled Data

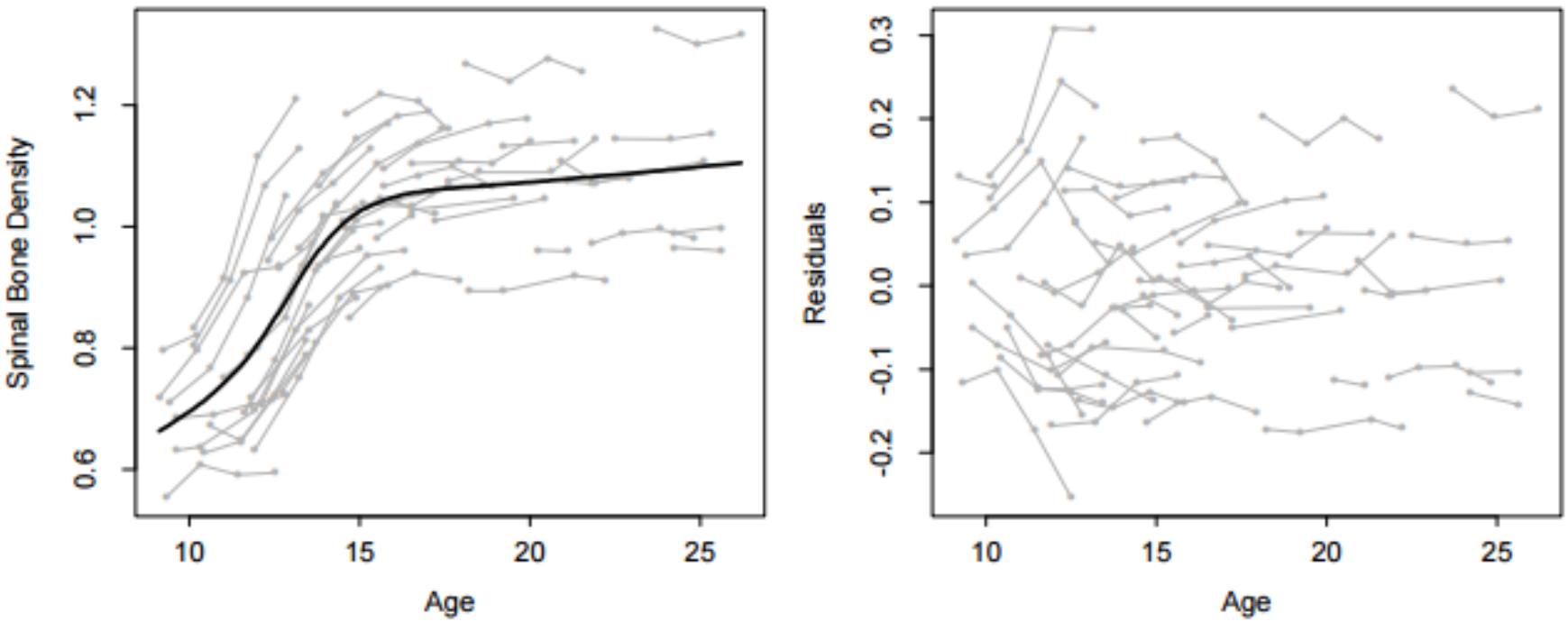
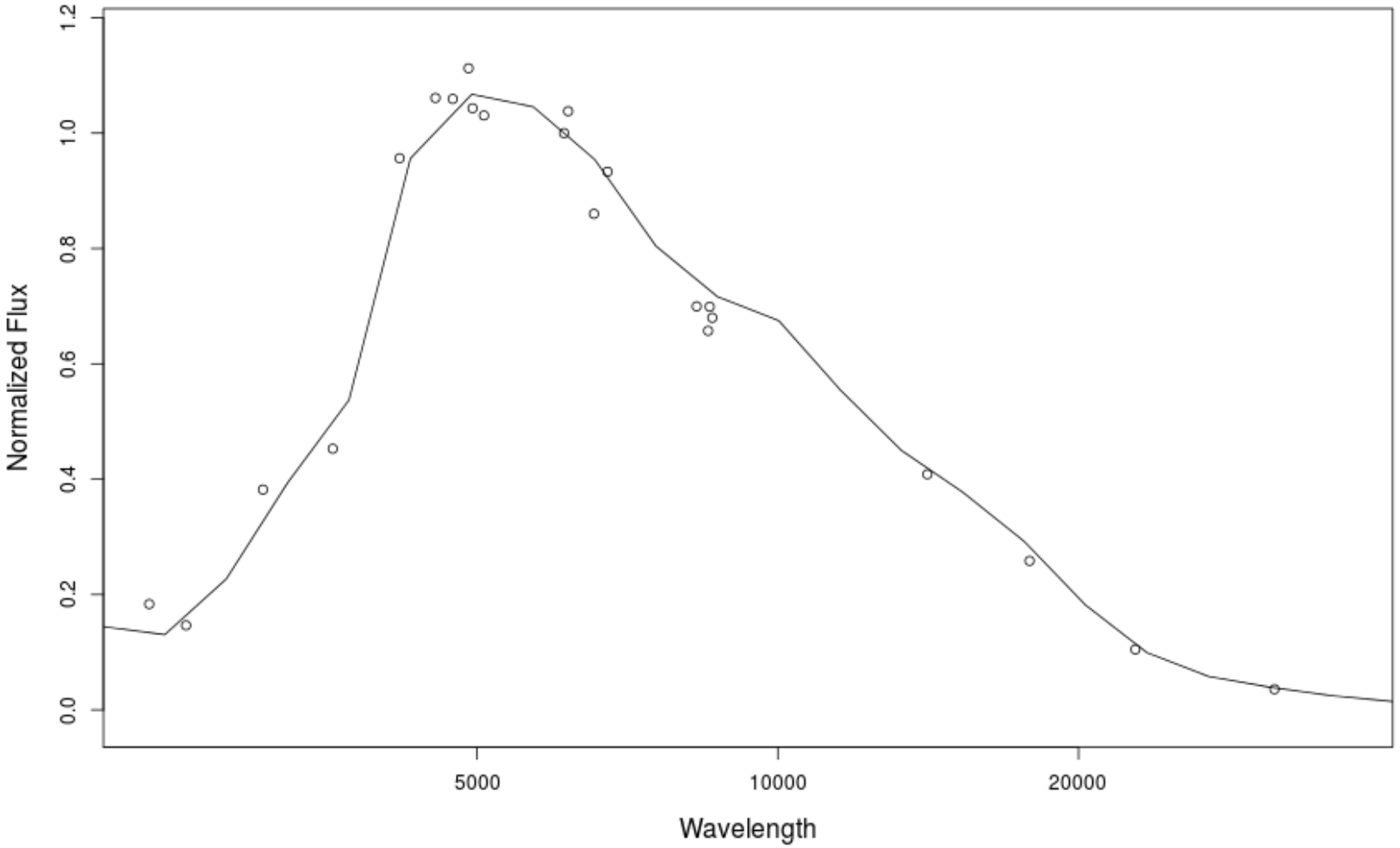


Figure 1: The data are measurements of spinal bone mineral density for forty-eight white females. There are between 2 and 4 measurements per subject (160 in all) indicated by the growth curve fragments in the plots. The solid line in (a) is an estimate for the population mean growth curve. The residuals are shown in (b). The variability of the residuals is smallest in childhood and increases slightly during the period associated with the adolescent growth spurt.

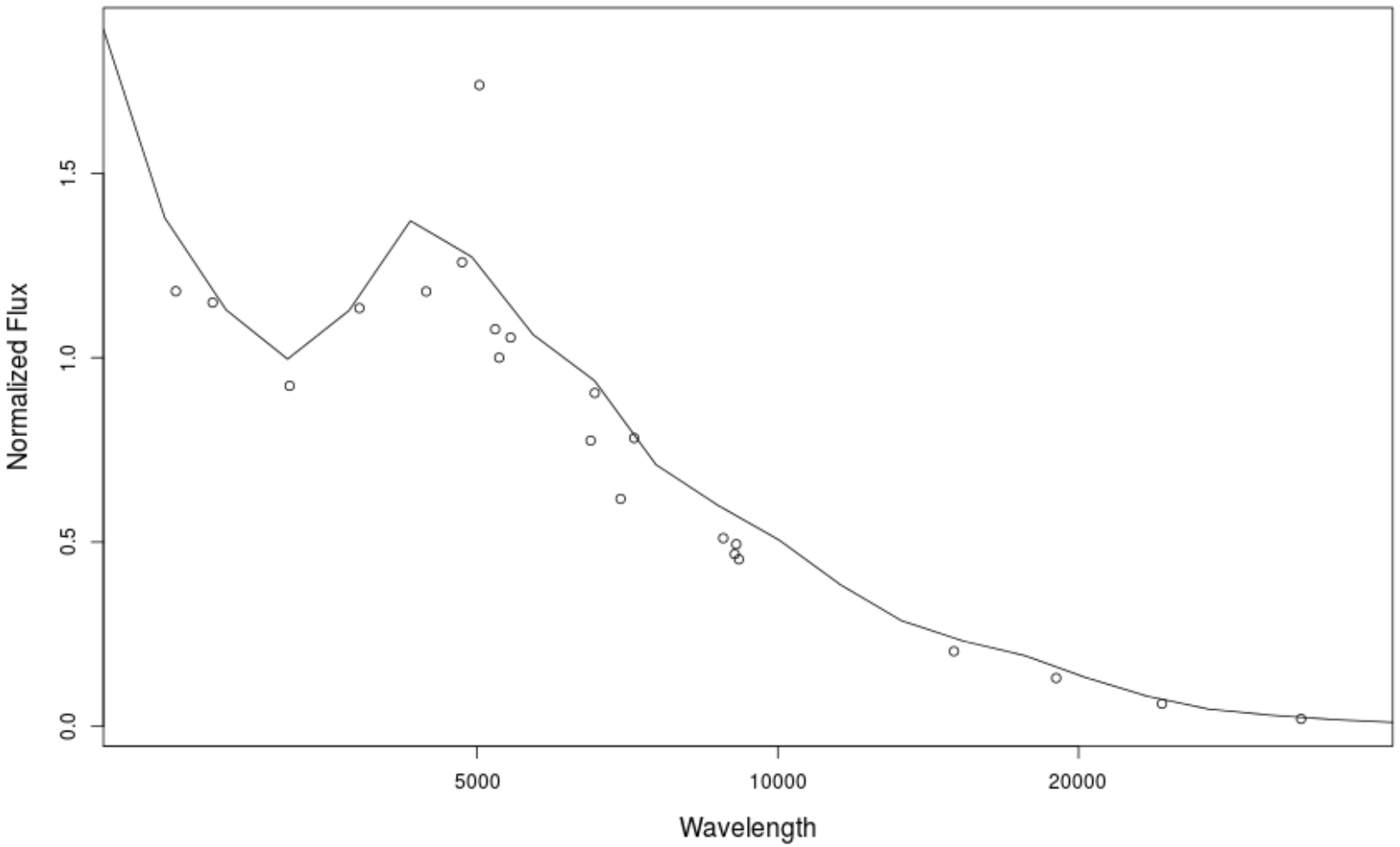
Cannot smooth these curves before applying methodology (eg PCA).

"Principal component models for sparse functional data." James, Hastie, Sugar. Biometrika 2000

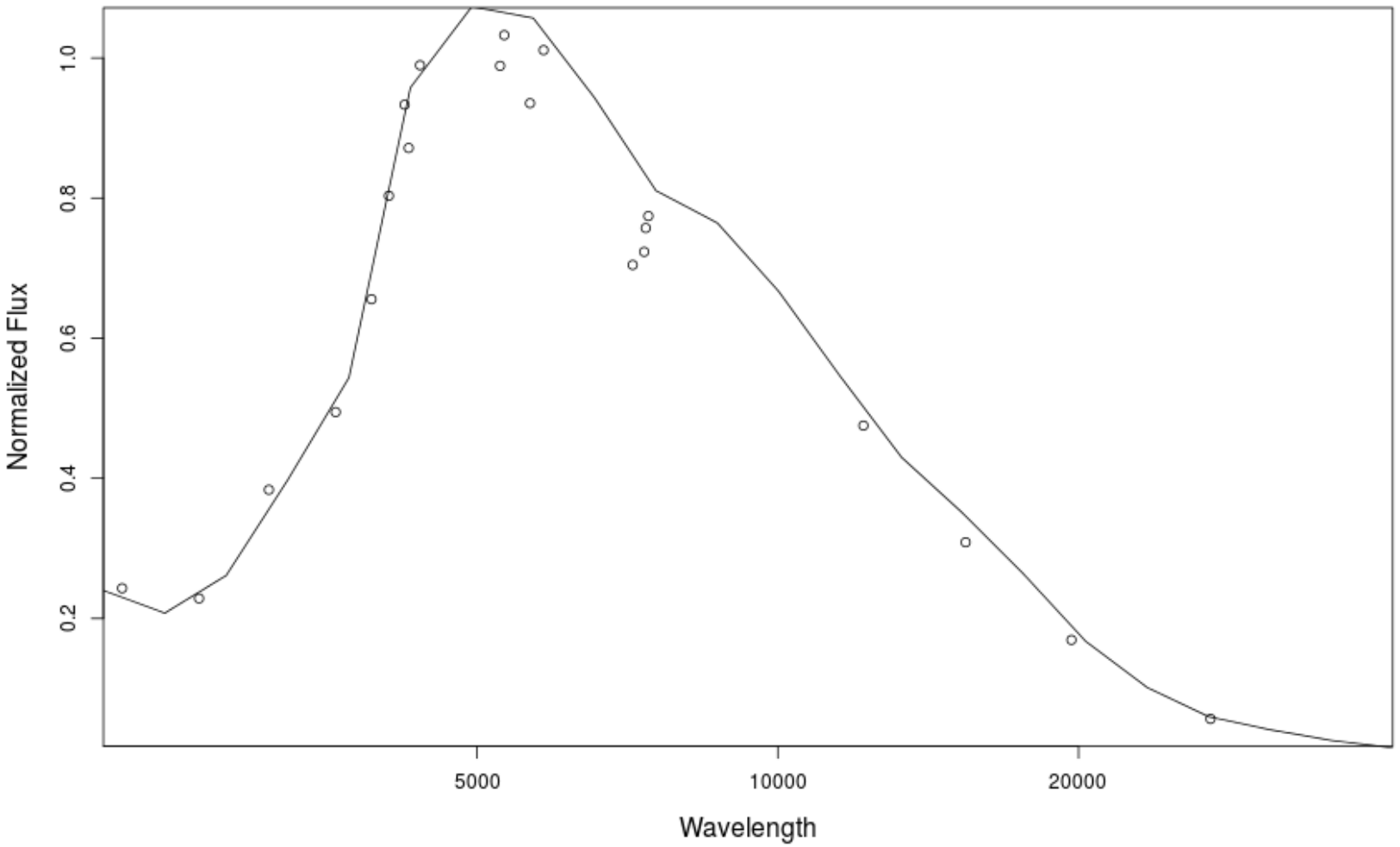
ZFORGE Data: In Between Case



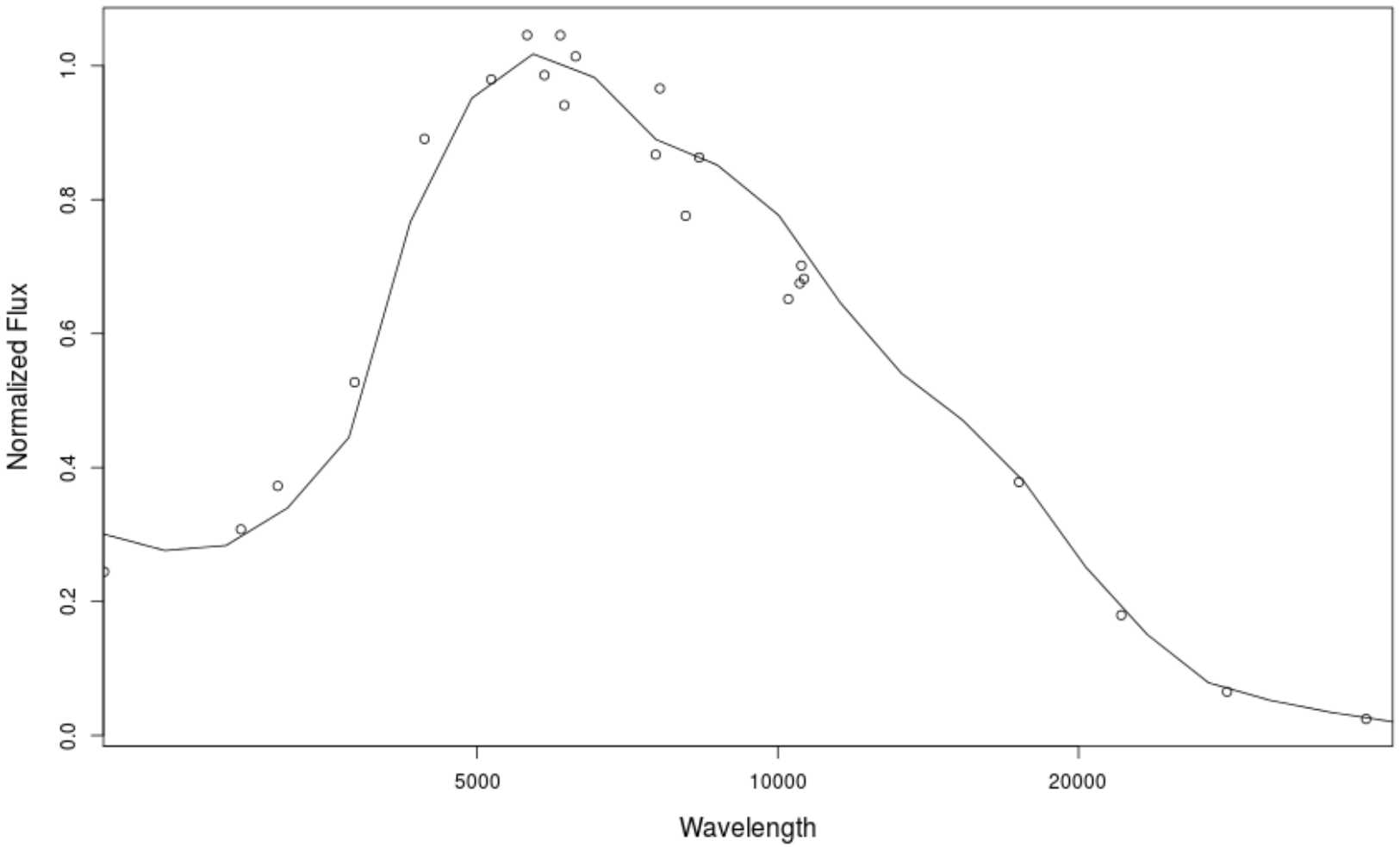
ZFORGE Data: In Between Case



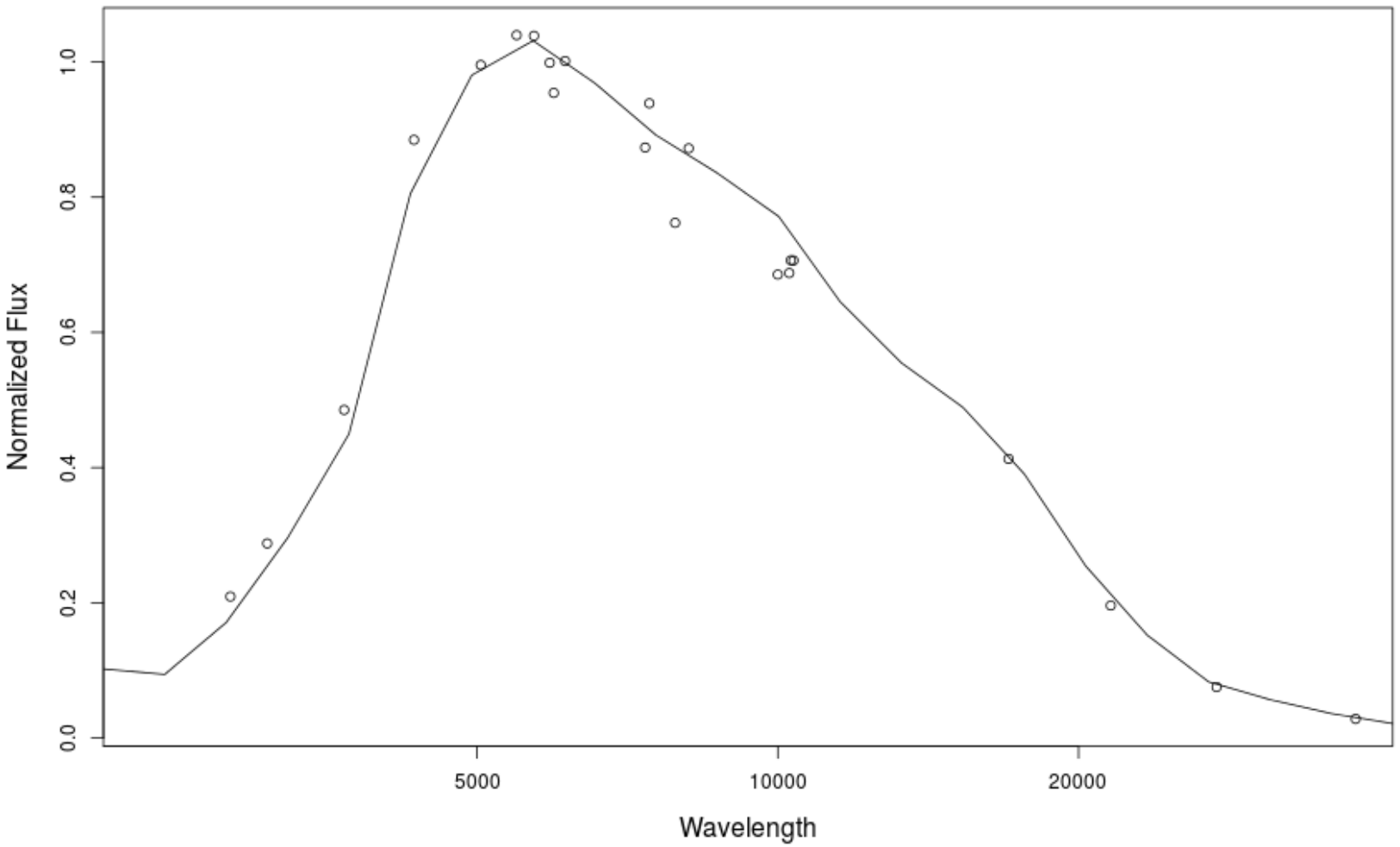
ZFORGE Data: In Between Case



ZFORGE Data: In Between Case



ZFORGE Data: In Between Case

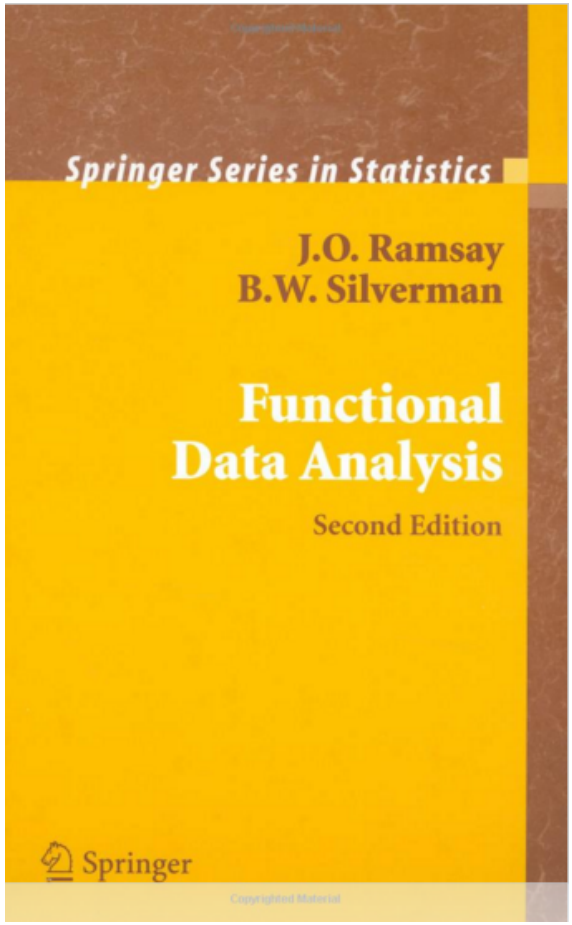


EAZY

- Easy and Accurate Photometric Redshifts from Yale (EAZY) algorithm produces synthetic photometry.
  - "EAZY: A fast, public photometric redshift code" ApJ 2008 Brammar et al.
- Algorithm makes use of physical models for galaxy spectra.
- Perhaps introduces biases into PCA results / clustering.
- May want to compare:
  - Methodology applied to synthetic photometry.
  - Methodology applied to actual photometry.

Notes

- Most functional data analysis techniques designed for well–sampled data.
- Sparse, irregularly sampled case often called longitudinal data.
  - For some discussion of terms see: "Properties of principal component methods for functional and longitudinal data analysis." Annals of Statistics. Hall et al.



PCA – Well Sampled Functions

- **Option 1:** Pretend data is multivariate (see Tuesday lecture)
- **Option 2:** Enforce smoothness constraints on extracted principal components.
  - Most appropriate when original functions are not particularly smooth.
  - This is a functional data analysis procedure, would not make any sense to apply to multivariate data.

Example: Original Data

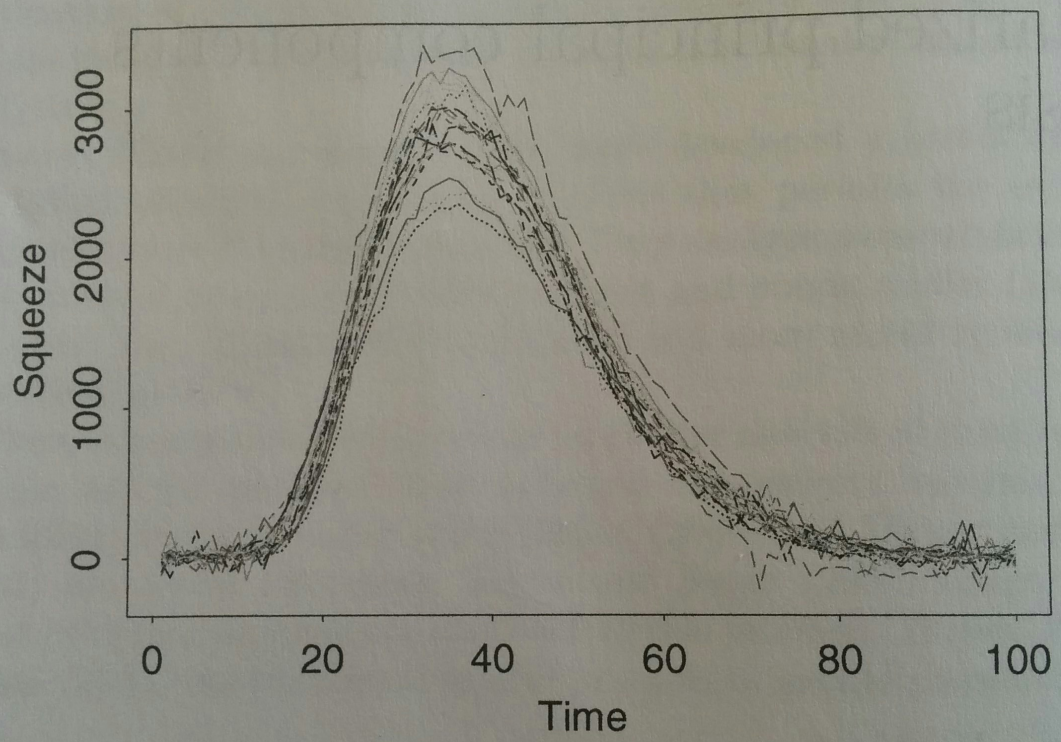


Figure 9.1. The aligned original recordings of the force relative to a baseline value exerted during each of 20 brief pinches.

Example: Principal Components from Method 1

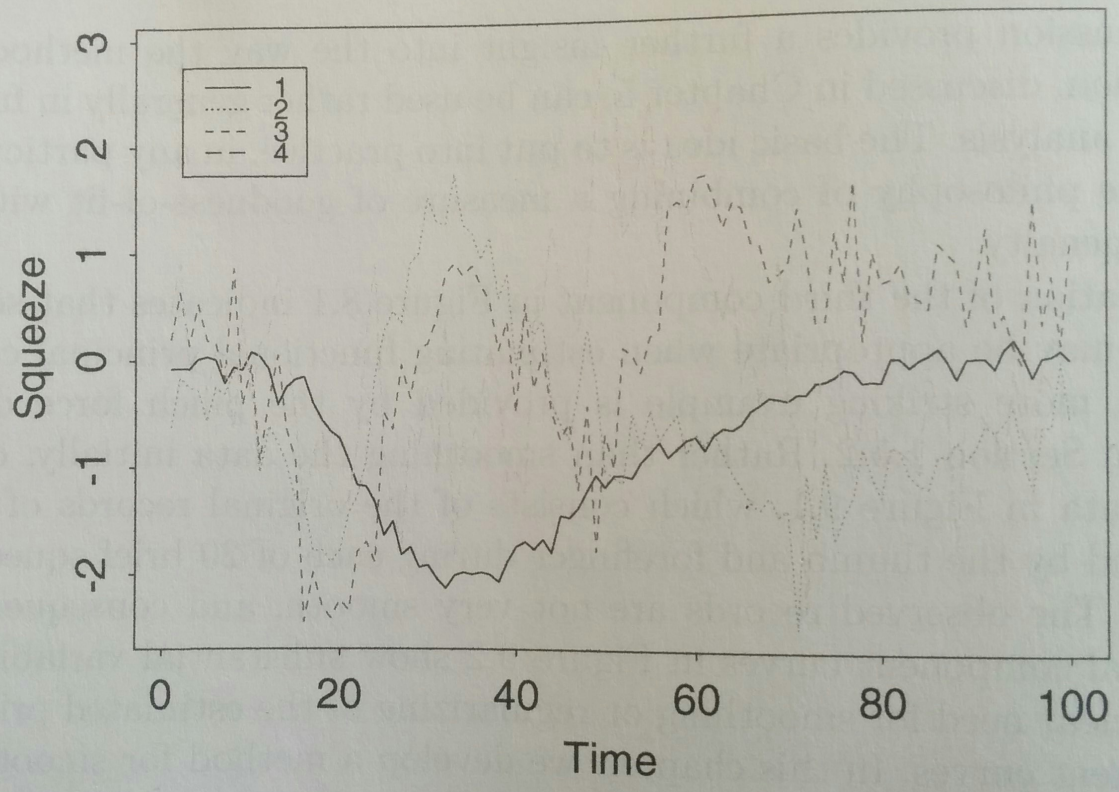


Figure 9.2. The first four principal component curves for the pinch force data without regularization.

Example: Principal Components from Method 2

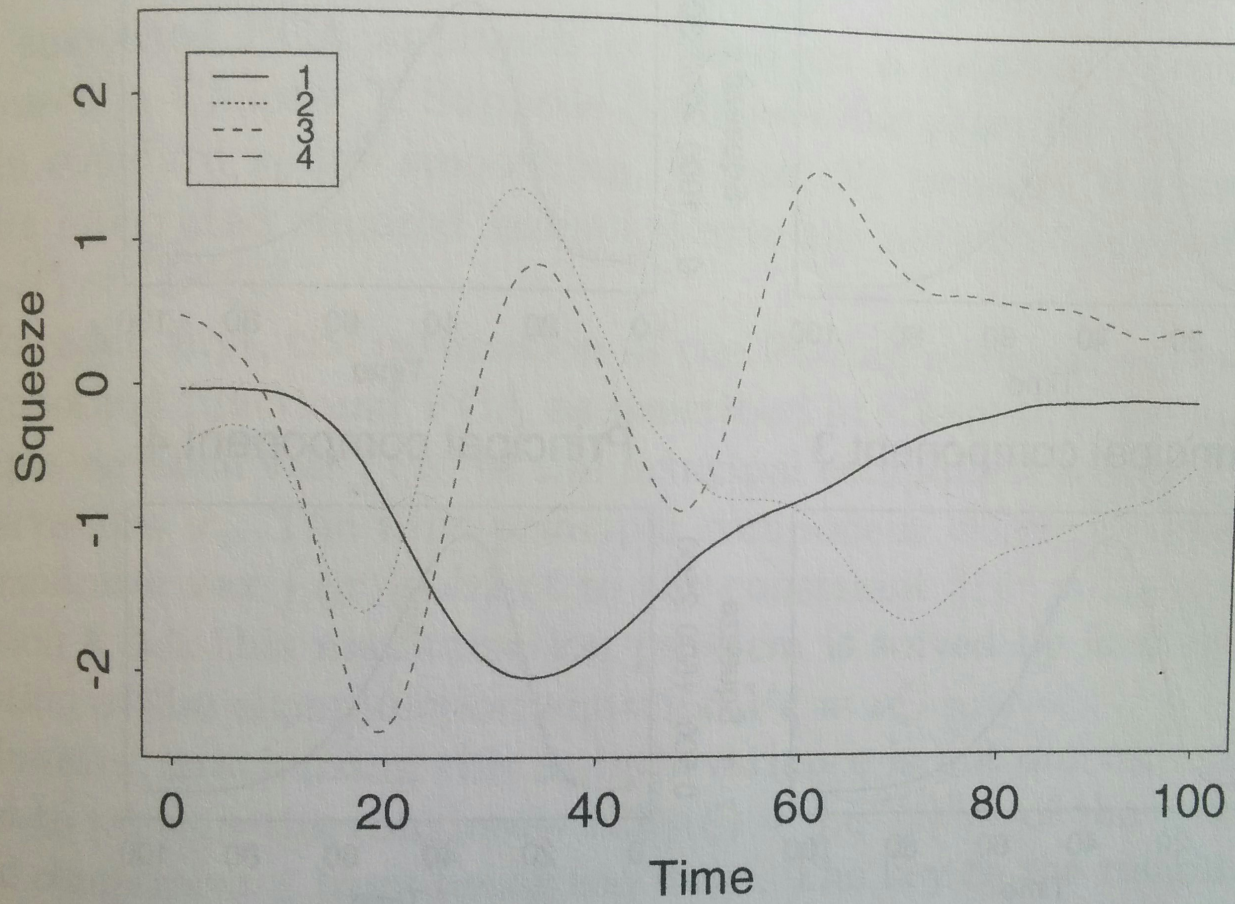
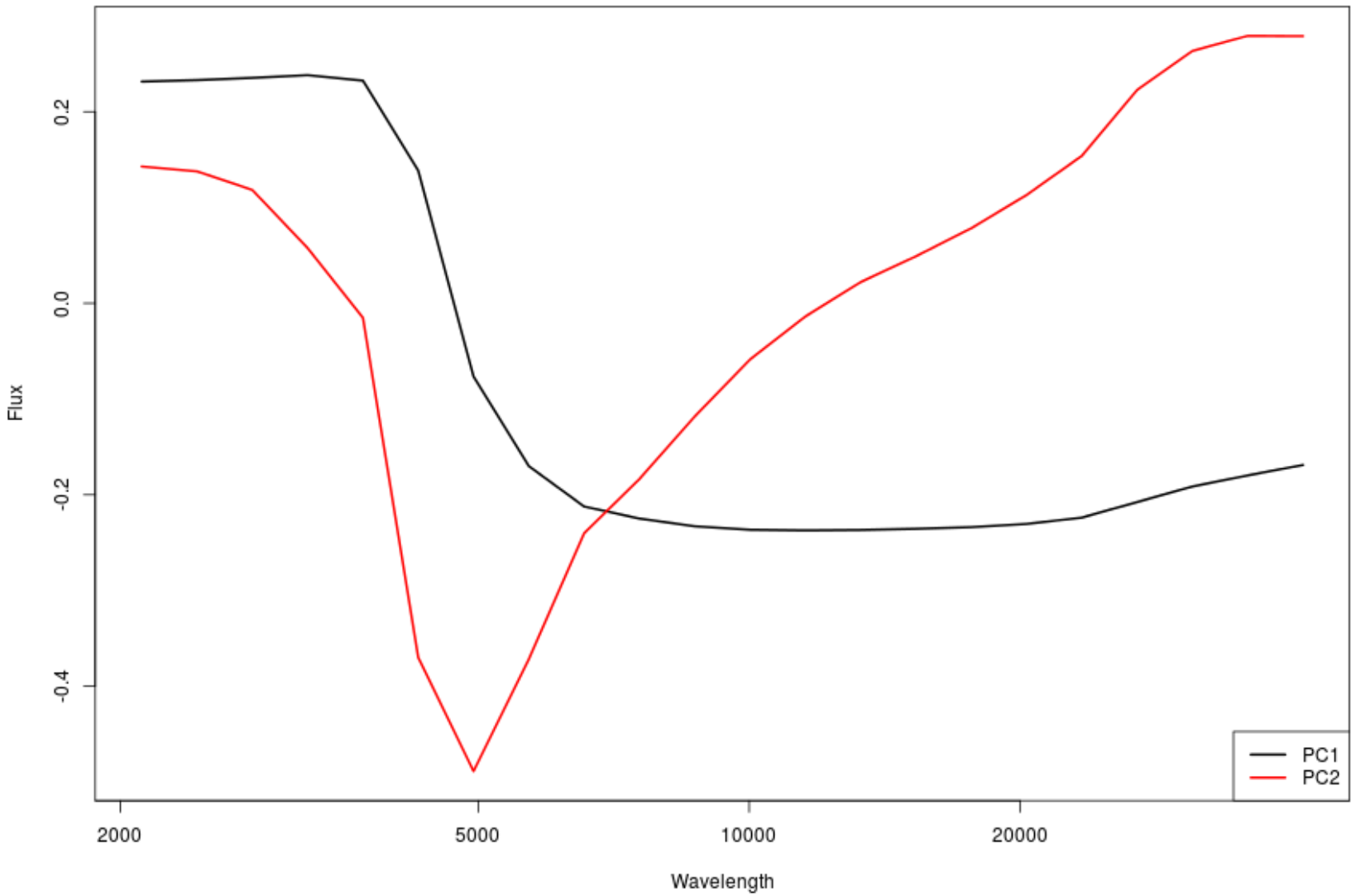


Figure 9.3. The first four smoothed principal components for the pinch force data, smoothed by the method of Section 9.3. The smoothing parameter is chosen by cross-validation.

Method 1 on ZFORGE Data



Already fairly smooth, perhaps not much to gain by using FPCA.

FPCA – Sparsely Sampled Functions

Several Recent Approaches / Theoretical Discussion

- "Principal component models for sparse functional data." James, Hastie, Sugar. Biometrika 2000.
- "Functional data analysis for sparse longitudinal data." Yao, Muller, Wang. JASA 2005.
- "Functional Modeling and Classification of Longitudinal Data." Muller. Scand. J of Stat. 2004.
- "Properties of principal component methods for functional and longitudinal data analysis." Hall et al. Annals of Stat. 2006.

Will discuss these papers on Tuesday.