Principal Components for Sparse and Irregularly Sampled Functional Data

Trevor Hastie and Lukasz Kidzinski

and our Mobilize colleagues
Apoorva Rajagopal, Eni Halilaj, Jennifer Hicks, Ya Le, Michael Schwartz and Scott Delp,

and based on earlier work with Stanford Statistics graduates
Gareth James and Catherine Sugar.
What are principal components?

Suppose our data consists of a large collection of variables with lots of correlations. Principal components allow us to reduce the set to a smaller number of derived variables — linear combinations of the originals — that capture the information (variability) in the data. This is the traditional framework.

- The first principal component of a set of variables \( X = (X_1, X_2, \ldots, X_p)' \) is the (normalized) linear combination \( Z_1 = X'v_1 = \sum_{j=1}^{p} X_j v_{1j} \) having largest variance.

- The second principal component is the linear combination \( Z_2 = X'v_2 \), uncorrelated with \( Z_1 \), having largest variance.

- \( \ldots \)
What are principal components?

Suppose our data consists of a large collection of variables with lots of correlations.
Principal components allow us to reduce the set to a smaller number of derived variables — linear combinations of the originals — that capture the information (variability) in the data. This is the traditional framework.

- The first principal component of a set of variables $X = (X_1, X_2, \ldots, X_p)'$ is the (normalized) linear combination $Z_1 = X'v_1 = \sum_{j=1}^{p} X_j v_{1j}$ having largest variance.
- The second principal component is the linear combination $Z_2 = X'v_2$, uncorrelated with $Z_1$, having largest variance.
- \ldots

*technical detail: $v_j$ are eigenvectors of covariance matrix $\Sigma$ of $X$*
Approximate each data point $x_i$ by the closest point in an affine hyperplane spanned by $v_1$ and $v_2$:

$$x_i \approx \bar{x} + z_{1i}v_1 + z_{2i}v_2$$
Approximate each data point $x_i$ by the closest point in an affine hyperplane spanned by $v_1$ and $v_2$:

$$x_i \approx \bar{x} + z_{1i}v_1 + z_{2i}v_2$$

technical detail: SVD solution solves $\min \|\tilde{X} - ZV'\|_F$ with $\tilde{X}$ the centered data matrix; $V$ also eigenvectors of $\Sigma$ as before.
PCA when observations are sampled functions?
Many sampled functions! (Cerebral Palsy Data)
Now the second view of PCA is more intuitive

\[ x_i \approx \bar{x} + z_{1i}v_1 + z_{2i}v_2 \]

- \( x_i \) is the \( i \)th sampled function — a vector here of 51 values
- \( \bar{x} \) is the mean function (light blue) — again a vector of 51 values
- \( v_1 \) and \( v_2 \) are deviation template functions, that capture the variation of the collection of \( x_i \) about the mean function.
- \( z_{i1} \) and \( z_{i2} \) are the first two principal component scores (numbers) for function \( x_i \), and summarize this function relative to the collection in terms of their representation wrt \( v_1 \) and \( v_2 \).
First two principal components $v_1$ and $v_2$
First two principal component scores
Issues

1. Data are noisy functions, leading to noisy PC functions?
Issues

1. Data are noisy functions, leading to noisy PC functions? *First fit smooth curve to each data function, then proceed as before.*
Issues

1. Data are noisy functions, leading to noisy PC functions? *First fit smooth curve to each data function, then proceed as before.*

2. Some data functions are missing a few values randomly?
Issues

1. Data are noisy functions, leading to noisy PC functions? 
   *First fit smooth curve to each data function, then proceed as before.*

2. Some data functions are missing a few values randomly? 
   *First fit smooth function to observed values, fill in the missing values, and proceed as before.*
Issues

1. Data are noisy functions, leading to noisy PC functions? First fit smooth curve to each data function, then proceed as before.

2. Some data functions are missing a few values randomly? First fit smooth function to observed values, fill in the missing values, and proceed as before. Alternatively, use matrix-completion methods to both fill in missing values and compute SVD.
Issues

1. Data are noisy functions, leading to noisy PC functions? *First fit smooth curve to each data function, then proceed as before.*

2. Some data functions are missing a few values randomly? *First fit smooth function to observed values, fill in the missing values, and proceed as before.* Alternatively, use matrix-completion methods to both fill in missing values and compute SVD.

3. Some data functions are missing a few values, and the observed values are noisy?
1. Data are noisy functions, leading to noisy PC functions? *First fit smooth curve to each data function, then proceed as before.*

2. Some data functions are missing a few values randomly? *First fit smooth function to observed values, fill in the missing values, and proceed as before.* Alternatively, *use matrix-completion methods to both fill in missing values and compute SVD.*

3. Some data functions are missing a few values, and the observed values are noisy? *As above. For matrix-completion methods, impose smoothness constraint on right singular vectors.*

Note: matrix completion methods can work in 2. and 3. even if many values are missing, while smoothing methods deteriorate.
Bigger issues

In the above, each curve was measured at the *same* set of index points.
Often this is not the case. We now deal with the case where each function consists of a small number of noisy measurements at arbitrary and different index points.
Bone mineral density as a function of age

Spinal Bone Mineral Density

Residual Variation

Age

0.6 0.8 1.0 1.2 1.4

-0.3 -0.2 -0.1 0.0 0.1 0.2 0.3

10 15 20 25

10 15 20 25
Gait summary score at different ages

Note: each fragment measured at arbitrary ages!
Stochastic process model

We assume a stochastic measurement process for subject $i$

$$X_i(s) = \mu(s) + z_{1i}f_1(s) + z_{2i}f_2(s) + \varepsilon_i(s)$$

- Measurements made at $n_i$ index values $s_{i1}, s_{i2}, \ldots, s_{in_i}$ for subject $i$. Different for each subject.
Stochastic process model

We assume a stochastic measurement process for subject $i$

$$X_i(s) = \mu(s) + z_{1i}f_1(s) + z_{2i}f_2(s) + \varepsilon_i(s)$$

- Measurements made at $n_i$ index values $s_{i1}, s_{i2}, \ldots, s_{in_i}$ for subject $i$. Different for each subject.
- $\mu(s)$ is a mean function.
Stochastic process model

We assume a stochastic measurement process for subject $i$

$$X_i(s) = \mu(s) + z_{1i}f_1(s) + z_{2i}f_2(s) + \varepsilon_i(s)$$

- Measurements made at $n_i$ index values $s_{i1}, s_{i2}, \ldots, s_{in_i}$ for subject $i$. Different for each subject.
- $\mu(s)$ is a mean function.
- $f_1(s)$ and $f_2(s)$ are (smooth) principal component functions, suitably normalized/orthogonalized, that must be estimated. We used two here for illustration — in general the rank must be chosen (as in PCA).
Stochastic process model

We assume a stochastic measurement process for subject $i$

$$X_i(s) = \mu(s) + z_{1i}f_1(s) + z_{2i}f_2(s) + \varepsilon_i(s)$$

- Measurements made at $n_i$ index values $s_{i1}, s_{i2}, \ldots, s_{in_i}$ for subject $i$. Different for each subject.
- $\mu(s)$ is a mean function.
- $f_1(s)$ and $f_2(s)$ are (smooth) principal component functions, suitably normalized/orthogonalized, that must be estimated. We used two here for illustration — in general the rank must be chosen (as in PCA).
- $z_{1i}$ and $z_{2i}$ are equivalent to the principal component scores.
Stochastic process model

We assume a stochastic measurement process for subject $i$

$$X_i(s) = \mu(s) + z_{1i}f_1(s) + z_{2i}f_2(s) + \varepsilon_i(s)$$

- Measurements made at $n_i$ index values $s_{i1}, s_{i2}, \ldots, s_{in_i}$ for subject $i$. Different for each subject.
- $\mu(s)$ is a mean function.
- $f_1(s)$ and $f_2(s)$ are (smooth) principal component functions, suitably normalized/orthogonalized, that must be estimated. We used two here for illustration — in general the rank must be chosen (as in PCA).
- $z_{1i}$ and $z_{2i}$ are equivalent to the principal component scores.
- $\varepsilon_i(s)$ is noise process.
Stochastic process model

We assume a stochastic measurement process for subject $i$

$$X_i(s) = \mu(s) + z_{1i}f_1(s) + z_{2i}f_2(s) + \varepsilon_i(s)$$

- Measurements made at $n_i$ index values $s_{i1}, s_{i2}, \ldots, s_{in_i}$ for subject $i$. Different for each subject.
- $\mu(s)$ is a mean function.
- $f_1(s)$ and $f_2(s)$ are (smooth) principal component functions, suitably normalized/orthogonalized, that must be estimated. We used two here for illustration — in general the rank must be chosen (as in PCA).
- $z_{1i}$ and $z_{2i}$ are equivalent to the principal component scores.
- $\varepsilon_i(s)$ is noise process.

*We treat $z_{1i}$ and $z_{2i}$ as random — a type of factor-analysis version of PCA for functional data.*
Stochastic process model — some details

- Represent $\mu(s)$, $f_1(s)$ and $f_2(s)$ in a basis $h(s)$ of spline functions (with designated knots)
  
  $h(s) = (h_1(s), h_2(s), \ldots, h_L(s))'$:

  $\begin{align*}
  \mu(s) &= h(s)'\theta_0 \\
  f_1(s) &= h(s)'\theta_1 \\
  f_2(s) &= h(s)'\theta_2,
  \end{align*}$

  with $\theta_0$, $\theta_1$ and $\theta_2$ each $L$-vectors of parameters.
Stochastic process model — some details

• Represent $\mu(s)$, $f_1(s)$ and $f_2(s)$ in a basis $h(s)$ of spline functions (with designated knots)
  
  $h(s) = (h_1(s), h_2(s), \ldots, h_L(s))'$:

  $$\begin{align*}
  \mu(s) &= h(s)'\theta_0 \\
  f_1(s) &= h(s)'\theta_1 \\
  f_2(s) &= h(s)'\theta_2,
  \end{align*}$$

  with $\theta_0$, $\theta_1$ and $\theta_2$ each $L$-vectors of parameters.

• Hence for subject $i$, if $x_i$ is the vector of measurements (of length $n_i$), the model amounts to

  $$x_i = H_i\theta_0 + z_{1i}H_i\theta_1 + z_{2i}H_i\theta_2 + \varepsilon_i,$$

  where $H_i$ is a $n_i \times L$ matrix of evaluations of $h(s)$ at the $n_i$ values of $s$, and $\varepsilon_i$ is a $n_i$-vector of residuals (assume iid $N(0, \sigma^2)$). We also assume $z_k \sim N(0, D_k)$, $k = 1, 2$. 

Stochastic process model — more details

• Hence marginally

\[ x_i \sim N(H_i \theta_0, \sigma^2 I + H_i \Theta D \Theta' H_i'), \]

where \( \Theta = [\theta_1, \theta_2] \) and \( D = \text{diag}[D_1, D_2] \).
Stochastic process model — more details

• Hence marginally

\[ x_i \sim N(H_i\theta_0, \sigma^2 I + H_i\Theta D\Theta' H_i'), \]

where \( \Theta = [\theta_1, \theta_2] \) and \( D = \text{diag}[D_1, D_2] \).

• Model is fit by Gaussian maximum likelihood, with help from EM algorithm*.

• Lukasz Kidzinski preparing an updated R package for working with such models.

• Extensions of this model to clustering, classification and regression with sparse functional data.

* James, Hastie, Sugar (2000) Biometrika, “Principal component models for sparse functional data”
First two principal component functions $f_1(s)$ and $f_2(s)$
First two principal component scores
Summary

We saw:

- Principal components from two different viewpoints:
  - High variance feature extraction.
  - Data approximation by linear manifold.
- Principal components of functional data, such as gait curves.
- Principal components of very sparse and irregular functional data.

The last provides a way of extracting features from irregular repeated measures data, which can be used in other analyses.
Summary

We saw:

- Principal components from two different viewpoints:
  - High variance feature extraction.
  - Data approximation by linear manifold.
- Principal components of functional data, such as gait curves.
- Principal components of very sparse and irregular functional data.

The last provides a way of extracting features from irregular repeated measures data, which can be used in other analyses.

Thank You!