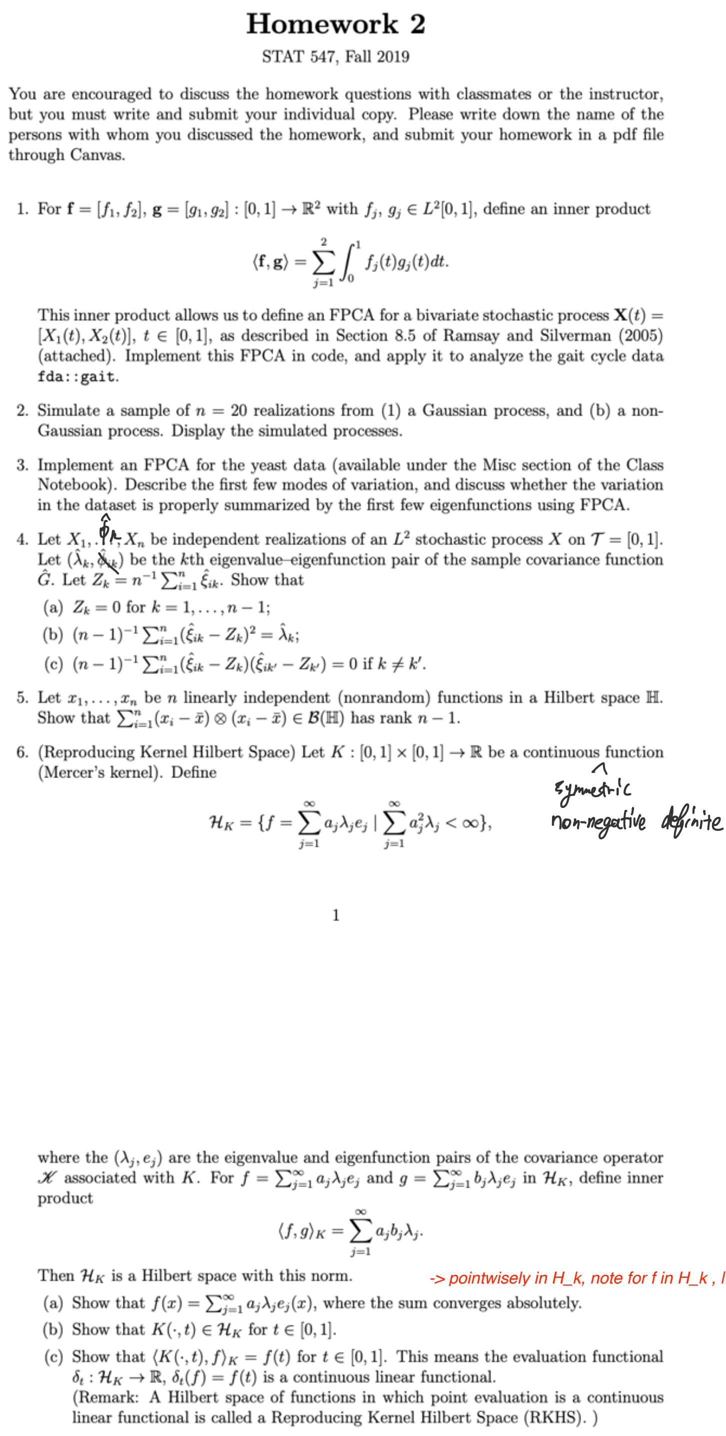
HW2

Tuesday, September 17, 2019 11:55 PM

Due on Sep 26 before class

HW2

Due Tue Oct 1 (before class)



Homework 2

STAT 547, Fall 2019

You are encouraged to discuss the homework questions with classmates or the instructor, but you must write and submit your individual copy. Please write down the name of the persons with whom you discussed the homework, and submit your homework in a pdf file through Canvas.

1. For $\mathbf{f} = [f_1, f_2]$, $\mathbf{g} = [g_1, g_2] : [0,1] \to \mathbb{R}^2$ with $f_j, g_j \in L^2[0,1]$, define an inner product
$$\langle \mathbf{f}, \mathbf{g} \rangle = \sum_{j=1}^{2} \int_0^1 f_j(t) g_j(t) dt.$$
This inner product allows us to define an FPCA for a bivariate stochastic process $\mathbf{X}(t) = [X_1(t), X_2(t)]$, $t \in [0,1]$, as described in Section 8.5 of Ramsay and Silverman (2005) (attached). Implement this FPCA in code, and apply it to analyze the gait cycle data `fda::gait`.

2. Simulate a sample of $n = 20$ realizations from (1) a Gaussian process, and (b) a non-Gaussian process. Display the simulated processes.

3. Implement an FPCA for the yeast data (available under the Misc section of the Class Notebook). Describe the first few modes of variation, and discuss whether the variation in the dataset is properly summarized by the first few eigenfunctions using FPCA.

4. Let $X_1, \dots, X_n$ be independent realizations of an L^2 stochastic process X on $\mathcal{T} = [0,1]$. Let $(\hat{\lambda}_k, \cancel{\hat{\phi}_k})$ $\hat{\phi}_k$ be the kth eigenvalue–eigenfunction pair of the sample covariance function $\hat{G}$. Let $Z_k = n^{-1}\sum_{i=1}^n \hat{\xi}_{ik}$. Show that
   (a) $Z_k = 0$ for $k = 1, \dots, n-1$;
   (b) $(n-1)^{-1}\sum_{i=1}^n (\hat{\xi}_{ik} - Z_k)^2 = \hat{\lambda}_k$;
   (c) $(n-1)^{-1}\sum_{i=1}^n (\hat{\xi}_{ik} - Z_k)(\hat{\xi}_{ik'} - Z_{k'}) = 0$ if $k \neq k'$.

5. Let $x_1, \dots, x_n$ be n linearly independent (nonrandom) functions in a Hilbert space $\mathbb{H}$. Show that $\sum_{i=1}^n (x_i - \bar{x}) \otimes (x_i - \bar{x}) \in \mathcal{B}(\mathbb{H})$ has rank $n-1$.

6. (Reproducing Kernel Hilbert Space) Let $K : [0,1] \times [0,1] \to \mathbb{R}$ be a continuous function (Mercer's kernel). Define
$$\mathcal{H}_K = \{f = \sum_{j=1}^{\infty} a_j \lambda_j e_j \mid \sum_{j=1}^{\infty} a_j^2 \lambda_j < \infty\},$$
symmetric non-negative definite

1

where the (λ_j, e_j) are the eigenvalue and eigenfunction pairs of the covariance operator $\mathscr{K}$ associated with K. For $f = \sum_{j=1}^{\infty} a_j \lambda_j e_j$ and $g = \sum_{j=1}^{\infty} b_j \lambda_j e_j$ in $\mathcal{H}_K$, define inner product
$$\langle f, g \rangle_K = \sum_{j=1}^{\infty} a_j b_j \lambda_j.$$
Then $\mathcal{H}_K$ is a Hilbert space with this norm.

   (a) Show that $f(x) = \sum_{j=1}^{\infty} a_j \lambda_j e_j(x)$, where the sum converges absolutely.
   (b) Show that $K(\cdot, t) \in \mathcal{H}_K$ for $t \in [0,1]$.
   (c) Show that $\langle K(\cdot, t), f \rangle_K = f(t)$ for $t \in [0,1]$. This means the evaluation functional $\delta_t : \mathcal{H}_K \to \mathbb{R}$, $\delta_t(f) = f(t)$ is a continuous linear functional.
   (Remark: A Hilbert space of functions in which point evaluation is a continuous linear functional is called a Reproducing Kernel Hilbert Space (RKHS).)

-> pointwisely in H_k, note for f in H_k , ||F||_k = \sum_i^n \alpha_j^2 \lambda_j

from Ramsay and Silverman 2005 Functional Data Analysis Second Edition

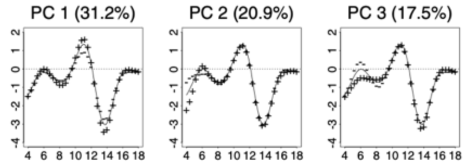


Figure 8.7. The solid curve in each panel is the mean acceleration in height in cm/year^2 for girls in the Zurich growth study. Each principal component is plotted in terms of its effect when added (+) and subtracted (−) from the mean curve.

marker events. Full details of this process can be found in Ramsay, Bock and Gasser (1995). The curves are for 112 girls who took part in the Zurich growth study (Falkner, 1960).

Figure 8.7 shows the first three eigenfunctions or harmonics plotted as perturbations of the mean function. Essentially, the first principal component reflects a general variation in the amplitude of the variation in acceleration that is spread across the entire curve, but is particularly marked during the pubertal growth spurt lasting from 10 to 16 years of age. The second component indicates variation in the size of acceleration only from ages 4 to 6, and the third component, of great interest to growth researchers, shows a variation in intensity of acceleration in the prepubertal period around ages 5 to 9 years.

8.5 Bivariate and multivariate PCA

We often wish to study the simultaneous variation of more than one function. The hip and knee angles described in Chapter 1 are an example; to understand the total system, we want to know how hip and knee angles vary jointly. Similarly, the handwriting data require the study of the simultaneous variation of the X and Y coordinates; there would be little point in studying one coordinate at a time. In both these cases, the two variables being considered are measured relative to the same argument, time in both cases. Furthermore, they are measuring quantities in the same units (degrees in the first case and cm in the second). The discussion in this section is particularly aimed towards problems of this kind.

8.5.1 Defining multivariate functional PCA

For clarity of exposition, we discuss the extension of the PCA idea to deal with bivariate functional data in the specific context of the hip and knee data. Suppose that the observed hip angle curves are $\mathtt{Hip}_1, \mathtt{Hip}_2, \ldots, \mathtt{Hip}_n$ and the observed knee angles are $\mathtt{Knee}_1, \mathtt{Knee}_2, \ldots, \mathtt{Knee}_n$. Let $\mathtt{Hipmn}$ and $\mathtt{Kneemn}$ be estimates of the mean functions of the $\mathtt{Hip}$ and $\mathtt{Knee}$ processes. Define $\upsilon_{\mathtt{HH}}$ to be the covariance operator of the $\mathtt{Hip}_i$, $\upsilon_{\mathtt{KK}}$ that of the $\mathtt{Knee}_i$, $\upsilon_{\mathtt{HK}}$ to be the cross-covariance function, and $\upsilon_{\mathtt{KH}}(t,s) = \upsilon_{\mathtt{HK}}(s,t)$.

A typical principal component is now defined by a 2-vector $\xi = (\xi^{\mathtt{H}}, \xi^{\mathtt{K}})'$ of weight functions, with $\xi^{\mathtt{H}}$ denoting the variation in the $\mathtt{Hip}$ curve and $\xi^{\mathtt{K}}$ that in the $\mathtt{Knee}$ curve. To proceed, we need to define an inner product on the space of vector functions of this kind. Once this has been defined, the principal components analysis can be formally set out in exactly the same way as previously.

The most straightforward definition of an inner product between bivariate functions is simply to sum the inner products of the two components. Suppose ξ_1 and ξ_2 are both bivariate functions each with hip and knee components. We then define the inner product of ξ_1 and ξ_2 to be

$$\langle \xi_1, \xi_2 \rangle = \int \xi_1^{\mathtt{H}} \xi_2^{\mathtt{H}} + \int \xi_1^{\mathtt{K}} \xi_2^{\mathtt{K}}. \tag{8.20}$$

The corresponding squared norm $\|\xi\|^2$ of a bivariate function ξ is simply the sum of the squared norms of the two component functions $\xi^{\mathtt{H}}$ and $\xi^{\mathtt{K}}$.

What all this amounts to, in effect, is stringing two (or more) functions together to form a composite function. We do the same thing with the data themselves: define $\mathtt{Angles}_i = (\mathtt{Hip}_i, \mathtt{Knee}_i)$. The weighted linear combination (8.4) becomes

$$f_i = \langle \xi, \mathtt{Angles}_i \rangle = \int \xi^{\mathtt{H}} \mathtt{Hip}_i + \int \xi^{\mathtt{K}} \mathtt{Knee}_i. \tag{8.21}$$

We now proceed exactly as in the univariate case, extracting solutions of the eigenequation system $V\xi = \rho\xi$, which can be written out in full detail as

$$\begin{aligned} \int v_{\mathtt{HH}}(s,t)\xi^{\mathtt{H}}(t)\,dt + \int v_{\mathtt{HK}}(s,t)\xi^{\mathtt{K}}(t)\,dt &= \rho\xi^{\mathtt{H}}(s) \\ \int v_{\mathtt{KH}}(s,t)\xi^{\mathtt{H}}(t)\,dt + \int v_{\mathtt{KK}}(s,t)\xi^{\mathtt{K}}(t)\,dt &= \rho\xi^{\mathtt{K}}(s). \end{aligned} \tag{8.22}$$

In practice, we carry out this calculation by replacing each function $\mathtt{Hip}_i$ and $\mathtt{Knee}_i$ with a vector of values at a fine grid of points or coefficients in a suitable expansion. For each i these vectors are concatenated into a single long vector Z_i; the covariance matrix of the Z_i is a discretized version of the operator V as defined in (8.7). We carry out a standard principal components analysis on the vectors Z_i, and separate the resulting principal component vectors into the parts corresponding to $\mathtt{Hip}$ and to $\mathtt{Knee}$. The

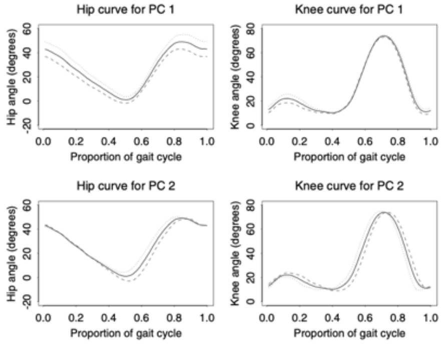


Figure 8.8. The mean hip and knee angle curves and the effects of adding and subtracting a multiple of each of the first two vector principal components.

analysis is completed by applying a suitable inverse transform to each of these parts if necessary.

If the variability in one of the sets of curves is substantially greater than that in the other, then it is advisable to consider down-weighting the corresponding term in the inner product (8.20), and making the consequent changes in the remainder of the procedure. In the case of the hip and knee data, however, both sets of curves have similar amounts of variability and are measured in the same units (degrees) and so there is no need to modify the inner product.

8.5.2 Visualizing the results

In the bivariate case, the best way to display the result depends on the particular context. In some cases it is sufficient to consider the individual parts $\xi_m^{\mathtt{H}}$ and $\xi_m^{\mathtt{K}}$ separately. An example of this is given in Figure 8.8, which displays the first two principal components. Because $\|\xi_m^{\mathtt{H}}\|^2 + \|\xi_m^{\mathtt{K}}\|^2 = 1$ by definition, calculating $\|\xi_m^{\mathtt{H}}\|^2$ gives the proportion of the variability in the mth principal component accounted for by variation in the hip curves.

For the first principal components, this measure indicates that 85% of the variation is due to the hip curves, and this is borne out by the presentation in Figure 8.8. The effect on the hip curves of the first combined principal component of variation is virtually identical to the first principal

component curve extracted from the hip curves considered alone. There is also little associated variation in the knee curves, apart from a small associated increase in the bend of the knee during the part of the cycle where all the weight is on the observed leg. The main effect of the first principal all the weight is on the observed leg. The main effect of the first principal component remains an overall shift in the hip angle. This could be caused by an overall difference in stance; some people stand up more straight than others and therefore hold their trunks at a different angle from the legs through the gait cycle. Alternatively, there may simply be variation in the angle of the marker placed on the trunk.

For the second principal component, the contributions of both hip and knee are important, with somewhat more of the variability (65%) due to the knee than to the hip. We see that this principal component is mainly a distortion in the timing of the cycle, again correlated with the way in which the initial slight bend of the knee takes place. There is some similarity to the second principal component found for the hip alone, but this time there is very substantial interaction between the two joints.

A particularly effective method for displaying principal components in the bivariate case is to construct plots of one variable against the other. Suppose we are interested in displaying the mth principal component function. For equally spaced points t in the time interval on which the observations are taken, we indicate the position of the mean function values $(\texttt{Hipmn}(t), \texttt{Kneemn}(t))$ by a dot in the (x, y) plane, and we join this dot by an arrow to the point $(\texttt{Hipmn}(t) + C\xi^{\texttt{H}}_m(t),\ \texttt{Kneemn}(t) + C\xi^{\texttt{K}}_m(t))$. We choose the constant C to give clarity. Of course, the sign of the principal component functions, and hence the sense of the arrows, is arbitrary, and plots with all the arrows reversed convey the same information.

This technique is displayed in Figure 8.9. The plot of the mean cycle alone demonstrates the overall shape of the gait cycle in the hip-knee plane. The portion of the plot between time points 11 and 19 (roughly the part where the foot is off the ground) is approximately half an ellipse with axes inclined to the coordinate axes. The points on the ellipse are roughly at equal angular coordinates — somewhat closer together near the more highly curved part of the ellipse. This demonstrates that in this part of the cycle, the joints are moving roughly in simple harmonic motion but with different phases. During the other part of the cycle, the hip angle is changing at a approximately constant rate as the body moves forward with the leg approximately straight, and the knee bends slightly in the middle.

Now consider the effect of the first principal component of variation. As we have already seen, this has little effect on the knee angle, and all the arrows are approximately in the x-direction. The increase in the hip angle due to this mode of variation is somewhat larger when the angle itself is larger. This indicates that the effect contains an exaggeration (or diminution) in the amount by which the hip joint is bent during the cycle, and is also related to the overall angle between the trunk and the legs.

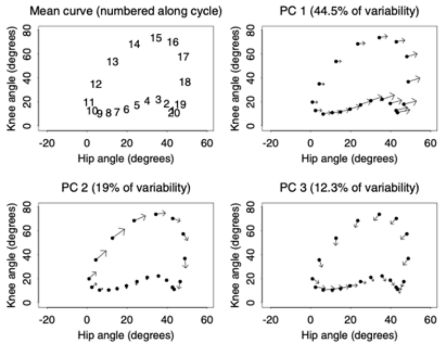


Figure 8.9. A plot of 20 equally spaced points in the average gait cycle, and the effects of adding a multiple of each of the first three principal component cycles in turn.

The second principal component demonstrates an interesting effect. There is little change during the first half of the cycle. However, during the second half, individuals with high values of this principal component would traverse roughly the same cycle but at a roughly constant time ahead. Thus this component represents a uniform time shift during the part of the cycle when the foot is off the ground.

A high score on the third component indicates two effects. There is some time distortion in the first half of the cycle, and then a shrinking of the overall cycle; an individual with a high score would move slowly through the first part of the cycle, and then perform simple harmonic motion of knee and hip joints with somewhat less than average amplitude.

8.5.3 Inner product notation: Concluding remarks

One of the features of the functional data analysis approach to principal components analysis is that, once the inner product has been defined appropriately, principal components analysis looks formally the same, whether the data are the conventional vectors of multivariate analysis, scalar functions as considered in Section 8.2.2, or vector-valued functions as in Section 8.5.1. Indeed, principal component analyses for other possible forms of functional data can be constructed similarly; all that is needed is a suitable inner product, and in most contexts the definition of such an inner product will be a natural one. For example, if our data are functions defined over a region $\mathcal{S}$ in two-dimensional space, for example temperature profiles over a geographical region, then the natural inner product will be given by

$$\int_{\mathcal{S}} f(\mathbf{s})g(\mathbf{s})d\mathbf{s},$$

and the principal component weight functions will also be functions defined over $\mathbf{s}$ in $\mathcal{S}$.

Much of our subsequent discussion of PCA, and of other functional data analysis methods, will use univariate functions of a single variable as the standard example. This choice simplifies the exposition, but in most or all cases the methods generalize immediately to other forms of functional data, simply by substituting an appropriate definition of inner product.

8.6 Further readings and notes

An especially fascinating and comprehensive application of functional principal components analysis can be found in Locantore, Marron, Simpson, Tripoli, Zhang and Cohen (1999). These authors explore abnormalities in the curvature of the cornea in the human eye, and along the way extend functional principal components methodology in useful ways. Since the variation is over the spherical or elliptical shape of the cornea, they use Zernicke orthogonal basis functions. Their color graphical displays and the importance of the problem make this a showcase paper.

Viviani, Grön and Spitzer (2005) apply PCA to repeated fMRI scans of areas in the human brain, where each curve is associated with a specific voxel. They compare the functional and multivariate versions, and find that the functional approach offers a rather better image of experimental manipulations underlying the data. They also find that the use of the GCV criterion is particularly effective in choosing the smoothing parameter prior to applying functional PCA.

While most of our examples have time as the argument, there are many important problems in the physical and engineering sciences where spectral analysis is involved. An example involving elements of both registration and principal components analysis is reported in Liggett, Cazares and Semmes (2003). Kneip and Utikal (2001) apply functional principal components analysis to the problem of describing a set of density curves where the argument variable is log income.

Besse, Cardot and Ferraty (1997) studied the properties of estimates of curves where these are assumed to lie within a finite-dimensional subspace, and where principal components analysis is used in the estimation process, and Cardot (2004) extended this work.