19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Bootstrapping functional data: a study of distributional property of sample eigenvalues

Han Lin Shang [a]

[a] *Department of Econometrics & Business Statistics, Monash University, Melbourne, Australia*
Email: HanLin.Shang@monash.edu.au

Abstract: Modern computer technology has facilitated the presence of high-dimensional data, whose graphical representations are curves, images or shapes. Because of the high-dimensionality, a dimension reduction such as functional principal component analysis or singular value decomposition is often employed. By using functional principal component analysis, a set of observed high-dimensional data can be decomposed into functional principal components and their uncorrelated principal component scores, that is,

$$f_t(x_i) = \bar{f}(x_i) + \sum_{k=1}^{\infty} \xi_{t,k}\phi_k(x_i), \quad t = 1, \dots, n, \quad i = 1, \dots, p,$$

where $\bar{f}(x_i)$ is the sample mean, $\xi_{t,k}$ is the k^{th} principal component score of observation t, and $\phi_k(x_i)$ is the k^{th} functional principal component observed at data point $\{x_1, \dots, x_p\}$.

With the aim of investigating distributional property of sample eigenvalues, we present two bootstrap methods for resampling functional data. The difference between these two techniques stem from the difference of resampling principal component scores. The bootstrap procedures can be briefly summarized as follows:

1. Hold the mean $\bar{f}(x_i)$ and $\phi_k(x_i)$ fixed at their realizations.
2. This step differs between two techniques:
   (*a*) For $t = 1, \dots, n$, generate bootstrap replication $\{\xi^*_{t,1}, \dots, \xi^*_{t,K}\}$ by sampling with replacement from $\{\xi_{t,1}, \dots, \xi_{t,K}\}$.
   (*b*) Since each set of the principal component scores follows a standard normal distribution asymptotically, generate bootstrap replication $\{\xi^*_1, \dots, \xi^*_K\}$ by sampling from independent identically distributed (i.i.d) standard multivariate normal distribution.
3. Construct the bootstrap sample $\{f^b_1(x_i), \dots, f^b_n(x_i)\}$ from the bootstrapped principal component scores.

The proposed two bootstrap procedures are applied to i.i.d functional data in simulation and to dependent functional data in empirical example. As a result, generating principal component scores from $N(0, 1)$ produces more accurate bootstrap accuracy than resampling principal component scores, and thus more accurately mimicking the behavior of sample eigenvalues and the amount of explained variation.

Keywords: Bootstrap, Eigenvalues, Functional data, Functional time series, Fertility rates

1 Introduction

Recent computer technology facilitates the presence of functional data, whose graphical representations are in the form of curve, image or shape. Many parametric statistical methods have been extended to functional data framework (see for example, Ramsay and Silverman, 1997, 2002, 2005), while Ferraty and Vieu (2006) is an excellent reference on the adaption of many nonparametric statistical methods. In functional data analysis, it is commonly assumed that random curves or functions are sampled from a stochastic process $f(x)$ in $L^2[0,r]$, where $L^2[0,r]$ is the Hilbert space of square-integrable functions on the interval $[0,r]$. The stochastic process $f(x)$ satisfies the inner product $< f, g >= \int_0^r f(t)g(t)dt$ for any two functions, $f, g \in L^2[0,r]$ and induced squared norm $\|\cdot\| =< \cdot, \cdot >$.

Among different techniques for capturing the characteristics of functional data, methods based on Karhunen-Loève decomposition (also known as functional principal component analysis) are quite popular (see, Cardot et al. 1999, 2003; Cai and Hall 2006; Hall and Hosseini-Nasab 2006; Hall and Horowitz 2007), and we have also considered this kind of method in this work. A Karhunen-Loève expansion of $f(x)$ is expressed by

$$f(x) = \mu(x) + \sum_{k=1}^{\infty} \xi_k \phi_k(x), \tag{1}$$

with the mean function $\mu(x) = \mathrm{E}[f(x)]$ and the basis functions $\phi_k(x)$ are the orthonormal eigenfunctions of the covariance kernel $\Gamma(x,z) = \mathrm{Cov}[f(x), f(z)]$. The covariance kernel $\Gamma(x,z)$ can also be approximated as

$$\int_0^r \Gamma(x,z)\phi_k(x)dx = \lambda_k \phi_k(z),$$

$$\Gamma(x,z) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x)\phi_k(z),$$

where λ_k is a set of eigenvalues in a decreasing order. The coefficient ξ_k in (1) is given by the projection of $f(x) - \mu(x)$ in the direction of the k^{th} eigenfunction $\phi_k(t)$, i.e., $\xi_k =< f(x) - \mu(x), \phi_k(x) >$. The ξ_k consists of an uncorrelated sequence of random variables with zero mean and finite variance. Asymptotically, the distribution of ξ_k can be captured by $N(0, \lambda_k^2)$.

In multivariate data analysis, the use of bootstrap techniques has been popularized since the work of Efron (1979) and Efron and Tibshirani (1993). In functional data analysis, there is comparably little work that has been done on bootstrapping functional data. In functional linear regression with functional predictors and scalar responses, Hall and Vial (2006) used the bootstrap technique by randomly sampling from i.i.d functional data to determine the optimal number of principal components, while González-Manteiga and Martínez-Calvo (2011) obtained the bootstrap samples by randomly sampling from the residuals, in order to estimate the variance of regression coefficient and construct its confidence intervals.

Hall and Vial (2006), González-Manteiga and Martínez-Calvo (2011) are two examples of bootstrapping functional data in the context of functional linear regression, and the use of bootstrap techniques allow the estimation of regression coefficient and assess its uncertainty. However, our aim is not to estimate regression coefficient, but to study the distribution of functional data. In this direction, Poskitt and Sengarapillai (2011) put forward the idea of bootstrapping functional data by randomly sampling from principal component scores (also known as basis coefficients). We follow this direction and extend it where possible.

The contribution of this paper is to propose a novel technique for bootstrapping a set of i.i.d or dependent functional data, by randomly drawing principal component scores from a standard multivariate normal distribution. Through a series of Monte Carlo studies and a real data set, the bootstrap accuracy is evaluated and compared with the existing approaches of Poskitt and Sengarapillai (2011).

The outline of this paper is given as follows. In Section 2, we describe the basic structure of functional data and its approximation by singular value decomposition. Section 3 introduces a number of nonparametric bootstrapping techniques. Via simulation studies, we compare the bootstrap accuracy among these three approaches in Section 4. Section 5 presents a real data example. Section 6 provides a summary of our main results, and some thoughts on how the bootstrap techniques might be extended.

2 BASIC STRUCTURE

Although the function $f(x)$ is of infinite dimension on the interval $[0, r]$, it is seldom observed at infinitely many data points. Instead, a finite number of data points are observed. Here, we presume that each curve is observed on a common and dense grid of p data points $x_i; i = 1, \dots, p$ with $0 \leq x_1 < x_2 < \dots, x_p \leq r$. Thus, a set of raw functional data $\boldsymbol{\mathcal{X}} = \{f_1(x), \dots, f_n(x)\}$ of n observations on some function space $\boldsymbol{\mathcal{X}}$ will consist of an $n \times p$ data matrix, where

$$f_t(x_i) = \bar{f}(x_i) + \sum_{k=1}^{\infty} \xi_{t,k}\phi_k(x_i), \quad \text{for} \quad t = 1, \dots, n, \quad i = 1, \dots, p,$$

where $\bar{f}(x_i) = \frac{1}{n}\sum_{t=1}^{n} f_t(x_i), i = 1, \dots, p$ as the sample mean of functions.

A standard approach to estimating the covariance kernel is to replace population mean function by sample mean function,

$$\hat{\Gamma}(x_i, z_j) = \frac{1}{n}\sum_{t=1}^{n}\{f_t(x_i) - \bar{f}(x_i)\}\{f_t(z_j) - \bar{f}(z_j)\}, \tag{2}$$

as an estimator of $\Gamma(x_i, z_j)$. Let $\boldsymbol{f}^c(x_i) = \boldsymbol{f}(x_i) - \bar{f}(x_i)$ be the centered functional data. The $\boldsymbol{f}^c(x_i)$ can then be approximated by singular value decomposition expressed as

$$\boldsymbol{f}^c(x_i) = n^{1/2}\boldsymbol{U}\boldsymbol{L}\boldsymbol{R}^{'}, \tag{3}$$

where $\boldsymbol{U}^{'}\boldsymbol{U} = \boldsymbol{R}^{'}\boldsymbol{R} = \boldsymbol{I}_K$, and $K = \text{rank}[f(x_i)] = \min(n-1, p)$, and the diagonal matrix $L = \text{diag}(\sqrt{l_1}, \sqrt{l_2}, \dots, \sqrt{l_K})$, where $l_1, l_2, \dots, l_K$ are non-negative singular values of $\boldsymbol{f}^c(x_i)'\boldsymbol{f}^c(x_i)/n$. The n columns of $\boldsymbol{U}$ and p columns of $\boldsymbol{V}$ are called the left singular vector and right singular vector of $\boldsymbol{f}^c(x_i)$. Notice that singular value is equal to the square of eigenvalue.

Equation (3) provides an empirical counterpart to the Karhunen-Loève expansion in that a curve in $\boldsymbol{\mathcal{X}}$ can be written as

$$f_t(x_i) = \bar{f}(x_i) + \sqrt{n}\sum_{k=1}^{K} u_{t,k}w_k\phi_k(x_i),$$

where $w_k = \sqrt{\frac{r}{p}}\sqrt{l_k}$, $\phi_k(x_i) = \frac{R_{ik}}{\sqrt{\frac{r}{p}}}$, and $\sqrt{\frac{r}{p}}$ is the adjustment made to discretization of an interval on a dense grid. Hereafter, $(u_{t,k}, w_k, \phi_k(x_i))$ will form the foundation of our subsequent methodology.

3 THE BOOTSTRAP

3.1 Random sampling with replacement from the principal component scores

Given the raw data $\boldsymbol{\mathcal{X}} = \{f_1(x), \dots, f_n(x)\}$ of n observations on $\boldsymbol{\mathcal{X}}$, an obvious way to get some idea of the sampling variability of a statistics of interest is to re-sample from $\boldsymbol{\mathcal{X}}$ and construct a bootstrap replication $\boldsymbol{\mathcal{X}^*} = \{f_1^*(x), \dots, f_n^*(x)\}$,

$$\begin{aligned}\boldsymbol{f}(x_i) &= \bar{f}(x_i) + \sqrt{n}\boldsymbol{U}\boldsymbol{L}\boldsymbol{R}^{'}, \\ \boldsymbol{f}^*(x_i) &= \bar{f}(x_i) + \sqrt{n}\boldsymbol{U}^*\boldsymbol{L}\boldsymbol{R}^{'}.\end{aligned} \tag{4}$$

From (4), we can see that the bootstrap replications of the process can be generated in the following manner. Bootstrap step

B1: Hold the mean $\bar{f}(x_i)$, the singular values $l_j, j = 1, \dots, K$, and basis functions $\phi_k(x_i), k = 1, \dots, K$ and $i = 1, \dots, p$ fixed at their realized values.

B2: For $t = 1, \dots, n$, generate bootstrap replication $u^*_{t,k}, k = 1, \dots, K$ by taking i.i.d random draw from $u_{t,k}, k = 1, \dots, K$.

B3: Construct the bootstrap sample $\boldsymbol{\mathcal{X}}^* = \{f^*_1(x), \dots, f^*_n(x)\}$ where the bootstrap realization $f^*_t(x_i), t = 1, \dots, n$ and $i = 1, \dots, p$, is constructed as in

$$f_t(x_i) = \bar{f}(x_i) + \sqrt{n}\sum_{k=1}^{K} u^*_{t,k} w_k \phi_k(x_i)$$

by replacing $u_{t,k}, k = 1, \dots, K$ by $u^*_{t,k}, k = 1, \dots, K$.

3.2 Random drawing the principal component scores from a standard multivariate normal distribution

The following adaption, for example, indicates how we can simulate different realizations of a process whose stochastic structure approximates that of the process giving rise to the original data in $\boldsymbol{\mathcal{X}}$. Simulation step

S1: Hold the mean $\bar{f}(x_i)$, the singular values $l_k, k = 1, \dots, K$, and basis functions $\phi_k(x_i), k = 1, \dots, K, i = 1, \dots, p$ fixed at their realized values.

S2: For $t = 1, \dots, n$, generate new realizations $u^*_{t,k}, k = 1, \dots, K$ by taking i.i.d random draws from a standard multivariate normal distribution.

S3: Construct the bootstrap sample $\boldsymbol{\mathcal{X}}^* = \{f^*_1(x), \dots, f^*_n(x)\}$ where for $t = 1, \dots, n$ the bootstrap realization $f^*_t(x_i), i = 1, \dots, p$, is constructed as in

$$f_t(x_i) = \bar{f}(x_i) + \sqrt{n}\sum_{k=1}^{K} u^*_{t,k} w_k \phi_k(x_i)$$

by replacing $u_{t,k}, k = 1, \dots, K$ by $u^*_{t,k}, k = 1, \dots, K$ from a multivariate normal distribution.

4 Simulation study

Since sample eigenvalues are of great importance in determining the amount of explained variation, we concentrate on the bootstrap accuracy of the sample eigenvalues. The performance of these two bootstrap methods is evaluated and compared based on the difference between the holdout sample eigenvalues and bootstrapped sample eigenvalues. To measure such a difference, we calculate the mean error (ME), mean absolute error (MAE) and mean squared error (MSE) given by

$$ME = \frac{1}{AB}\sum_{r=1}^{A}\sum_{i=1}^{B}(L^\delta - L^\delta_{r,i}),$$

$$MAE = \frac{1}{AB}\sum_{r=1}^{A}\sum_{i=1}^{B}\left|L^\delta - L^\delta_{r,i}\right|, \quad MSE = \frac{1}{AB}\sum_{r=1}^{A}\sum_{i=1}^{B}(L^\delta - L^\delta_{r,i})^2$$

where A represents the total number of simulated samples, B represents the total number of bootstrap replications, and δ indicates a specific eigenvalue.

With $A = 200$ replications, each simulated curve sample consists $n = 500$ observations and $p = 4$ dimensions. We have considered the following simulation setup

$$\text{Case (1)} \begin{cases} (l_1, l_2, l_3, l_4) = (0.9, 0.09, 0.009, 0.001) & \text{Fixed eigenvalues} \\ \boldsymbol{U} \text{ and } \boldsymbol{V} & \text{simulated from normal distribution} \end{cases}$$

$$\text{Case (2)} \begin{cases} (l_1, l_2, l_3, l_4) = (0.9, 0.09, 0.009, 0.001) & \text{Fixed eigenvalues} \\ \boldsymbol{U} \text{ and } \boldsymbol{V} & \text{simulated from uniform distribution} \end{cases}$$

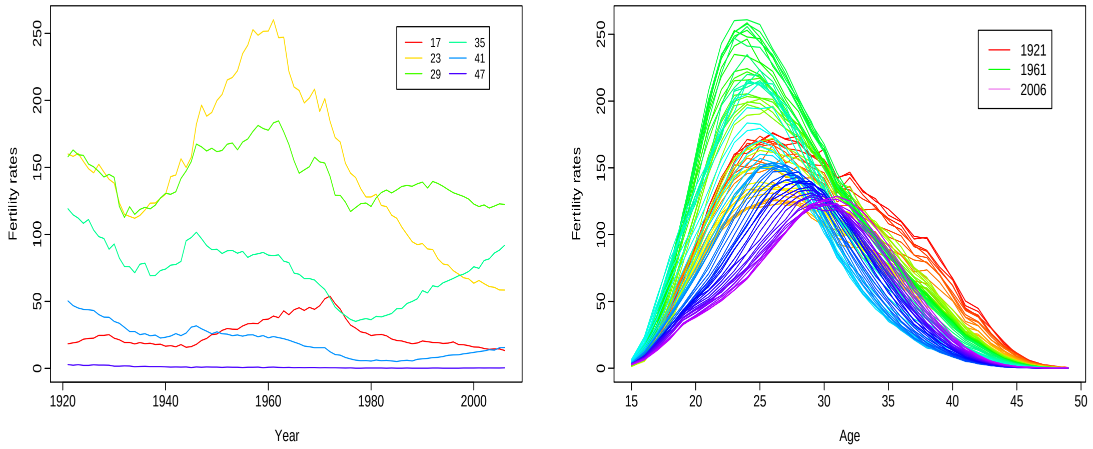
As in both simulation settings, we know the true data generating process from which we construct random matrices from $\boldsymbol{UDV}^{'}$ and apply the aforementioned bootstrap procedures to each simulated sample. To measure the performance of bootstrap procedures, Tables 1 shows the ME, MAE and MSE. As a result, the bootstrap procedure of randomly drawing principal component scores from a multivariate normal distribution performs better uniformly than the bootstrap procedure that samples with replacement from estimated principal component scores.

Table 1: ME, MAE and MSE between the holdout sample eigenvalues and bootstrap sample eigenvalues, based on $A = 200$ sample replications and $B = 1000$ bootstrap replications. Algorithm B represents the bootstrap procedure shown in Section 3.1, while algorithm S represents the bootstrap procedure shown in Section 3.2. The $\boldsymbol{U}$ and $\boldsymbol{V}$ matrices were generated from $N(0,1)$ or $U(0,1)$ distribution.

		ME		MAE		MSE	
		algorithm B	algorithm S	algorithm B	algorithm S	algorithm B	algorithm S
$N(0,1)$	$l_1 - l_1^b$	-3.0354e-03	-1.2057e-05	3.6930e-03	2.6775e-03	2.4636e-05	1.3863e-05
	$l_2 - l_2^b$	2.6688e-03	2.5331e-06	3.4005e-03	2.5498e-03	2.1587e-05	1.2765e-05
	$l_3 - l_3^b$	3.3536e-04	8.4214e-06	3.5684e-04	2.5697e-04	2.3611e-07	1.4476e-07
	$l_4 - l_4^b$	3.1228e-05	1.1028e-06	3.17928e-05	1.7086e-05	2.2459e-09	8.9076e-10
$U(0,1)$	$l_1 - l_1^b$	-1.9126e-03	-4.5128e-06	2.0513e-03	1.9457e-03	7.2403e-06	7.1412e-06
	$l_2 - l_2^b$	1.6986e-03	8.7968e-08	1.8764e-03	1.8643e-03	6.8081e-06	6.0859e-06
	$l_3 - l_3^b$	1.9668e-04	3.9063e-06	1.9884e-04	1.3561e-04	7.2884e-08	3.8775e-08
	$l_4 - l_4^b$	1.7285e-05	5.1854e-07	1.7318e-05	8.8508e-06	6.0126e-10	1.9439e-10

5 Empirical example

Fertility rates in Australia, as in most other developed countries have been falling for a considerable time. Consider annual Australian fertility rates from 1921 to 2006 for ages 15-49, obtained from the Australian Demographic Data Bank (Hyndman, 2007). These are defined as the number of live births during each calendar year, according to the age of the mother, per 1000 female resident population of the same age at 30 June.



(a) Australian fertility rates from 1921 to 2006 viewed as univariate time series for ages $\{17, 23, 29, 35, 41, 47\}$.

(b) Age-specific Australian fertility rates viewed as functional time series for ages 15–49, observed from 1921 to 2006.

Figure 1: Age-specific Australian fertility rates from 1921 to 2006 for ages 15-49.

Figure 1a shows the fertility rates viewed as a univariate time series for ages 17, 23, 29, 35, 41 and 47, while a rainbow plot (Hyndman and Shang, 2010) of functional time series is presented in Figure 1b. The colors of the functional time series plot indicate the time ordering of the curves in the same order as the colors in a rainbow with the distant past curves shown in red and most recent curves shown in purple.

Figures 1a and 1b reflect the changing social conditions affecting fertility rates. For instance, there was an increase in fertility rates in all age groups around the end of World War II, achieving a peak in 1961, followed by a rapid decrease during the 1970s due to the increasing use of contraceptive pills, and then an increase in fertility rates at higher ages in most recent years caused by a tendency to postpone childbearing while pursuing careers.

For any observed data matrix, we decompose it by singular value decomposition to obtain $\boldsymbol{U}$, $\boldsymbol{L}$ and $\boldsymbol{R}$ matrices. Here, we chose to study only the first four singular values. By holding $\boldsymbol{L}$ and $\boldsymbol{R}$ constant, we can either bootstrap $\boldsymbol{U}$ matrix by sampling with replacement or drawing random sample from a standard multivariate normal distribution. The difference in how to bootstrap $\boldsymbol{U}$ matrix leads to two different algorithms, denoted by algorithm B and algorithm S. The ME, MAE and MSE are tabulated in Table 2. The performance of algorithm S is uniformly better than algorithm B for mimicking the behavior of the sample eigenvalues and the amount of explained variation.

Table 2: ME, MAE and MSE between the holdout sample eigenvalues and bootstrap sample eigenvalues, based on $B = 1000$ bootstrap replications. Algorithm B represents the bootstrap procedure shown in Section 3.1, while algorithm S represents the bootstrap procedure shown in Section 3.2.

	ME		MAE		MSE	
	algorithm B	algorithm S	algorithm B	algorithm S	algorithm B	algorithm S
$l_1 - l_1^b$	-8.0000e-03	-6.7054e-03	1.9309e-02	1.7275e-02	5.6274e-04	4.7694e-04
$l_2 - l_2^b$	-4.1240e-04	-1.1790e-03	1.2650e-02	1.2016e-02	2.4244e-04	2.2500e-04
$l_3 - l_3^b$	-4.9186e-04	-4.2957e-04	5.6891e-03	5.2798e-03	5.1202e-05	4.4443e-05
$l_4 - l_4^b$	-3.2290e-04	1.3774e-04	3.3396e-03	3.0107e-03	1.7605e-05	1.4538e-05

6 Conclusion

In this paper, we have introduced two bootstrap methods to investigate distributional property of sample eigenvalues in functional data. These two bootstrap procedures have been applied to i.i.d data in simulation and dependent data in empirical example. As measured by mean error, mean absolute error and mean squared error, the bootstrap method of randomly drawing principal component scores from a standard multivariate normal distribution performs better in approximating sample eigenvalues than the bootstrap method that samples with replacement from the estimated principal component scores. Therefore, the former one should be used for bootstrapping functional data, since it is better to approximate the sample eigenvalues and the amount of explained variation. Although this paper mainly concentrates on studying the distributional property of sample eigenvalues, the bootstrap procedures can be applied in other contexts, such as hypothesis testing and construction of confidence intervals.

References

Cai, T. and P. Hall (2006). Prediction in functional linear regression. *The Annals of Statistics 34*(5), 2159–2179.

Cardot, H., F. Ferraty, and P. Sarda (1999). Functional linear model. *Statistics & Probability Letters 45*(1), 11–22.

Cardot, H., F. Ferraty, and P. Sarda (2003). Spline estimators for the functional linear model. *Statistica Sinica 13*, 571–591.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics 7*(1), 1–26.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis*. New York: Springer.

González-Manteiga, W. and A. Martínez-Calvo (2011). Bootstrap in functional linear regression. *Journal of Statistical Planning and Inference 141*(1), 453–461.

Hall, P. and J. Horowitz (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics 35*(1), 70–91.

Hall, P. and M. Hosseini-Nasab (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society (Series B) 68*(1), 109–126.

Hall, P. and C. Vial (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society (Series B) 68*(4), 689–705.

Hyndman, R. J. (2007). *addb: Australian demographic data bank*. R package version 3.222.

Hyndman, R. J. and H. L. Shang (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics 19*(1), 29–45.

Poskitt, D. S. and A. Sengarapillai (2011). Description length and dimensionality reduction in functional data analysis. *Computational Statistics and Data Analysis in press*.

Ramsay, J. and B. Silverman (1997). *Functional data analysis*. New York: Springer.

Ramsay, J. and B. Silverman (2002). *Applied Functional Data Analysis*. New York: Springer Series in Statistics.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis* (2nd ed.). New York: Springer Series in Statistics.