RUNNING HEAD: *PCA of Waveforms and fPCA*

PCA of Waveforms and Functional PCA: A Primer for Biomechanics

John Warmenhoven[1,2], Norma Bargary[3], Dominik Liebl[4], Andrew Harrison[5], Mark Robinson[6], Edward Gunning[3] & Giles Hooker[7,8]

[1] Exercise & Sport Science, University of Sydney

[2] People Development & Wellbeing, Australian Institute of Sport

[3] Mathematics Applications Consortium for Science and Industry (MACSI), University of Limerick

[4] Bonn Graduate School of Economics, University of Bonn

[5] Physical Education & Sport Science, University of Limerick

[6] Sport & Exercise Sciences, Liverpool John Moores University

[7] Department of Statistics and Data Science, Department of Computational Biology, Cornell University,

[8] Research School of Finance, Actuarial Science and Statistics, Australian National University

Submitted to:

Journal of Biomechanics

Corresponding Author:

Dr. John Warmenhoven

School of Exercise & Sport Science

University of Sydney

75 East Street, Lidcombe, 2141

Phone: +61 412 891 680

Email: john.warmenhoven@hotmail.com

Abstract

Principal components analysis (PCA) of waveforms and functional PCA (*f*PCA) are statistical approaches used to explore patterns of variability in biomechanical curve data, with *f*PCA being an accepted statistical method grounded within the functional data analysis (FDA) statistical framework. This technical note demonstrates that PCA of waveforms is the most rudimentary form of FDA, and consequently can be rationalised within the FDA framework of statistical processes. Mathematical proofing applied demonstrations of both techniques, and an example of when *f*PCA may be of greater benefit to control over smoothing of functional principal components is provided using an open access motion sickness dataset. Finally, open access software is provided with this paper as means of priming the biomechanics community for using these methods as a part of future functional data explorations.

Key Words (3-8): PCA, curves, statistics, variability.

PCA of Waveforms and Functional PCA: A Primer for Biomechanics

Introduction

Principal component analysis (PCA) is a classical multivariate statistical technique used for dimension reduction in human movement data (Deluzio et al., 1997). When applied to entire curves or time-series in biomechanics, it has been referred to as *PCA of waveforms* (Deluzio and Astephen, 2007; Deluzio et al., 1997; Landry et al., 2007), and has been used to transform time-series data into a smaller set of linear combinations, referred to as principal components (PCs), that account for most of the variability in the original data. These linear combinations demonstrate different patterns of variation present in the original time-series data. Similarly, the application of Functional Data Analysis (FDA) has also become common in biomechanics, particularly functional PCA (*f*PCA) (Dona et al., 2009; Kipp and Harris, 2014; Ryan et al., 2006; Warmenhoven et al., 2017). *f*PCA is an extension of PCA of waveforms tailored for use with functional data. While both methods aim to achieve the same analytical outcome, there is limited theoretical or experimental evidence demonstrating the similarities and differences between these techniques in biomechanical literature. This technical note bridges this gap, providing mathematical descriptions of both methods, supporting statistical literature and concurrent exemplar applications on two datasets via two separate experiments: *Experiment 1* (Exp. 1): a direct comparison of PCA and *f*PCA is carried out using an open access gait data dataset to demonstrate equivalence of methods; *Experiment 2* (Exp. 2): the potential benefit of using *f*PCA, smoothing of functional principal components (*f*PCs) rather than the raw data, is demonstrated using a motion sickness posturography dataset. Associated software (Matlab and R) is also supplied as supplementary information for the biomechanics community to use PCA and *f*PCA.

Datasets

Dataset 1 (DS1): Similarities in PCA of waveforms and *f*PCA were demonstrated experimentally (in Exp. 1) through application of both approaches to a data-set on children's gait collected (Olshen et al., 1989). The dataset is also openly available at www.functionaldata.org and consisted of 39 male children participating in walking, with a single gait cycle for each child. For

demonstrational purposes, only the knee joint angle was analysed and only the first two modes of variation were explored.

Dataset 2 (DS2): Demonstration of smoothing *f*PCs (in Exp. 2) was undertaken using an open access dataset that examined the relationship between reported susceptibility to motion sickness and postural control, where postural fluctuations while standing quietly were related to motion sickness history. PCA was originally conducted on the power spectrum density (PSD) curves for antero-posterior (AP) and medio-lateral (ML) axes of centre-of-pressure (CoP) measures during a static posture task. These curves were composed of amplitudes sampled at successive frequency values, which like time-points are not independent, thus making this scenario similar to PCA applied to time-series data. This data set was selected due to the combination of low and frequency content displayed along the PSD waveforms (Laboissière et al., 2015).

Statistical Techniques

PCA of Waveforms

A comprehensive mathematical description of PCA can be found in the supplementary information provided with this article. Briefly, PCA consists of an orthogonal transformation that converts the p variables (in this case time points) into p new uncorrelated principal components. The principal components are mutually uncorrelated in the sample and are arranged in decreasing order of their sample variances. The principal component model is $Z = U^tX$ where the columns of $U = u_1, u_2, \ldots, u_p$ are called principal component loading vectors, and are the eigenvectors of the covariance matrix of X (which is defined as the original data matrix). The eigenvector matrix is orthonormal; therefore, the principal component model can be inverted so that, $X = UZ$. That is, the original data can be reconstructed from the principal components. The principal component score (PC score) vectors, z_i (which represent the columns of Z), are composed of the coefficients which measure the contribution of the principal components to each individual waveform. In this way, the original waveform data for a particular subject is transformed into a set of PC scores that measure the degree to which the shape of their waveform corresponds to each feature.

As *f*PCA is mathematically an extension of PCA of waveforms for use with functional data, several preliminary steps are required prior to application. An initial step usually involves representing each time-series as a function using a suitably chosen basis expansion and then smoothing these functions, with these processes being linked. The choice of expansion is often dependent on the properties of the data being analysed (inclusive of B-splines, Fourier series, wavelets, etc.).

For spline basis functions, the interval over which a function is divided for approximation (with these referred to as breakpoints or knots) can be identified prior to analysis. The derived functions are smoothed by adding a roughness penalty to the fitting procedure, with the influence of the roughness penalty controlled through a smoothing parameter 'λ'. This penalty ensures that the smoothness of each fitted curve is controlled by minimising the penalised residual sum of squares term, where fit to the data is balanced by the smoothness of the resulting curves (Dona et al., 2009; Ramsay and Silverman, 2005). Generalised cross-validation (GCV) is often used as a starting point for determining possible values of λ before a final subjective choice is made (Ramsay and Silverman, 2005).

When applying *f*PCA computationally, it is necessary to convert functional data to a finite number of dimensions (i.e. data points), rather than analyse infinite dimensional object (i.e. mathematical functions). For *f*PCA, one way of reducing the infinite dimensional eigenequation to a discrete or matrix form is to express each function x_i as a linear combination of k basis functions Φ. These are often used when practically implementing *f*PCA.

An *f*PCA then consists of an orthogonal transformation that converts these curves now represented in the functional domain to a new set of uncorrelated principal component functions. In *f*PCA, eigenfunctions rather than eigenvectors are used to represent principal components (also referred to functional principal components or *f*PCs). Similar to PCA of waveforms, *f*PC scores measure the degree to which the shape of their waveform corresponds to each feature, with a comprehensive mathematical description for deriving *f*PCs and *f*PC scores being provided in the supplementary material.

Smoothing fPCs

We can evaluate the effectiveness of smoothing *f*PC's from their ability to represent unseen data; we might expect roughness in the *f*PCs to reflect random variation in the data that was used to

obtain them and that might therefore be different in new data. Smoothing over this roughness will then let us more closely match unseen data (Silverman, 1996). To obtain a smoothed *f*PCA, control over the roughness of the *f*PCs is accounted for by a roughness penalty. The amount of smoothing is controlled by a smoothing parameter (α),which is applied to the roughness penalty (Ramsay and Silverman, 2002). A cross-validation (CV) process was trialled for a grid of values for αranging from $1e^{-12}$ to $1e^{-1}$, to assess performance for smoothing *f*PCs (Ramsay and Silverman, 2005). Practically, this involved examination of the ability to approximate a test set of held-out curves using 1, 2, 3, 4 or 5 principal components at different values of α.That is, linear regression was performed to predict the value of the held-out curves from the smoothed principal components and summed the squared error associated with this reconstruction (Ramsay and Silverman, 2005).

Results

Exp. 1

For *f*PCA a roughness penalty was selected using GCV (see Figure 1), with a range of possible values of λ being trialled, before a final choice was made (λ = 3). Given the GCV criterion relates directly to estimating the predictive error for different values of λ, a starting point for selecting λ is the smallest GCV value of those trialled. No additional smoothing was applied before applying PCA of waveforms.

The first two modes of variation are displayed in Figure 2 (top and middle subplots). Positive scoring curves for both PCs and *f*PCs are illustrated by the plus (+) signs, and negative scorers are indicated by the minus (−) signs. These plots show clear descriptive similarities between the two techniques. Positive scorers on the first mode of variation (PC1 and *f*PC1) demonstrated greater knee flexion through the first half of the gait cycle, with this switching to a reduction in knee angle displacement leading into peak knee flexion angle (~75% of the gait cycle). The reverse was true for negative scorers for both approaches. For the second mode of variation (PC2 and *f*PC2), positive scorers demonstrated a consistently greater knee flexion angle across the whole cycle, with the reverse being true for negative scorers.

Scores for both PCA of Waveforms and *f*PCA were nearly identical (bottom two subplots; Figure 2). This was consistent across both the first and second modes of variation, where R^2 values of 0.99 were noted between PC and *f*PC scores for both components. It is likely any subtle differences between PCA of Waveforms and *f*PCA in the correlation between scores is attributable to the smoothing during function fitting for *f*PCA.

Exp. 2

For all of 1, 2, 3, 4, or 5 *f*PCs, test-set reconstruction error was lower at some positive value of alpha compared with performing no smoothing. To choose smoothing parameters, we applied cross-validation to the original training data (results displayed in Figure 3), leaving one curve out in turn for each smoothing parameter value and measuring our ability to represent it from principal components estimated from the remaining data. Using the value of alpha chosen by cross validation, we find an improvement in test set performance for all but using 2 principal components (see Table 1).

In the case of 4 principal components, representation error decreased by 14%. Thus, smoothing has benefits both for visual interpretation as well as for representing future data. How much improvement is obtained will depend on the smoothness of the underlying functional data, but it may be substantial. A demonstration of a smoothed and un-smoothed PC can be found in Figure 4, with an exemplar comparison for the first two PCs being demonstrated in Figure 5.

Discussion

Why do we consider movements as functional data rather than time-series?

Functions are viewed as time-continuous stochastic processes. When evaluated at some regular grid of points, $t_1, \ldots, t_p$, we can get a time-series $x_1(t_1), \ldots, x_i(t_p)$ in discrete time, resembling a classical time-series process. However, it is also possible to evaluate the time-continuous process, $x_1(t)$, at some other (finer or rougher) grid $t_1, \ldots, t_k$, where k is a different total number of points to p. As such, denoting $x_1(t_1), \ldots, x_i(t_p)$ as a time-series process philosophically could be viewed as confusing, given the majority of literature related to time-series analysis methods considers their application in *discrete* rather than *continuous* time.

In practice, PCA of waveforms applies PCA to the time series of observations with no smoothing applied. This makes PCA of waveforms one method of implementing *f*PCA as already demonstrated in the biomechanics literature (Liebl et al., 2014), and as the results from Exp 1 show. However, the representation of biomechanics data as functions provides considerable advantages when going beyond the application of *f*PCA. This includes incorporating data taken at different temporal resolutions, across different sampling rates, with the possibility of incorporating data from multiple sources. It also allows for the analysis of derivatives – velocity or acceleration – and the alignment of, or analysis of, the timing of events between curves.

Why incorporate smoothing as a part of the analysis process?

Since the mid 1980's splines have been suggested as a useful alternative for smoothing and processing biomechanical data (Woltring, 1985). Woltring identified that optimally regularized, natural quintic spline functions were useful for smoothing and differentiation up to at least the second derivative, but admitted that a potentially better approach would be to use B-splines, which are more stable than the piecewise polynomials (i.e. quintic splines). B-splines are integrated into the analysis framework for applying *f*PCA as demonstrated in the present study. Further to this, there are demonstrated benefits for using spline based smoothing approaches for handling curve endpoint distortions (Zin et al., 2020). Vint and Hinrichs (Vint and Hinrichs, 1996) compared four popular smoothing methods, Butterworth digital filter, Fourier series, cubic spline, and quintic spline, in terms of root mean squared (RMS) residual errors of acceleration in endpoint regions using a modification (Lanshammar, 1982) of Pezzack et al.'s (Pezzack et al., 1977) raw angular displacement data. Quintic splines produced the most accurate acceleration values, which were close to the endpoints the original (Pezzack et al., 1977) dataset, when compared to the other three methods. One cautionary note, is that B-splines may not be suitable for all types of data in human movement, and FDA offers a suite of other options for function fitting depending on the properties of the data (i.e. Fourier and wavelet approaches) (Ramsay and Silverman, 2005; Warmenhoven et al., 2017). Finally, as demonstrated within the current study, smoothing *f*PCs rather than the raw data can lead to more interpretable modes of variability (see Figure 3) and also more accurate reconstruction of the original signals (Table 1).

Software

This demonstration was conducted using the Matlab software (provided from www.functionaldata.org). For researchers looking to explore the application of PCA or *f*PCA on their own data, supplementary files with R and Matlab code are available with this article, and comprehensive tutorials in Markdown (R) and Publisher (Matlab) are available at: https://github.com/johnwarmenhoven/PCA-FPCA.

Conclusion

Theoretically and experimentally, PCA of waveforms was demonstrated as a form of *f*PCA, providing very similar results to *f*PCA of functional data represented by a basis expansion (i.e. B-Splines). PCA of waveforms can therefore be viewed as a part of the FDA family of statistical processes, as the most basic form of *f*PCA. There are however benefits to using *f*PCA, which have been outlined, with B-splines also being demonstrated as a useful expansion process for human movement data. It should also be noted that equivalence of statistical methods occur commonly in applied research, with examples in biomechanics already being demonstrated in the equivalence of statistical non-parametric mapping and FDA hypothesis testing (Warmenhoven et al., 2018). Importantly this article provides the biomechanics community with tutorials that accompany this article for applying these techniques in future research.

Funding

Dr. Norma Bargary is supported in part by Grants from Science Foundation Ireland (Grant No. 12/RC/2289-P2, 16/RC/3918, 12/RC/2275_P2, 18/CRT/6049) and co-funded under the European Regional Development Fund. Edward Gunning is supported in part Science Foundation Ireland (Grant No. 18/CRT/6049) and co-funded under the European Regional Development Fund.

Conflict of Interest Statement

The authors have no conflict of interest to declare.

References

Deluzio, K., Astephen, J., 2007. Biomechanical features of gait waveform data associated with knee osteoarthritis: an application of principal component analysis. Gait & posture 25, 86-93.

Deluzio, K.J., Wyss, U.P., Zee, B., Costigan, P.A., Serbie, C., 1997. Principal component models of knee kinematics and kinetics: normal vs. pathological gait patterns. Human Movement Science 16, 201-217.

Dona, G., Preatoni, E., Cobelli, C., Rodano, R., Harrison, A.J., 2009. Application of functional principal component analysis in race walking: an emerging methodology. Sports Biomechanics 8, 284-301.

Kipp, K., Harris, C., 2014. Patterns of barbell acceleration during the snatch in weightlifting competition. Journal of sports sciences 33, 1467-1471.

Laboissière, R., Letievant, J.-C., Ionescu, E., Barraud, P.-A., Mazzuca, M., Cian, C., 2015. Relationship between spectral characteristics of spontaneous postural sway and motion sickness susceptibility. PloS one 10, e0144466.

Landry, S.C., McKean, K.A., Hubley-Kozey, C.L., Stanish, W.D., Deluzio, K.J., 2007. Neuromuscular and lower limb biomechanical differences exist between male and female elite adolescent soccer players during an unanticipated side-cut maneuver. The American journal of sports medicine 35, 1888-1900.

Lanshammar, H., 1982. On practical evaluation of differentiation techniques for human gait analysis. Journal of Biomechanics 15, 99-105.

Liebl, D., Willwacher, S., Hamill, J., Brüggemann, G.-P., 2014. Ankle plantarflexion strength in rearfoot and forefoot runners: A novel clusteranalytic approach. Human movement science 35, 104-120.

Olshen, R.A., Biden, E.N., Wyatt, M.P., Sutherland, D.H., 1989. Gait analysis and the bootstrap. The annals of statistics, 1419-1440.

Pezzack, J., Norman, R., Winter, D., 1977. An assessment of derivative determining techniques used for motion analysis. Journal of biomechanics 10, 377-382.

Ramsay, J.O., Silverman, B.W., 2002. Applied functional data analysis: methods and case studies. Springer, New York.

Ramsay, J.O., Silverman, B.W., 2005. Functional data analysis. Wiley Online Library.

Ryan, W., Harrison, A., Hayes, K., 2006. Functional data analysis of knee joint kinematics in the vertical jump. Sports Biomechanics 5, 121-138.

Silverman, B.W., 1996. Smoothed functional principal components analysis by choice of norm. The Annals of Statistics 24, 1-24.

Vint, P.F., Hinrichs, R.N., 1996. Endpoint error in smoothing and differentiating raw kinematic data: an evaluation of four popular methods. Journal of biomechanics 29, 1637-1642.

Warmenhoven, J., Cobley, S., Draper, C., Harrison, A.J., Bargary, N., Smith, R., 2017. Considerations for the use of functional principal components analysis (fPCA) in sports biomechanics: examples from on-water rowing. Sports Biomechanics In Press.

Warmenhoven, J., Harrison, A., Robinson, M.A., Vanrenterghem, J., Bargary, N., Smith, R., Cobley, S., Draper, C., Donnelly, C., Pataky, T., 2018. A force profile analysis comparison between functional data analysis, statistical parametric mapping and statistical non-parametric mapping in on-water single sculling. Journal of science and medicine in sport 21, 1100-1105.

Woltring, H.J., 1985. On optimal smoothing and derivative estimation from noisy displacement data in biomechanics. Human Movement Science 4, 229-245.

Zin, M.A.M., Rambely, A.S., Ariff, N.M., Ariffin, M.S., 2020. Smoothing and Differentiation of Kinematic Data Using Functional Data Analysis Approach: An Application of Automatic and Subjective Methods. Applied Sciences 10, 2493.

Figure Captions

Figure 1. GCV for selection of λ as a smoothing parameter for function fitting using B-splines.

Figure 2. The first (top) and second (middle) modes of variation for PCA of Waveforms and *f*PCA. Scores (bottom) for both approaches are also compared.

Figure 3. Cross validation on the original training data (first subplot), leaving one curve out in turn for each smoothing parameter and measuring our ability to represent it from principal components estimated from the remaining data (using the summed squared error of reconstruction). This was then validated on a test-set (second sub-plot).

Figure 4. The original spectral density curves from experiment 2 (first subplot), with the an unsmoothed PC1 (second subplot) and smoothed PC1 (third subplot) using a smoothing parameter derived via CV.

Figure 5. PC1 and PC2 from experiment 2 unsmoothed (subplot1), and smoothed using the smoothing parameter derived via CV (subplot 2).

Figure 1

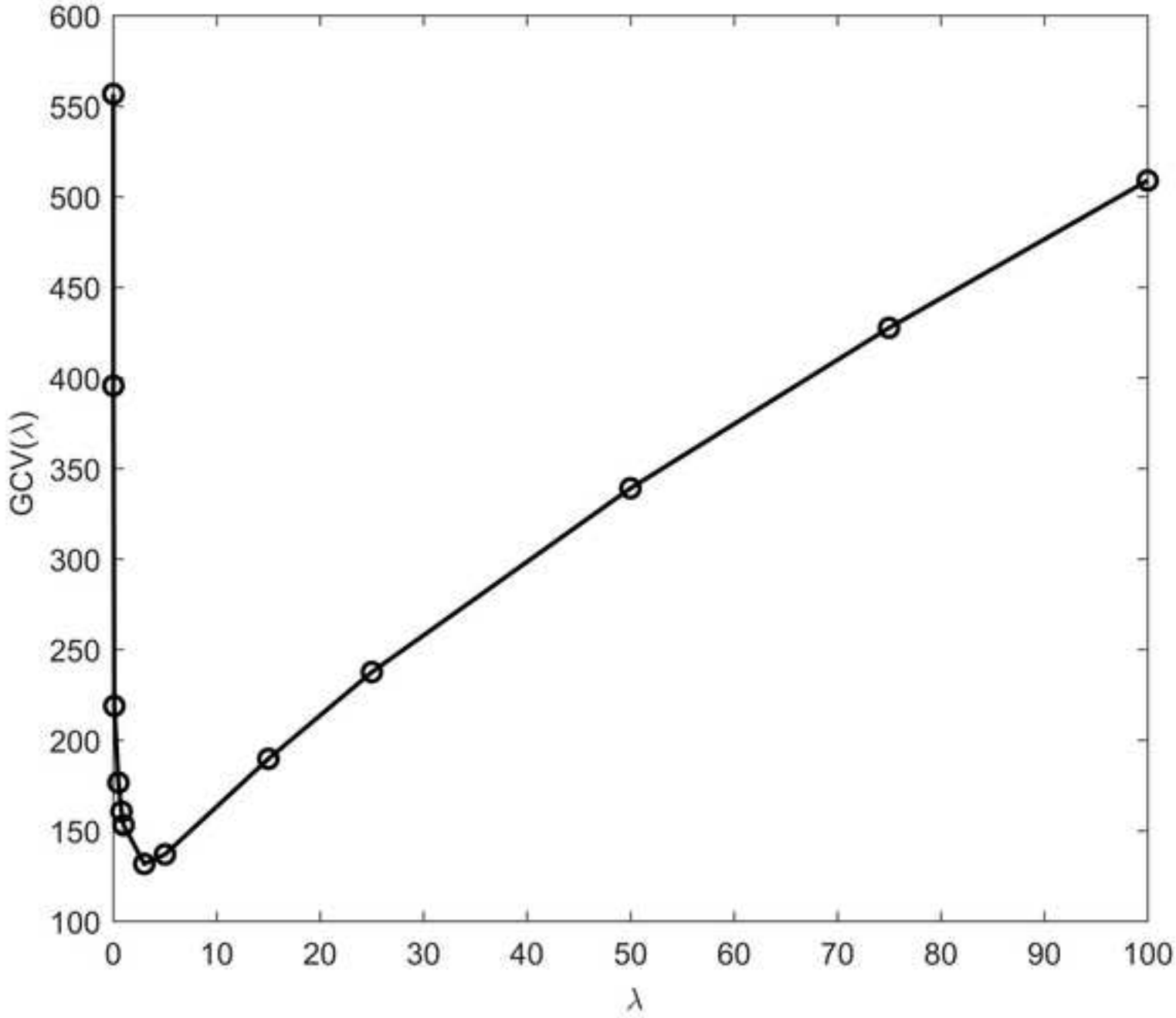

600
550
500
450
400
350
300
250
200
150
100
GCV(λ)
0
10
20
30
40
50
60
70
80
90
100
λ

Figure 2

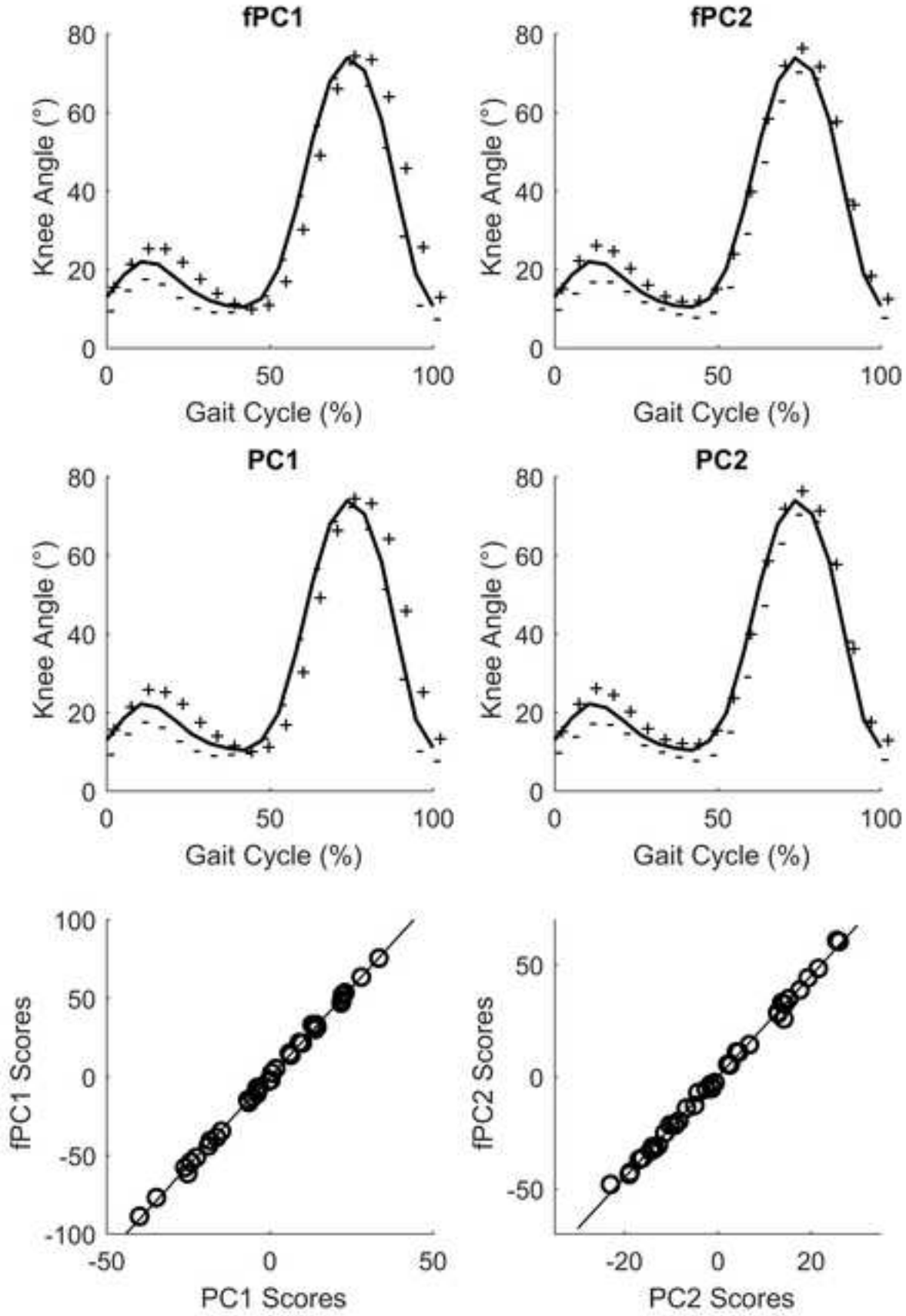
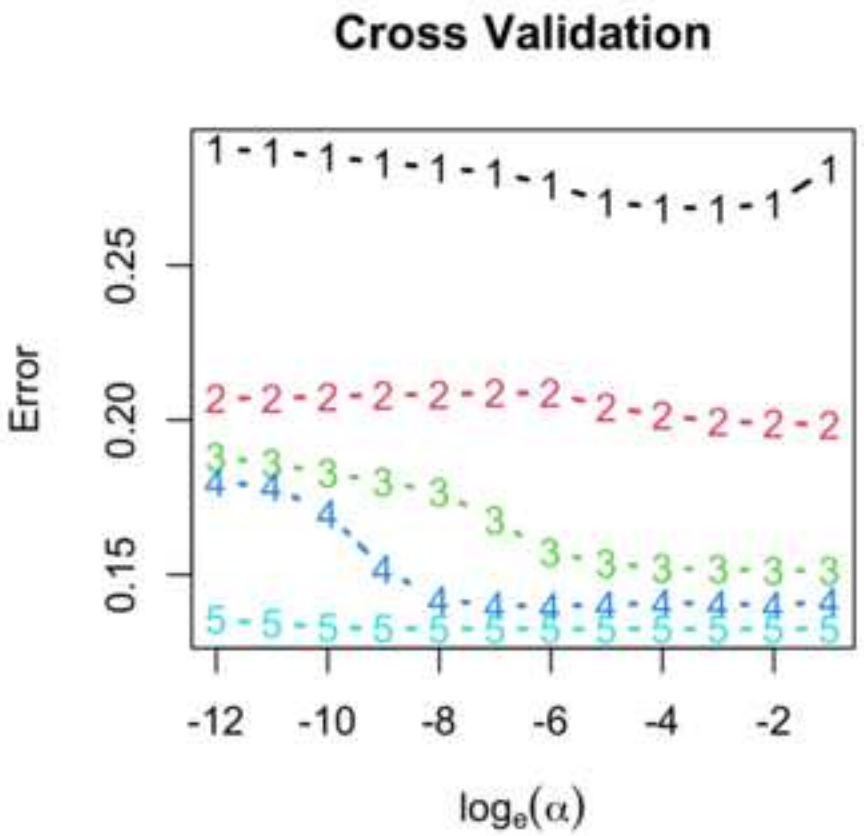

fPC1
fPC2
PC1
PC2
Knee Angle (°)
Gait Cycle (%)
fPC1 Scores
PC1 Scores
fPC2 Scores
PC2 Scores

Figure 3


Cross Validation
Error
0.15
0.20
0.25
-12
-10
-8
-6
-4
-2
log$_e(\alpha)$

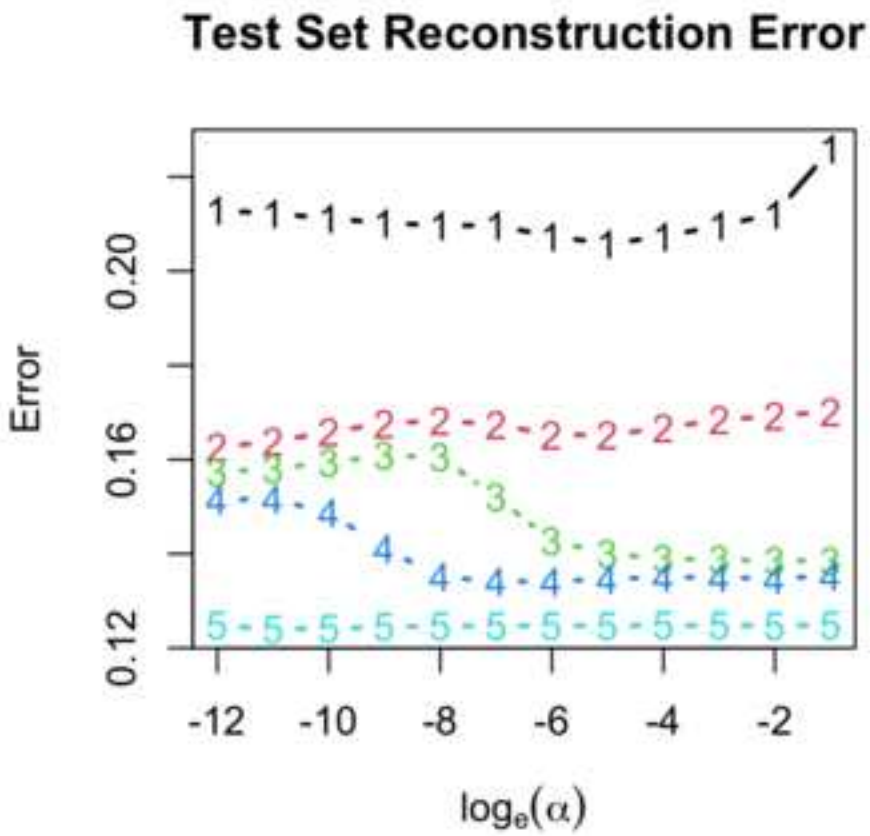

Test Set Reconstruction Error
Error
0.12
0.16
0.20
-12
-10
-8
-6
-4
-2
log$_e(\alpha)$

Figure 4

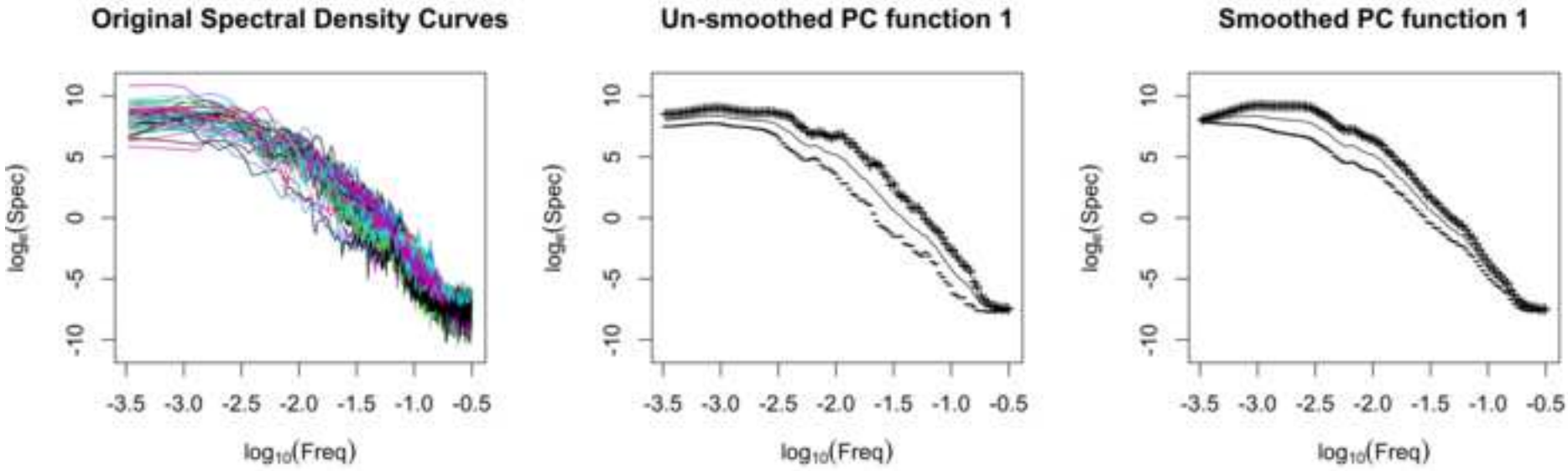
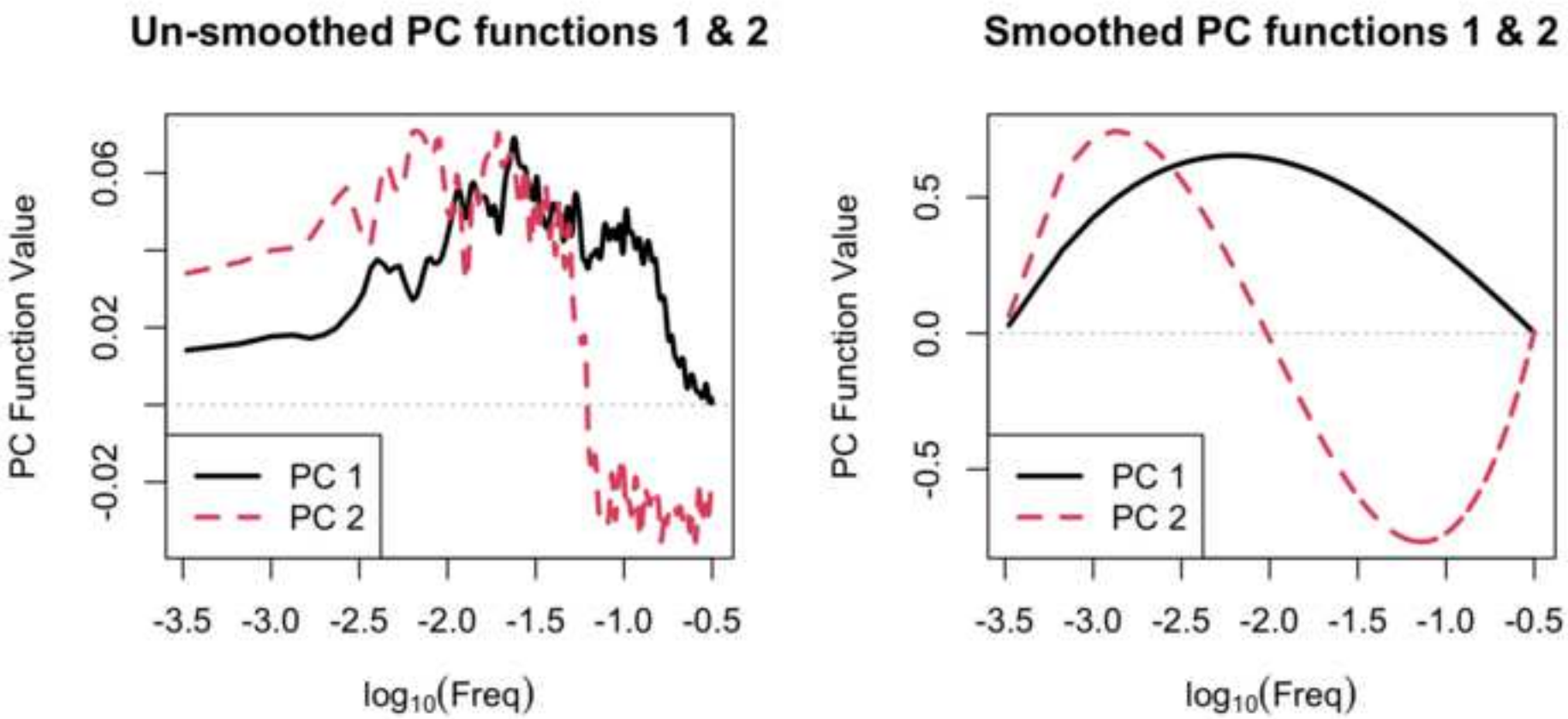

Original Spectral Density Curves
Un-smoothed PC function 1
Smoothed PC function 1
log10(Spec)
log10(Freq)
10
5
0
-5
-10
-3.5
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5

Figure 5


Un-smoothed PC functions 1 & 2
PC Function Value
0.06
0.02
-0.02
PC 1
PC 2
-3.5
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
log10(Freq)
Smoothed PC functions 1 & 2
PC Function Value
0.5
0.0
-0.5
PC 1
PC 2
-3.5
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
log10(Freq)

Table

Table 1. Summed the squared error (SSE) associated with reconstruction of unsmoothed and CV alpha selected smoothed *f*PCs. Bold indicates better error in each instance.

	1 *f*PC	**2 *f*PCs**	**3 *f*PCs**	**4 *f*PCs**	**5 *f*PCs**
No smoothing	0.219	**0.169**	0.159	0.155	0.129
CV selected alpha	**0.210**	0.170	**0.139**	**0.134**	**0.125**

Conflict of Interest Statement

The authors have no conflict of interest to declare.