First International Workshop on
Functional and Operatorial Statistics.
Toulouse, June 19-21, 2008

On the effect of curve alignment and functional PCA

Juhyun Park*

Department of Mathematics and Statistics, Lancaster University, Lancaster, U.K.
juhyun.park@lancaster.ac.uk

Abstract

When dealing with multiple curves as functional data, it is a common practice to apply functional PCA to summarise and characterise random variation in finite dimension. Often functional data however exhibits additional time variability that distorts the assumed common structure. This is recognized as the problem of curve registration. While the registration step is routinely employed, this is considered as a preprocessing step prior to any serious analysis. Consequently, the effect of alignment is mostly ignored in subsequent analyses and is not well understood. We revisit the issue by particularly focusing on the effect of time variability on the FPCA and illustrate the phenomena from a borrowed perturbation viewpoint. The analysis further suggests an iterative estimating procedure to optimise FPCA.

Introduction

Repeated measurements in the form of curves are increasingly common in various scientific applications including biomedicine and physical sciences (Ramsay and Silverman, 2002, 2005). Individual measurements are taken at consecutive time points (index set) and repeatedly observed for different subjects. Usually the sample of curves is assumed to have some homogeneous structure in the functional shape, while allowed for individual variability. It is desirable that the additional variability is summarised with a few components which are able to extract most variability and which are easy to interpret (Park et al. 2007).

Functional PCA utilises the well-known Karhunen-Loève expansion to provide an optimal representation of the function with a small number of common components. This is based on the assumption that the underlying random function shares the common mean

and covariance function. To fix the idea, consider a stochastic process $X \in L^2(\mathcal{T})$ with compact support $\mathcal{T} = [0, T]$, with the mean function $\mu(t)$ and the covariance function $\gamma(s,t) = Cov(X(s), X(t))$. Assume that $\int_{\mathcal{T}} E[X(t)^2] < \infty$. Let $\lambda_1 \geq \lambda_2 \geq \ldots$ be the ordered eigenvalues of the covariance operator defined through γ with the corresponding eigenfunctions $\phi_1, \phi_2, \ldots$. We assume that $\sum_k \lambda_k < \infty$. Then

$$X(t) = \mu(t) + \sum_k \xi_k \phi_k(t)\,, \tag{1}$$

where $E[\xi] = 0$ and $E[\xi_j \xi_k] = \lambda_j I(j = k)$.

With a sample of curves available, these quantities are replaced by their estimates and a finite number of components are usually considered sufficient to extract *significant* observed variation. Theoretical properties of estimators are studied in Dauxois et al. (1982), Rice and Silverman (1991), Kneip (1994) and Hall et al. (2006).

Often functional data exhibits additional time variability, which is mainly dealt with in pre-processing step, by aligning curves to eliminate the time variability prior to any serious analysis. This is known as registration problem and there are several methods developed. Basically when the functions exhibit identifiable features, curves can be aligned to match those features, which is known as landmark registration (Gasser and Kneip, 1995). This works well as long as features are correctly identified. Several other methods have been developed to automate the procedure when the features are less prominent. An overview can be found in Ramsay and Silverman (2005).

Although the issue has been rightly acknowledged, because most analysis treats registration as a preprocessing step, its carry-on effects on later analysis was not well studied. A recent work of Kneip and Ramsay (2007) address a similar problem and propose a new procedure to combine registration to fit functional PCA models, extending the covex averaging idea of registration (Liu and Müller, 2004).

Instead we focus on quantifying our misconduct. What happens then if registration was not carried out or was made improperly ? The obvious problem arises when estimating global mean structure. Generally, how does the time variability propagate through to functional PCA analysis ? Some issues with interpretability in functioanl PCA may also be attributed to the improper registration. We concentrate on relations of eigenvalues and eigenfunctions between unregistered and registered curves, in the sense that we do not want our registrations step to be *perfect* but we would like to be able to *correct* the residual difference from our imperfect analysis later.

Assume that the observed variable is $\tilde{X}(t) = X(\eta(t))$ for a monotone transformation $\eta(t)$ with $E[\eta(t)] = t$ for $t \in \mathcal{T}$. Suppose that we proceed to functional PCA without correcting η at the earlier stage to obtain $\tilde{\lambda}$ and $\tilde{\phi}$. How much do we lose by ignoring η ?

We may start with the representation in (1) as

$$\tilde{X}(t) = \mu(\eta(t)) + \sum_k \xi_k \phi_k(\eta(t))\,.$$

Now $E[\tilde{X}(t)] = \mu(\eta(t))$ but note that the series is not any longer orthonormal decomposition. Write $\tilde{\gamma}(s,t) = Cov(\tilde{X}(s), \tilde{X}(t))$. Then

$$\tilde{\gamma}(s,t) = \gamma(s,t) + \tilde{\gamma}(s,t) - \gamma(s,t)\,.$$

With some Taylor appromixation argument, it may be shown that $\tilde{\gamma}(s,t) - \gamma(s,t) = \varepsilon v(s,t)$ for some ε and v, then, under some regularity conditions and for small ε, we would have

$$\begin{aligned} \tilde{\lambda}_k &= \lambda_k + \varepsilon < \phi_k, V\phi > + O(\varepsilon^2)\,, \\ \tilde{\phi}_k &\propto \phi + \varepsilon \sum_{l \neq k} \frac{< \phi_k, V\phi_l >}{\lambda_k - \lambda_l} + O(\varepsilon^2)\,, \end{aligned}$$

where V denotes the corresponding operator for v. A similar derivation is made in Hall et al. (2006) to quantify sampling variability. We extend the idea to include time variability. Our interest is to recover λ and ϕ from $\tilde{\lambda}$ and $\tilde{\phi}$ using a sample of curves and a registration. Our estimators will be obtained from the estimators of unregistered curves with some correction made based on a registration. The precision of registration will be reflected on that of V and thus the correction terms in general. Based on these relations some properties of estimators will be studied and illustrated.

References

Dauxois, J. Pousse, A. and Romain, Y. (1982) Asymptotic theory for the principal component analysis of a vector random function : some applications to statistical inference. *Journal of Multivariate Analysis*, **12**, 136-154.

Liu, X. and Müller, H. G. (2004) Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association*, **99**, 687-699.

Gasser, T. and Kneip, A. (1995) Searching for structure in curve samples. *Journal of the American Statistical Association*, **90**, 1179-1188.

Hall, P, Müller, H. G. and Wang, J. L. (2006). Properties of principal componenet methods for functional and longitudinal data analysis. *Annals of Statistics*, **34**, 1493-1517.

Kneip, A. (1994) Nonparametric estimation of common regressors for similar curve data. *Annals of Statistics*, **22**, 1386-1427.

Kneip, A. and Ramsay, J. O. (2007) Combining registration and fitting for functional models. *technical report.*

Park, J. Gasser, T. and Rousson, V. (2007) Structural components in functional data. *technical report.*

Ramsay, J. O. and Silverman, B. W. (2002) *Applied functional data analysis*, New York : Springer.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional data analysis*, New York : Springer.

Rice, J. W. and Silverman, B. W. (1991) Estimating the mean and the covariance structure nonparametrically when the data are curves. *Journal of Royal Statistical Society, B*, **53**, 233-243.