

Principal components selection for estimating the functional linear model with scalar response ¹

Manuel Febrero-Bande

Dept. de Estadística e Investigación Operativa, Universidad de Santiago de Compostela,
e-mail: manuel.febrero@usc.es

Pedro Galeano ²

Dept. de Estadística, Universidad Carlos III de Madrid, e-mail: pedro.galeano@uc3m.es

Wenceslao González-Manteiga

Dept. de Estadística e Investigación Operativa, Universidad de Santiago de Compostela,
e-mail: wenceslao.gonzalez@usc.es

Abstract: In this work we analyze the selection of eigenfunctions of the functional predictor in the functional linear model with scalar response when the model is estimated through the principal components technique. Several properties on the selection of eigenfunctions to include in estimation are analyzed. In particular, it is shown that a wrong selection of the eigenfunctions included in estimation may lead, either to a biased estimation of the slope function of the model or to a estimation with large mean square error. Moreover, it is shown that the contribution to the fit of each eigenfunction is measured in terms of the correlation between the scalar response and the corresponding functional principal component score. However, it is shown that including eigenfunctions in estimation just because its autocorrelation is large with the scalar response may not be the more convenient approach. Indeed, a eigenfunction associated with a small eigenvalue should not be included in estimation unless its associated functional principal component score has a very large correlation with the scalar response. Finally, the performance of several strategies based on model selection criteria for selecting eigenfunctions are compared in terms of several aspects such as mean square error of estimation of the slope function of the model, mean square prediction error, error variance and coefficient of determination.

Keywords: Eigenfunctions; Eigenvalues; Functional linear model; Functional principal components; Model selection criteria.

1. Introduction

Functional data analysis deals with data consisting of repeated measurements of the same subject densely taken over an ordered grid of points belonging to a finite interval of the real line. The collection of this kind of data is becoming progressively frequent due to technological advances in measurement machines in fields such as chemometrics, meteorology, economy and image analysis, among others. As a consequence, there is a recent interest in methods dealing with functional data. The books of Ramsay and Silverman (2002, 2005) and Ferraty and Vieu (2006) summarizes several methods and case studies

¹The authors acknowledge financial support from MCI grant MTM2008-00166. The second author acknowledges financial support from Xunta de Galicia grant PGIDIT06PXIB207009PR.

²Address of correspondence: Dept. de Estadística, Universidad Carlos III de Madrid, c/Madrid, 126 28903 Getafe, Spain

for handling functional data from different approaches. Methods for functional data include, among others, functional principal components, functional linear models in which the predictors and/or the response are of a functional nature, functional analysis of variance and outlier detection for functional datasets.

Among all the available methods for handling functional data, the functional linear model in which the predictor is functional and the response is scalar have attracted considerable attention in the recent literature. Estimation of this model can be carried out by means of different approaches. For instance, estimation based on the use of restricted basis functions and penalization methods has been analyzed by Hastie and Mallows (1993), Marx and Eilers (1999), Cardot, Ferraty and Sarda (2003), Ramsay and Silverman (2005), Li and Hsing (2007) and Crambes, Kneip and Sarda (2009), among others. A different point of view based on functional kernel regression estimates defined through semi-metrics of functional spaces has been considered by Ferraty and Vieu (2006). On the other hand, Cardot, Ferraty and Sarda (1999) proposed an estimator based on regressing the response on the principal component scores of the functional predictor which has been posteriorly analyzed in several papers including Cardot, Ferraty and Sarda (2003), Hall and Hosseini-Nasab (2006), Cai and Hall (2006) and Hall and Horowitz (2007), among others. Hypothesis testing on this last estimator can be done using the test statistics proposed in Cardot, Ferraty, Mas and Sarda (2003), while diagnosis via residual processes was undertaken in Chiou and Müller (2007). Finally, Febrero-Bande, Galeano and González-Manteiga (2009) proposed influence measures for the functional linear model with scalar response estimated through this approach.

The estimation method based on the functional principal components technique has become very popular because, as mentioned by Hall and Hosseini-Nasab (2006) and Hall and Horowitz (2007), this method gives the most rapidly convergent representation of the functional dataset predictor when speed of convergence is defined in an L_2 sense. Estimation of the model proceed as follows. First, the functional principal components of the predictor variable are obtained. This is accomplished by extending the set of predictor curves in the basis of eigenfunctions of the covariance operator of the functional predictor, which is known as the Karhunen–Loève expansion, in order to obtain a small dimension space which represents as well as possible the main variations of the predictor. Then, the first principal component scores of the functional predictor are regressed against the scalar response in order to obtain an estimate of the functional slope of the model. Afterwards, one can estimate the residual variance of the process and obtain fitted values and residuals, which are useful for model assessment.

Cardot, Ferraty and Sarda (2003) pointed out that the estimator of the slope function of the model based on the functional principal components procedure may be sometimes too rough even for large sample sizes. This is because the inclusion in estimation of eigenfunctions associated with small eigenvalues may dramatically increase the mean square error of the estimator. As a consequence, several papers, as the mentioned above, have considered that are more reliable methods which includes penalization terms trying to avoid the inclusion of a large number of eigenfunctions. The first purpose of this work is to show that, indeed, a wrong selection of the eigenfunctions considered for estimation may lead, either to a biased estimation of the slope function of the model or to a estimation with large mean square error. In particular, it is shown that the contribution to the fit of each eigenfunction is measured in terms of the correlation between the scalar response and the corresponding functional principal component score. As a consequence, if a eigenfunction associated with a principal component score highly correlated with the

scalar response is not included in estimation, the estimate of the functional slope of the model may be strongly biased. On the other hand, if a eigenfunction associated with a principal component score with small correlation with the scalar response and with a small associated eigenvalue is included in estimation, the estimate of the functional slope of the model will have a large mean square error. Therefore, the selection of an adequate number of eigenfunctions in estimation is a crucial task. Indeed, an appropriate selection method joint with a convenient smoothing of the estimated slope function may avoid the two problems previously stated.

The results mentioned previously shows that the eigenfunctions selected accordingly with the magnitudes of their associated eigenvalues do not contribute monotonically to explain the total variance of the scalar response. This result suggests that selection of eigenfunctions linked with the functional principal scores with larger correlation with the scalar response may be more appropriate that selecting the eigenfunctions linked with the largest eigenvalues. However, the second purpose of this work is to show that including eigenfunctions in estimation just because its autocorrelation is large with the scalar response may not be the more convenient approach. This is because, the effect due to the estimation of the eigenfunctions and eigenvalues is not taken into account in the correlation analysis. Indeed, a eigenfunction associated with a small eigenvalue should not enter in estimation unless its associated functional principal component score be strongly correlated with the scalar response. Thus, in general situations, it may be more appropriate the conventional approach of including the eigenfunctions linked with the larger eigenfunctions and not to include eigenfunctions linked with the smaller eigenfunctions unless the correlation with the scalar response is significantly large. In fact, it is shown that the bias included in estimation of the functional slope, the error variance and the coefficient of determination of the model using the conventional approach is minimal.

Usually, the selection of the number of eigenfunctions to include in estimation is carried out by means of model selection criteria. For instance, Hall and Hosseini-Nasab (2006) used the predictive cross-validation criterion (PCV), Cardot, Ferraty and Sarda (2003) used the generalized cross-validation criterion (GCV), Chiou and Müller (2007) considered the Akaike Information Criterion (AIC), while Febrero-Bande, Galeano and González-Manteiga (2009) used the Corrected Akaike Information Criterion (AICc) and the Bayesian Information Criterion (BIC). The third purpose of this work is to compare the performance of these model selection criteria in selecting an adequate number of eigenfunctions to include in estimation in terms of several aspects such as mean square error of the estimation of the slope function of the model, mean square prediction error, error variance and coefficient of determination. In particular, it is shown that the BIC appears to have the better performance for selecting the number of eigenfunctions, specially for small sample sizes.

The structure of the talk will be as follows. First, we present the functional linear model with scalar response and analyzes estimation through the principal components methodology. Then, we analyze several aspects regarding the selection of eigenfunctions to include in estimation. Afterwards, we present several model selection criteria usually considered for estimating the functional linear model with scalar response. Finally, we carry out a comparison of several strategies for selecting eigenfunctions through different measures by means of different Monte-Carlo experiments, which considers different models, sample sizes and error distributions.

References

- Cai T.T. and Hall P. (2006) Prediction in functional linear regression, *Annals of Statistics*, 34, 2159–2179.
- Cardot H., Ferraty F., Mas A. and Sarda P. (2003a) Testing hypothesis in the functional linear model, *Scandinavian Journal of Statistics*, 30, 241–255.
- Cardot H., Ferraty F. and Sarda P. (1999) Functional linear model, *Statistics and Probability Letters*, 45, 11–22.
- Cardot H., Ferraty F. and Sarda P. (2003b) Spline estimators for the functional linear model, *Statistica Sinica*, 13, 571–591.
- Chiou J.M. and Müller H.G. (2007) Diagnostics for functional regression via residual processes, *Computational Statistics and Data Analysis*, 51, 4849–4863.
- Crambes C., Kneip A. and Sarda P. (2009) Smoothing splines estimators for functional linear regression, *Annals of Statistics*, 37, 35–72.
- Febrero-Bande M., Galeano P. and González-Manteiga W. (2009) Measures of influence in the functional linear model with scalar response, *Journal of Multivariate Analysis*.
- Ferraty F. and Vieu P. (2006) *Nonparametric functional data analysis*, Springer-Verlag, New York.
- Hall P. and Horowitz J.L. (2007) Methodology and convergence rates for functional linear regression, *Annals of Statistics*, 35, 70–91.
- Hall P. and Hosseini-Nasab M. (2006) On properties of functional principal components analysis, *Journal of the Royal Statistical Society, Series B*, 68, 109–126.
- Hastie T. and Mallows C. (1993) A discussion of “a statistical view of some chemometrics regression tools” by i. e. frank and j. h. friedman, *Technometrics*, 35, 140–143.
- Li Y. and Hsing T. (2007) On rates of convergence in functional linear regression, *Journal of Multivariate Analysis*, 98, 1782–1804.
- Marx B.D. and Eilers P.H. (1999) Generalized linear regression on sampled signals and curves: a p-spline approach, *Technometrics*, 41, 1–13.
- Ramsay J.O. and Silverman B.W. (2002) *Applied Functional Data Analysis*, Springer-Verlag, New York.
- Ramsay J.O. and Silverman B.W. (2005) *Functional data analysis, 2nd edition*, Springer-Verlag, New York.
- Yao F., Müller H.G. and Wang J.L. (2005) Functional linear regression analysis for longitudinal data, *Annals of Statistics*, 33, 2873–2903.