Functional Projection Pursuit

G. P. Nason
Department of Mathematics
University of Bristol
Bristol, BS8 1TW, England

Abstract

This article describes the adaption of exploratory projection pursuit for use with functional data. The aim is to find "interesting" projections of functional data: e.g. to separate curves into meaningful clusters. Functional data are projected onto low-dimensional subspaces determined by a projection function using a suitable inner product. Such a projection is rapidly computed by representing data and projection function in terms of a suitable orthogonal basis. Both Fourier and wavelet bases are considered and their advantages and disadvantages outlined. The concepts of interpretable projection solutions, centring and sphering are also discussed. Two examples are presented: one simulated and one real concerning the core temperature of a group of infants that sleep with and without their mothers. The aim for the real data is to find interesting projections that may separate the two groups.

1 Introduction

Functional projection pursuit (FPP) is about the exploratory investigation of functional data using projection pursuit. FPP marries functional data analysis (FDA) to exploratory projection pursuit (EPP).

FDA is concerned with the analysis and interpretation of data that arise as curves. Let $x_i(t)$, for $i = 1, \dots, n$ be our sample of n curves (assumed to be on some space $\mathcal{S}$, e.g. $L^2[0,1]$. We shall assume the functions are defined on some compact interval $[0,T]$). In some ways FDA is similar to multivariate analysis except that in FDA each observation is a curve not a vector although sometimes it is helpful to think of a curve as a vector of infinite dimension.

Figure 1 shows a picture of some functional data (colour versions of these figures can be found on the author's Web site: `www.stats.bris.ac.uk`). Each curve corresponds to a recording of infant core temperature from when the infant first went to sleep over a period of about 10 hours. Each infant underwent two recording sessions on different nights: once when sleeping alone (solitary sleep group) and once when sleeping with its mother (co-sleep group). Two questions are asked of this data:

1. Are there any differences between the two groups?
2. If there are differences, what are they?

Here we assume each observation is a curve: of temperature against time. This is a valid and important assumption because the core temperature exists at all time points through the night. It is only our present data recording technology that forces us to sample the curve in nearly all practical cases. In this infant example the temperature is recorded every 30 seconds throughout the night. Note also that the temperature curves are not smooth but exhibit a distinct staircase-like appearance due to the temperature sensor having a maximum resolution of 0.05°C. For further information and a recent survey of FDA see Ramsay and Silverman (1997). Many of the usual multivariate techniques such as principal components analysis, canonical correlations and linear model theory can be transferred to the functional context (see Rice and Silverman (1991), Ramsay (1996) and Leurgens, Moyeed and Silverman (1993)).

Exploratory projection pursuit is another multivariate technique for the exploratory analysis of high-dimensional data. Given a $K \times N$ data matrix X (assume N observations on K variables) EPP looks for interesting low-dimensional projections of high-dimensional data. One-dimensional projections of X may be obtained by choosing a K-dimensional projection vector a and forming the

$$Z = a^T X. \tag{1}$$

Usually the constraint $||a|| = 1$ is imposed as it is the direction of projection that matters, not the overall scale. Two- and higher dimensional projections may easily be obtained by choosing other projection vectors and it is

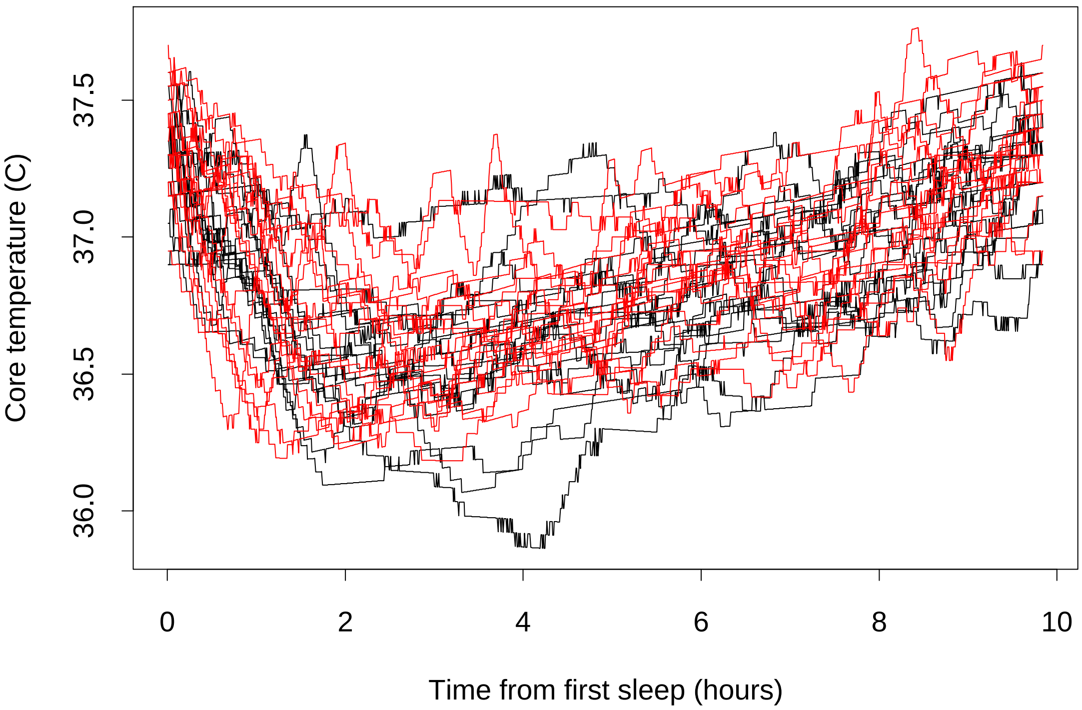

Figure 1: Core temperature versus time since infant first went to sleep for a group of 18 infants. There are two traces for each infant: once when the infant slept alone (black lines) and once when the infant slept with its mother (red lines).

usually the case that the projection vectors are chosen to be mutually orthogonal. To search for interesting views of the data each projection is assigned an "index of interestingness". This is done by defining a *projection index*, I, in terms of the projected data Z and hence the projection direction a. For example, we shall use the negative Shannon entropy index

$$I(Z(a)) = I(a) = -\int f_Z(z) \log f_Z(z)\, dz,$$

where $f_Z(z)$ is the density of the projected data Z. There are many other possible choices of projection index. The negative Shannon entropy measures the non-normality of the density of Z. For all *sphered* densities I is maximised by the standard normal density (a sphered density is one that has zero mean and unit variance). Thus I is a measure of non-normality and "small" values correspond to non-normal projections of the data. The objective of EPP is to numerically minimize estimates of I over all possible projections. Non-normal projections would suggest clustering or other interesting structure in high-dimensions. For our infant temperature data it is possible that EPP may find a projection that splits the two groups into solitary and co-sleepers. Good explanations of EPP may be found in Huber (1985), Jones and Sibson (1987) and Friedman (1987).

2 Projecting data

Adaption of EPP to functional data requires alteration of the projection step. Once we have projected data we can more or less proceed as for EPP. For FPP we project by means of an inner product $< \cdot, \cdot >$ defined on $\mathcal{S} \times \mathcal{S}$. In other words to obtain projected data z_i we form

$$z_i = < a, x_i > \tag{2}$$

for some projection *function* $a(t) \in \mathcal{S}$ and where $x_i(t)$ is the ith curve for $i = 1, \ldots, n$.

There are many possibilities for the type of inner product. The remainder of this article uses the "usual" integration inner product

$$< f, g > = \int_0^T f(t)\overline{g(t)}\, dt. \tag{3}$$

Other possibilities include the case where certain locations $t_1, \ldots, t_M \in [0, T]$ are known *a priori* to be of interest. Then we might use

$$< f, g > = \sum_{m=1}^{M} f(t_m)\overline{g(t_m)}.$$

This idea generalises to a strictly positive weight function $w(t) \in \mathcal{S}$ which puts more weight on some areas of $[0, T]$ than others and leads us to the following inner product:

$$< f, g >= \int w(t) f(t) \overline{g(t)}\, dt.$$

2.1 Orthogonal basis representation

Direct computation of projected data according to the integration inner product (3) is possible — say via numerical integration methods. However, computation in this manner is extremely computationally intensive. Further, index optimization is more effective if the derivatives of the index are available. For this we also require derivatives of the projected data with respect to the projection *function*. In principal, it might be possible to perform this differentiation using some form of numerical calculus of variations but the author has no idea of exactly how this might be carried out.

Fortunately, there is a much more elegant solution for the efficient computation of the projection which also permits easy differentiation. First choose a suitable orthogonal basis for $\mathcal{S}$: $\{\psi_\nu(t)\}_\nu$ say. Then write both the projection function and data in terms of this basis:

$$a(t) = \sum_\nu a_\nu \psi_\nu(t)$$

and

$$x_i(t) = \sum_\nu x_{i,\nu} \psi_\nu(t)$$

where $\{a_\nu\}_\nu$ and $\{x_{i,\nu}\}_\nu$ are the expansion coefficients of $a(t)$ and $x_i(t)$ respectively for $i = 1, \ldots, n$. This paper assumes that the transforms have n coefficients also. Plancheral's formula permits expression of the projected data by

$$z_i = \int a(t) x_i(t)\, dt = \sum_\nu a_\nu x_{i,\nu}$$

Now define a matrix $\hat{X}$ whose ith column consists of $\{x_{i,\nu}\}_\nu$ and write $\hat{a}$ for the vector of coefficients of $a(t)$. The projected data can be written as

$$z = \hat{a}^T \hat{X}.$$

In other words the projected data can be written as a matrix multiplication as in (1) even for functional data if the computation is performed with the orthogonal transform coefficients.

Use of orthogonal bases raises the interesting question of which basis to use — there is plenty of choice. Two possible choices discussed here are:

Fourier The Fourier basis is a natural choice for representing the projection functions. Additionally it is easy to incorporate smoothing as the projection function tends to the optimum by simply suppressing high frequencies. Computationally the transform (and inverse) algorithm is fast ($\mathcal{O}(n \log n)$) which is important not only for the inital computation of $\hat{a}$ and $\hat{X}$ but is also useful when one wishes to view representations of the projection whilst pursuing the optimum index value.

Wavelet Wavelet transforms and their inverses are extremely fast to compute ($\mathcal{O}(n)$). Moreover wavelets have good compression abilities and have the ability to represent many classes of function very sparsely. Wavelets should enhance the dimension reduction capabilities of projection pursuit because they should concentrate signals into fewer coefficients to begin with. It is also possible to produce smooth wavelet-based projection functions by progressively downweighting the higher resolution wavelets.

2.2 Index differentiation

The derivatives of a projection index aid its optimization. With the orthogonal basis representation we can easily find the derivatives of the index with respect to the basis coefficients $\hat{a}$ of the projection function. However, this method of differentiation deserves some comment. It is highly basis specific: the path the optimizer will take depends on the set of basis functions (and the initial projection function). However, the same thing happens in the multivariate case a specific basis is used: it just happens to be the one that the data were collected. A change of basis (rotation) is not usually performed but if it was it would only change the path of the optimizer (unless annoying local optima got in the way).

Here we shall estimate the density of the projected data by a kernel density estimate, $\tilde{f}_{\hat{a}}(z)$, with Gaussian kernel, ϕ, and bandwidth denoted by h. The derivatives of the negative Shannon index

$$I(\hat{a}) = \int \tilde{f}_{\hat{a}}(z) \log \tilde{f}_{\hat{a}}(z)\, dz$$

are given by

$$\frac{\partial I(\hat{a})}{\partial \hat{a}_k} = -\int \left\{1 + \log \tilde{f}_{\hat{a}}(z)\right\} \frac{\partial \tilde{f}_{\hat{a}}(z)}{\partial \hat{a}_k}\, dz,$$

where

$$\frac{\partial \tilde{f}_{\hat{a}}(z)}{\partial \hat{a}_k} = n^{-1} h^{-3} \sum_{i=1}^{n} \hat{X}_{ki}(z - Z_i) \phi\left(\frac{z - Z_i}{h}\right).$$

2.3 Computation, optimization and smoothing

The above index also requires a bandwidth h to be fully specified. In our algorithm we choose our bandwidth to be a fixed fraction of the 'rule-of-thumb' (but decidedly sub-optimal) bandwidth (see e.g. Silverman (1986))

$$h = 1.06\sigma n^{-1/5}.$$

Usually the algorithm uses 40% of this. This usually produces a rougher density estimate which means the pursuit algorithm falls into more local optima but these can be swiftly rejected.

The optimiser used is the Splus function `nlminb`. The starting projection function is one factor which determines how quickly we reach an optimum solution. At present we use random starts but we intend to use other possibilities in future: e.g. principal component starting solutions.

Whilst optimising the constraint $||\hat{a}|| = 1$ is satisified by a continual renormalisation of the $\hat{a}$ coefficients as in Jones and Sibson (1987). Likewise it is common for projection functions to be extremely non-smooth which prevents easy interpretation (non-smooth solutions also result from certain random starts). To ensure smoothness of our projection solutions we continually modify $\hat{a}$ by suppressing the high-frequencies by multiplying the Fourier coefficients by the weight function:

$$w(\omega) = \exp(-\lambda|\omega|)$$

where ω is frequency and λ is constant which is chosen manually at present. This method of enforcing smoothness is crude but simple and effective. Obviously smoothness can be more elegantly enforced by means of a constraint and an "ideal" solution would probably take the form of a penalty method which incorporates not only a smoothness penalty but the normalisation constraint as well.

Summary

To summarise the whole algorithm is as follows:

1. Form orthogonal basis coefficients of the data $\hat{X}$.
2. Choose initial projection function: $a(t)$.
3. Form orthogonal basis coefficients a_ν.
4. Compute projection index and derivatives.
5. Numerically optimize index I.
6. If optimal goto 4, else stop.

It is wise to repeat the algorithm using many different starting positions.

3 Simulated example

The 30 simulated data curves in Figure 2 were used to test out the method. Here curves were initially generated using a few Fourier basis functions and different sets were used for the two groups (one black, one red). In this case the curves are sampled at 40 equally spaced points. The optimal projection solution obtained from FPP is also shown on Figure 2 as a single dashed line (green on the colour version on the Web site). When the inner product is taken with each curve the projected data falls neatly into two completely separated groups — thus in this very obvious and simple example FPP is working as it should. In fact, the projection function is highly interpretable as well. Due to the simplicity of the example it is easy to see that the projection solution is large on parts of the interval where the two sets of curves differ (e.g. at about 6, 12 and 18 on the horizontal axis).

4 PP techniques

The world of EPP has promoted a set of interesting techniques: centring and sphering (see e.g. Tukey and Tukey (1981)) and the making of more interpretable projection solutions.

Interpretable PP

The aim of interpretable projection pursuit in the multivariate case is to make the optimal projection solution more interpretable by "adjusting" the projection vector so that more weight is put onto fewer original variables. Thus the optimal projection becomes easier to understand in terms of the original variables. Morton (1990) suggests two techniques for doing this:

1. Rotation of the projection coordinate axes within the plane of projection. If a rotationally invariant projection index is used then this does not change the projection index value but can often improve interpretability. The rotation can be achieved through using varimax techniques.

2. Rocking of the projection plane. Often it is possible to rock the plane of projection without changing the projected data view too much. Again, the goal is to rock the plane and improve interpretability.

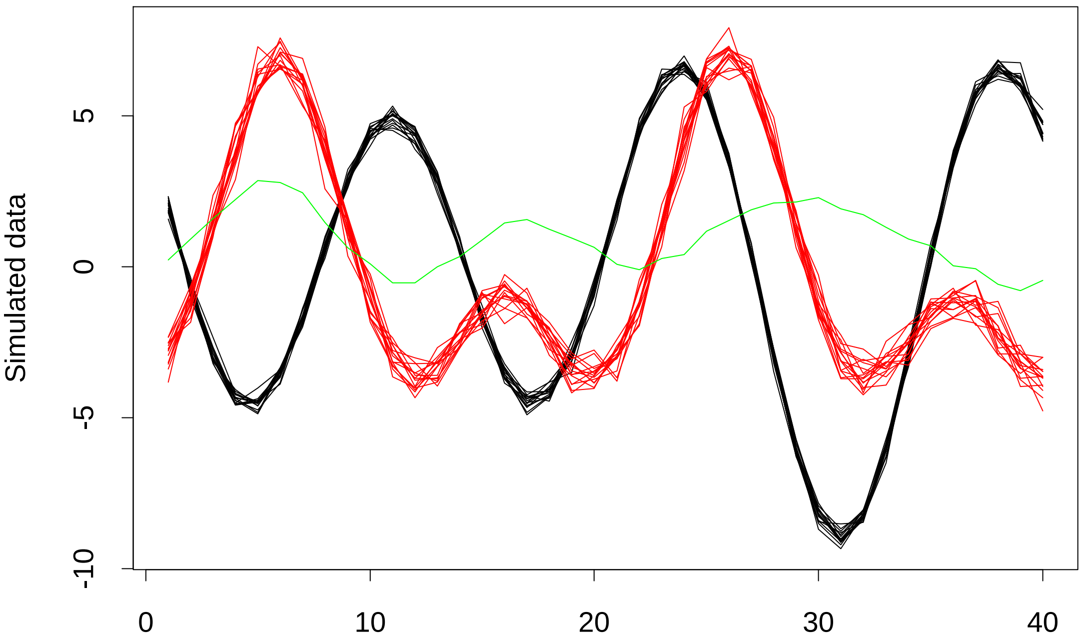

Figure 2: Simulated data. There are two sets of curves here indicated by the black and red colours. The optimal projection function is shown as a green line.

Both of these could, in principle, be adapted for use in the FPP situation. Interpretability in our setting would be with respect to the orthogonal basis functions rather than the overall projection function — this may or may not be useful. Indeed the desirability of finding sparse projection solutions is likely to be very problem specific: clusters "in-time" would probably be made more interpretable if a wavelet basis was used, clusters "in-frequency" would benefit from a Fourier basis.

Centring and sphering

In multivariate analysis centring is the act of transforming a random variable to have zero mean. Sphering is the act of transforming a random variable so that it has an identity covariance matrix. Sphering is an especially drastic transformation. The purpose of carrying out both of these transforms is so that EPP analysis of centred sphered data gives "independent" information from that produced by variance based methods such as principal components analysis (since the data has an identity covariance matrix the principal components "do not exist").

What happens in the functional case? Much the same thing and the transformation is still drastic. To centre a set of curves $\{x_i(t), i = 1, \ldots, n\}$ first compute the (pointwise) mean:

$$\overline{x}(t) = n^{-1} \sum_{i=1}^{n} x_i(t).$$

Then the centred curves are given by

$$y_i(t) = x_i(t) - \overline{x}(t)$$

for $i = 1, \ldots, n$. The covariance function of the centred $Y(t)$ is

$$\Gamma(s,t) = \mathrm{cov}\,[Y(s), Y(t)] = \mathbb{E}[Y(s)Y(t)]$$

for $s, t \in [0, T]$. The sample version of Γ can be estimated by

$$(n-1)^{-1} \sum_{i=1}^{n} y_i(s) y_i(t).$$

As with the multivariate case one way to sphere involves writing Γ in terms of its eigendecomposition: i.e. write

$$\Gamma(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu e_\nu(s) e_\nu(t)$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ and the e_ν are the orthogonal eigenfunctions of Γ (see e.g. Silverman (1996)). Then form the inverse square-root of Γ by

$$\Gamma^{-1/2}(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu^{-1/2} e_\nu(s) e_\nu(t)$$

The sphered curves $W(t)$ are formed by

$$W(t) = \int \Gamma^{-1/2}(s,t) Y(s)\, ds.$$

To show that the W are indeed sphered we need to show that Γ_W is the identity. To do this observe that

$$\Gamma_W(s,t) = \mathrm{cov}\,[W(s), W(t)] = \sum_{\nu=1}^{\infty} e_\nu(s) e_\nu(t).$$

It may not be immediately obvious to the reader that this is the identity. We show it is in the following operational sense. Let $f \in S$ be some function which we can represent in terms of the eigenfunctions $\{e_\nu\}_\nu$ with coefficients f_ν as

$$f(t) = \sum_{\nu=1}^{\infty} f_\nu e_\nu(t)$$

Then application of the covariance matrix of W yields:

$$\begin{aligned}
\int \Gamma_W(s,t) f(t)\, dt &= \int \sum_{\nu=1}^{\infty} e_\nu(s) e_\nu(t) \sum_{\mu=1}^{\infty} f_\mu e_\mu(t)\, dt \\
&= \sum_{\nu=1}^{\infty} e_\nu(s) \sum_{\mu=1}^{\infty} f_\mu \int e_\nu(t) e_\mu(t)\, dt \\
&= \sum_{\nu=1}^{\infty} e_\nu(s) \sum_{\mu=1}^{\infty} f_\mu \delta_{\nu\mu} \\
&= \sum_{\nu=1}^{\infty} f_\nu e_\nu(s) \\
&= f(s)
\end{aligned}$$

In other words $\mathcal{I} : \mathcal{S} \to \mathcal{S}$ defined by

$$\{\mathcal{I}(f)\}(s) = \int \Gamma_W(s,t) f(t)\, dt$$

is the identity operator.

5 Infant temperature

With our infant temperature example we have to report that, at the present time, we have not found any projections that separate the co-sleep/solitary sleep groups neatly. Higher dimensional projections obtained by combining one-dimensional ones "do better" but still no simple rule works. It could be that there is no projection function that will do the job or it could be that out method is not finding the best projection and getting stuck in local optima — we are investigating further. With this data there are other covariates that might also be interesting such as the sex of the baby. The other point to note is that this data is (effectively) of extremely high dimension thus making the discovery of good projection solutions difficult.

Note that the FPP algorithm *does* find other interesting multimodal projections, it is just that none of them coincide with the co-sleep/solitary sleep division.

6 Conclusions and Further Work

This article describes work in progress on a exploratory projection pursuit algorithm for functional data. It clearly works on simple simulated examples and produces "interesting" solutions with some complex data although without producing the "desired" group separation. For the data set it is likely that a functional discrimination method would also be useful in attempting to find out what projection functions can separate the two groups if any. The concepts of centring and sphering have been extended to the functional case and further experience needs to be gained in applying such techniques to real data. We are also in the process of using wavelet bases which should equip the FPP method with extra dimension reducing power.

Acknowledgements

I am grateful to Andrew Sawczenko of the Royal Hospital for Sick Children, Bristol for supplying the infant temperature data. I would like to thank Jim Ramsay and Bernard Silverman for letting me see advance copies of their forthcoming book. Thanks also to Bernard Silverman for supplying functions to compute the forward and reverse sine/cosine transforms. Finally I'd like also to thank Nancy Heckman and Trevor Hastie for interesting discussions and ideas. The author was supported in part by the U.K. EPSRC grant KR70236. He is also extremely grateful to David Scott and INTERFACE for financial support to attend.

References

Friedman, J.H. 1987. Exploratory projection pursuit. *J. Am. Statist. Ass.*, **82**, 249–266.

Huber, P.J. 1985. Projection pursuit (with discussion). *Ann. Statist.*, **13**, 435–525.

Jones, M.C., & Sibson, R. 1987. What is projection pursuit? (with discussion). *J. Roy. Statist. Soc. Series A*, **150**, 1–36.

Leurgans, S.E., Moyeed, R.A., & Silverman, B.W. 1993. Canonical correlation analysis when the data are curves. *J. Roy. Statist. Soc. Series B*, **55**, 725–740.

Morton, S.C. 1990. Interpretable exploratory projection pursuit. *Pages 470–474 of:* Page, C., & LePage, R. (eds), *Proc. Symp. Interface*. Berlin: Springer-Verlag.

Ramsay, J.O. 1996. Principal differential analysis: data reduction by differential operators. *J. Roy. Statist. Soc. Series B*, **58**, 495–508.

Ramsay, J.O., & Silverman, B.W. 1997. *Functional Data Analysis*. New York: Springer.

Rice, J.A., & Silverman, B.W. 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Series B*, **53**, 233–243.

Silverman, B.W. 1986. *Density estimation*. London: Chapman and Hall.

Silverman, B.W. 1996. Smoothed functional principal components — analysis by choice of norm. *Ann. Statist.*, **24**, 1–24.

Tukey, P.A., & Tukey, J.W. 1981. Preparation: prechosen sequences of views. *Pages 189–213 of:* Barnett, V. (ed), *Interpreting multivariate data*. Chichester: Wiley.