Principal Subspace Analysis and Some Extensions Using Parameter Cascading

J. O. Ramsay and A. Kneip

November 15, 2014

Abstract

Key Words: marginalization, regularization, nuisance parameters

1 Introduction

This paper investigates the generalization of principal components analysis (PCA) or, more descriptively, *principal subspace analysis*. A parameter estimation strategy call parameter cascading permits the computation the usual principal components of variation in either a multivariate or functional data context without using eigenanalysis; and, as a consequence, enables a wide range of useful extensions of the classical method. Among these is an improved version of partial least squares for predicting one or more dependent variables from a reduced rank approximation of the independent variable values, and the registration of functional data to its principal subspace of variation.

For an N by n multivariate data matrix $\mathbf{X}$, PCA is its approximation of rank K by the bilinear model $\mathbf{FA}$, where the N by K matrix $\mathbf{F}$ contains in its rows principal component scores for defining the approximation of the data for each case, and the K by n matrix $\mathbf{A}$ contains within its columns the principal component coefficients for approximating the variables. The classical estimation of these matrix parameters can be achieved in one step via the singular value decomposition of $\mathbf{X}$, but computational considerations favour the two-stage process of computing $\mathbf{A}$ through the eigenanalysis of $N^{-1}\mathbf{X}'\mathbf{X} = \mathbf{VDV}'$, $\mathbf{A}'$ being the leading K columns of $\mathbf{VD}^{1/2}$, and $\mathbf{F}$ defined as

$$\mathbf{F}(\mathbf{A}) = \mathbf{XA}'(\mathbf{AA}')^{-1}. \tag{1}$$

More generally, since it seems to us appropriate to reserve the use of "principal component" for the classical estimation of $\mathbf{F}$ and $\mathbf{A}$, we refer to these matrix parameters as *factor scores* and *factor loadings*, respectively. Elsewhere in the mathematical sciences, a plane embedded in higher dimensional Euclidean space is known as a Grassmann manifold.

Why would one want to amend such a well-established and clearly successful statistical method as PCA? Here are a few of its more serious deficiencies:

- PCA is tightly tied to least squares estimation. Either the eigenanalysis of $N^{-1}\mathbf{X}'\mathbf{X}$ or the singular value decomposition of $\mathbf{X}$ itself yields a least squares approximation of rank K to the respective matrix by eliminating all but the dominant K eigen- or singular-values, respectively. However, least squares approximation is often not appropriate; variables like the concentration of a chemical vary between pre-determined limits, other variables are intrinsically nonnegative, discrete or even binary, and least squares estimation is problematical in the presence of strongly skewed and long-tailed distributions of residuals in $\mathbf{X} - \mathbf{FA}$. Consequently the capacity to choose the fitting criterion separately for each variable is likely to improve the quality of what PCA estimates.

- The classical PCA problem is ill-posed. If order K matrix $\mathbf{W}$ is nonsingular, than PC-scores matrix $\mathbf{G} = \mathbf{FW}$ and PC-coefficients matrix $\mathbf{B} = \mathbf{W}^{-1}\mathbf{A}$ yield the same approximation to the data. It is well-known in the social science statistical literature that rotations of eigenvectors, where $\mathbf{W}$ is assumed to be orthonormal, can greatly improve the interpretability of the basis for the subspace. It is easy to see why many well-known texts on multivariate statistics fail to mention this issue; if we assume that the eigenvalues are distinct, then the eigenvector-eigenvalue pairs are uniquely defined. But as a basis system eigenvectors are merely byproducts of subspace identification, and in no sense are essential to the fundamental PCA problem, which is to define the best-fitting vector subspace of dimension K. Put another way, eigenvectors over-parameterize the subspace, which, as a Grassmann manifold, is known to have dimension $K(n-K)$ (Boothby, 1975), as opposed to the rank nK of t$\mathbf{V}$. The PCA problem therefore intrinsically leaves open the question of what basis is used to span the subspace. Put more succintly, it is the subspace that is the parameter in the PCA model, rather than its particular representation by eigenanalysis.

- Principal component scores and coefficients are treated symmetrically by the singular value decomposition, since it provides essentially the same solution for both $\mathbf{X}$ and $\mathbf{X}'$. But the dimensions of $\mathbf{X}$ differ in an important way: the n sets of PC-coefficients in $\mathbf{A}$ control the fit to the entire corpus of data, but each of the N sets of PC-scores, $f_{ik}, k = 1, \ldots, K$ controls the fit only to the relative small amount of data in the ith row of $\mathbf{X}$. Moreover, N can easily vary greatly from one set of data to another, whereas n is more apt to be considered fixed. As a consequence, the PC-coefficients are *structural* parameters and the PC-scores are *incidental* or *nuisance* parameters in the terminology of Neyman and Scott (1948), who first showed that the usual estimation methods for structural parameters in the presence of nuisance parameters are neither consistent nor efficient. The large literature on the nuisance parameter problem has made clear that diffeerent estimation strategies are required for these two parameter classes.

- The principle of *regularization* as a means of stabilizing estimates, exchanging squared bias for a reduction in sampling variance, and of incorporating some prior information about what is being estimated is now well established for most popular data analyses, and this also seems desirable for principal components analyses. Moreover, the notion of regularity or smoothness seems likely to differ for factor loadings and factor scores, further emphasizing the need for an asymmetric treatment of the two parameter classes.

- It is often essential to fit a model to data by a technique that is robust to outliers. Indeed, the PCA problem present robustness problems at more levels than that of the residual matrix $\mathbf{X} - \mathbf{FA}$; it is often the case in practice that a principal component emerges early in the sequence that essentially fits a single variable in the multivariate case, or a feature in a single curve in the functional context. There is already a considerable literature on robust covariance estimation where pure eigenanalysis has to be abandoned to achieve this end.

- Interval or confidence region estimation for the eigenvectors requires the assumption of separated eigenvalues, a property that seems artificial in the context of the processes that generate most data. This highlights the fact that interval estimation for basis systems is the wrong question; what we need is a measure of the sampling variation of the estimated subspace around some "population" manifold in frequentist terms, or the posterior probabilities of K-dimensional subspaces given $\mathbf{X}$ within a Bayesian framework.

Parameter cascading is a strategy for estimating structural parameters $\mathbf{A}$ in the presence of nuisance parameters $\mathbf{F}$ by defining the latter as a smooth function $\mathbf{F}(\mathbf{A})$ of the former, with user-control over "smoothness". This functional relationship can often be defined explicitly, but also more generally by specifying an inner data-fitting criterion $G(\mathbf{F}|\mathbf{A})$ that is re-optimized each time there is a change in $\mathbf{A}$ during the process of optimizing an outer criterion $H(\mathbf{A})$. When H and G are the same objective functions, as is the case for the least squares criterion in classical PCA, the use of this functional relationship is the well-known profiling algorithm. But we argue that there are important reasons for H and G to not to be the same, so that the resulting multi-criterion optimizing parameter cascading strategy is a generalization of profiling.

Parameter cascading was first used by Ramsay, Hooker, Campbell and Cao (2007) for the estimation of parameters defining dynamical systems, a context where an explicit expression of the model is usually not possible. Cao and Ramsay (2007) and Cao and Ramsay (2008) applied parameter cascading to data smoothing with adaptive regularization in order to adjust confidence intervals to take into account the uncertainty in data-determined smoothing parameters. Cao and Ramsay (2010) used parameter cascading for high-dimensional linear mixed effects models, including those for functional data.

Parameter cascading is already latent in a wide range of older methodologies, including the use of profiling in nonlinear least squares problems discussed in Bates and Watts (1988) and many other texts, and also in the large literature on the nuisance parameter estimation problem of Neyman and Scott (1948). Parameter cascading is also used in regularized data smoothing and functional parameter estimation in functional data analysis (Ramsay and Silverman, 2005), the REML method for estimating inter-record variance components in random effects models, and some versions of the EM algorithm.

The rest of this paper is organized as follows. Section 2 applies parameter cascading to the estimation of the classical principal components of $\mathbf{S}$, as well as of a set of N observed functions $x_i(t)$ observed over a common domain for continuous index t.

2 Parameter cascading for Multivariate PCA

We now consider the application of parameter cascading to a variety of types of principal components analyses, including both multivariate and functional versions, partial least squares, and the registration of functional observations to a principal components representation. For the most part, the path from the multivariate case to other situations is easy to follow, so that we consider parameter cascading for multivariate PCA in some depth.

2.1 Defining the parameter cascaded PCA for multivariate data

We define the factor scores in $\mathbf{F}$ as smooth functions of the factor coefficients in $\mathbf{A}$; that is, as $\mathbf{F}(\mathbf{A})$. In effect, this is already achieved in the classic case by the regression equation (1), but we now wish to add regularization features to the estimation of both $\mathbf{F}$ and $\mathbf{A}$, as well as the variation of fitting criteria over variables. Let g_j be a fitting criterion specific to variable j, which will often be a negative log likelihood tailored to that variable's known distributional characteristics. Let P_F be regularization function that quantifies one or more aspects of "roughness" in the estimate of $\mathbf{F}$. The inner fitting criterion will usually, but not inevitably, be

$$G(\mathbf{F}|\mathbf{A}, \mathbf{X}) = \sum_{j}^{n} g_j(\mathbf{x}_j, \mathbf{F}\mathbf{a}_j) + P_F(\mathbf{F}), \tag{2}$$

where $\mathbf{x}_j$ and $\mathbf{a}_j$ the jth column of $\mathbf{X}$ and $\mathbf{A}$, respectively.

Among the obvious choices for the roughness term P_F are the quadratic forms $\lambda_1 \text{trace}(\mathbf{F}\mathbf{R}_1\mathbf{F}')$, $\lambda_2 \text{trace}(\mathbf{F}'\mathbf{R}_2\mathbf{F})$ and $\lambda_{12} \text{vec}(\mathbf{F})'\mathbf{R}_{12}\text{vec}(\mathbf{F})$ that capture roughness in the rows, columns, and vectorized $\mathbf{F}$, respectively. The roughness penalty matrices $\mathbf{R}_1 \in \mathcal{S}^K$, $\mathbf{R}_2 \in \mathcal{S}^N$ and $\mathbf{R}_{12} \in \mathcal{S}^{NK}$, where $\mathcal{S}^m$ is the space of symmetric semidefinite matrices of order m, will often be projectors into the complement of some low-dimensional subspace spanned by the known columns of a pre-specified matrix $\mathbf{Z}$. For example, $\mathbf{R}_2 = \mathbf{I}_N - \mathbf{Z}_2(\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\mathbf{Z}_2'$ measures the extent to which columns of $\mathbf{F}$ lie in the subspace spanned by the columns of $\mathbf{Z}_2$ having N rows. Row-regularization using $\mathbf{R}_1 = \mathbf{I}_K$ can reign in the effects of outlying observations on the subspace identification. The regularization of vec($\mathbf{F}$) can be used to specify that entries in the matrix have fixed values, which in turn can imply removing some of the unidentified aspects of $\mathbf{A}$. We can also consider the use of $\mathcal{L}_1$ penalties on $\mathbf{F}$ and/or $\mathbf{A}$ in order to induce data-defined sparsity in either matrix.

In the specific case of pure least squares estimation, the minimizer of G when $\lambda_2 = \lambda_{12} = 0$ is a simple modification of (1):

$$\mathbf{F}(\mathbf{A}) = \mathbf{X}\mathbf{A}'(\mathbf{A}\mathbf{A}' + \lambda_1\mathbf{R})^{-1}, \tag{3}$$

but for $\lambda_2 > 0$ or $\lambda_{12} = 0$ as

$$\mathbf{F}(\mathbf{A}) = [\mathbf{I}_N \otimes (\mathbf{A}\mathbf{A}') + \lambda_1 \mathbf{I}_N \otimes \mathbf{R}_1 + \lambda_2 \mathbf{R}_2 \otimes \mathbf{I}_K + \lambda_{12}\mathbf{R}_{12}]^{-1}\mathrm{vec}(\mathbf{A}\mathbf{X}') \tag{4}$$

where $\mathbf{I}_N$ and $\mathbf{I}_K$ are identity matrices of order N and K, respectively and $\mathrm{vec}(\mathbf{A}\mathbf{X}')$ is matrix $\mathbf{A}\mathbf{X}'$ written column-wise as a vector.

Next we define the outer factor coefficient criterion

$$H(\mathbf{A}) = \sum_j^n h_j[\mathbf{x}_j, \mathbf{F}(\mathbf{A})\mathbf{a}_j] + P_H(\mathbf{A}) \tag{5}$$

The options for defining the $\mathbf{A}$ regularization term P_H contain quadratic row-, column- and vector-regularization strategies analogous to those defined above for $\mathbf{F}$.

Even with regularization, it may be that a simple rescaling of $\mathbf{A}$ can usually be matched by the inverse scaling of $\mathbf{F}$, and for more stable computation it may be useful to assume the scale restriction

$$\|\mathbf{A}\| = \rho. \tag{6}$$

The optimization H on the nK-hypersphere of radius ρ can be achieved by converting the cartesian coordinates $\mathrm{vec}(\mathbf{A})$ in nK space to hyperspherical angular coordinates of dimension $nK - 1$. Nevertheless, substantial non-identifiability may remain after the choice of G and H, and since the optimum may not be unique, second order optimization methods such as Newton-Raphson may fail. But quasi-Newton and conjugate gradient methods work fine in our experience, even when the hessian matrices associated with G and H are singular.

Of course, the nice explicit conditional minimization results (3) and (3) vanish if something other than least squares criteria are involved in the specification of G and H, and in this case parameter cascading requires that G be re-optimized each time $\mathbf{A}$ is modified. Moreover, the gradient of H must take into account that $\mathbf{F}$ is a function of $\mathbf{A}$. That is, the total derivative

$$\frac{dH}{d\mathbf{A}} = \frac{\partial H}{\partial \mathbf{A}} + \left[\frac{\partial H}{\partial \mathbf{F}}\right]\left[\frac{d\mathbf{F}}{d\mathbf{A}}\right]$$

From

$$\frac{d}{d\mathbf{A}}\left(\frac{J}{\partial \mathbf{F}}\right) = \frac{\partial^2 J}{\partial \mathbf{A} \partial \mathbf{F}} + \left[\frac{\partial^2 J}{\partial^2 \mathbf{F}}\right]\left[\frac{d\mathbf{F}}{d\mathbf{A}}\right] = 0$$

at the conditional optimum of J, and assuming that at that point the hessian matrix is nonsingular, we have that

$$\frac{d}{d\mathbf{A}}\left[\frac{\partial J}{\partial \mathbf{F}}\right] = \frac{d\mathbf{F}}{d\mathbf{A}} + \left[\frac{\partial^2 J}{\partial^2 \mathbf{F}}\right]^{-1}\left[\frac{d\mathbf{F}}{d\mathbf{A}}\right] = 0$$

and we arrive at the result known as the implicit function theorem,

$$\frac{dH}{d\mathbf{A}} = \frac{\partial H}{\partial \mathbf{A}} - \left[\frac{\partial H}{\partial \mathbf{F}}\right]\left[\frac{\partial^2 J}{\partial^2 \mathbf{F}}\right]^{-1}\left[\frac{d\mathbf{F}}{d\mathbf{A}}\right].$$

2.2 The evaluation and interpretation of the subspace

If the K dimensional subspace is the model, and the coordinates in the rows of $\mathbf{A}$ are in incidental to how the subspace is identified, then it becomes essential to use only quantities that are invariant with respect to a choice of coordinate system in evaluating the fit of a subspace to a known target, its sampling variance, or arriving at a description or interpretation of what space "means" in real-world terms. Fortunately, appropriate tools are readily at hand. Canonical correlation assesses the congruence of two subpaces of $\mathcal{R}^{\mathcal{N}}$ by the eigenanalysis of the product $\mathbf{P}_1\mathbf{P}_2$ of the order N projection matrices that define the image of each column of $\mathbf{X}$ on the respective subspace. The eigenvalues ρ_k^2 by order of size are the squared multiple correlations between the associated pair of canonical variables and, as in PCA, the sum of these eigenvalues is an overall measure relative to K of how closely the two subspaces coincide. Moreover, the smallest ρ_k^2 can be interpreted as the cosine of the smallest angle achievable between two vectors in the respective subspaces.

The interpretation of the subspace is necessarily within the context of the variables associated with the columns $\mathbf{x}_j$ of $\mathbf{X}$. A measure of the extent to which a column j is representable within the subspace is

$$h_j^2 = \frac{\|\mathbf{P}\mathbf{x}_j\|^2}{\|\mathbf{x}\|^2} \tag{7}$$

where $\mathbf{P} = \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'$ projects an N-vector on to the subspace. Squared multiple correlation measure h_j^2 is called the "communality" of the variable in the social science literature. A good description of a subspace will choose terms that take account of communalities that are near zero as well as those near one.

2.3 An illustration of cascaded multivariate PCA

Figure 1 illustrates the use of this regularization framework for an artificial problem with $N = 200, n = 5$ and $K = 2.$. The true factor scores in $\mathbf{F}_{pop}$ were 200 equally spaced points on a circle of unit radius, the 5 true factor loadings in the first row of $\mathbf{A}_{pop}$ were all one, and the loadings in its second row were (-2,-1,0,1,2). The data matrix was constructed by adding random normally distributed errors to $\mathbf{X}_{pop} = \mathbf{F}_{pop}\mathbf{A}_{pop}$ with mean zero and standard deviation 0.5. The left panel of Figure 1 shows the estimated and true factor scores using least squares criteria for an analysis with no regularization on either $\mathbf{F}$ or $\mathbf{A}$ but with the norm constraint (6) on $\mathbf{A}$. The error in the data results in factor scores having a roughly elliptical distribution. Using canonical correlation analysis to measure the agreement between the estimated and true $\mathbf{A}$'s results in squared canonical correlations $\rho_1^2 = 0.99991$ and $\rho_2^2 = 0.99775$, indicating an excellent level of agreement between the estimated and true two-dimensional subspaces. That is, although the estimated $\mathbf{A}$ differs substantially from its true counterpart, its rows span almost exactly the same subspace.

Next we set the column roughness penalty matrix for the estimate of $\mathbf{F}$ to $\mathbf{S} = \mathbf{I}_N - \mathbf{F}_{pop}(\mathbf{F}'_{pop}\mathbf{F}_{pop})^{-1}\mathbf{F}'_{pop}$. That is, we asked that the factor score estimates fall within the same subspace as is spanned by the two columns of its true counterpart. This, of course, does not force the estimate to be circular or constrain its orientation

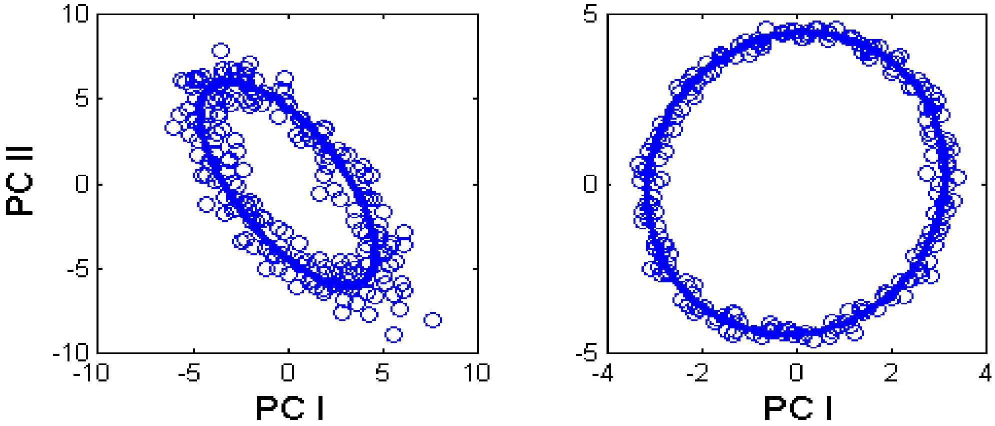

Figure 1: The left panel displays estimated factor scores as open circles and the scores used to generate the data as asterisks. No regularization was used for this analysis. The right panel shows estimated scores when the column roughness was defined by the projection of scores on to the complement of the space spanned by the true scores, using $\lambda_2 = 10$.

if it is not. The right panel Figure 1 shows the result of using a smoothing parameter $\lambda_2 = 10$ for this penalty. We see that the estimated scores now correspond more closely to a linear transformation of the true values. The two squared canonical correlations are now 0.99986 and 0.99808, still indicating a fine subspace recovery; and the values of $H(\mathbf{A})$ were 135.2 and 184.3 for the unregularized and regularized fits, respectively.

2.4 Partial least squares variance components

Estimating a K- dimensional subspace of the space of dimension n spanned by the columns of $\mathbf{X}$ can be generalized in many ways. Partial least squares regression (PLS) estimates a subspace of dimension K within which the approximation of an external set of data in N by m matrix $\mathbf{Y}$ is optimal in a least squares sense. PLS has been found to be preferable to principal components regression in situations where a substantial number of nearly multicollinear covariates are available, and where it considered likely that there is important explanatory variation in the complement of the PCA subspace. PLS has been found to be especially in chemical engineering and chemometrics where the goal is often to locate compounds imaged in some type of spectrum that play an important role in predicting continuous or classification outcomes.

The term "partial least squares" is usually identified with an algorithm developed by H. Wold (1975) and S. Wold (1976) that estimates such an approximation for a dependent vector $\mathbf{y}$ through K pairs of weighted averages followed by a deflation of an iterated approximation of $\mathbf{X}$:

$$\begin{aligned}\mathbf{p}^{(k)} &= \mathbf{X}^{(k)\prime}\mathbf{t}^{(k)}/\|\mathbf{t}^{(k)}\|^2 \\ \mathbf{t}^{(k)} &= \mathbf{X}^{(k)}(\mathbf{X}^{(k)\prime}\mathbf{y}) \\ \mathbf{X}^{(k+1)} &= \mathbf{X}^{(k)\prime} - \mathbf{t}^{(k)}\mathbf{p}^{(k)\prime}, k = 1, \ldots, K.\end{aligned} \tag{8}$$

These iterations are initialized by $\mathbf{X}^{(1)} = \mathbf{X}$ and $\mathbf{t}^{(1)} = \mathbf{X}^{(1)}(\mathbf{X}^{(1)\prime}\mathbf{y})/\|\mathbf{X}^{(1)\prime}\mathbf{y}\|$. In spite of its simple structure, the algorithm has performed well enough in applications to ensure its incorporation into a wide range of software environments, including Matlab and R. However, there is no assurance in the algorithm's design that the subspace spanned by $[\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(K)}]$ provides an optimal approximation of $\mathbf{y}$.

Keeping within the least squares framework and functional relationship (1), we generalize fitting criterion (5) using the approach of de Jong and Kiers (1992) as follows:

$$H(\mathbf{A}|\mathbf{X}, \mathbf{Y}) = (1-\gamma)\|\mathbf{X} - \mathbf{F}(\mathbf{A})\mathbf{A}\|^2 + \gamma\|\mathbf{Y}'\mathbf{Q}(\mathbf{A})\mathbf{Y}\|^2. \tag{9}$$

where the relaxation parameter $\gamma \in [0, 1]$ and

$$\mathbf{Q}(\mathbf{A}) = \mathbf{I} - \mathbf{F}(\mathbf{A})[\mathbf{F}(\mathbf{A})'\mathbf{F}(\mathbf{A})]^{-1}\mathbf{F}(\mathbf{A})'. \tag{10}$$

The second term in the criterion measures the extent to which $\mathbf{Y}$ is unpredictable from within the subspace defined by the factor loadings in $\mathbf{A}$. The boundary conditions $\gamma = 0$ and $\gamma = 1$ correspond to pure principal components analysis and pure optimal partial least squares, respectively; and the user has the potential to find a compromise between these objectives that satisfies other objectives such as the interpretability of the subspace itself. For unregularized lease squares estimation, de Jong and Kiers (1992) showed that a minimizing solution can be obtained by the eigenalysis of

$$\mathbf{S}_\gamma = (1-\gamma)\mathbf{X}\mathbf{X}' + \gamma\mathbf{Y}\mathbf{Y}'. \tag{11}$$

Again we see that this generalization of PCA and PLS is a nonlinear least squares problem indexed by parameter $\mathbf{A}$, so that inference methods available for such problems are readily available.

3 Functional PCA

Only minor modifications are required for the analysis of functional data, where the matrix $\mathbf{X}$ is replaced by a set of N functions x_i in N-vector $\mathbf{x}$, and the principal coefficient matrix $\mathbf{A}$ becomes the functional K-vector with values $\mathbf{a}(t)$. Let $\mathbf{a}$ be represented by an expansion in terms of L basis functions ϕ_ℓ, so that $\mathbf{a}(t) = \mathbf{C}\boldsymbol{\phi}(t)$, with $\mathbf{C}$ being the L by K matrix of coefficients of the expansion.

In effect, the computation is equivalent to that of a multivariate analysis in coefficient space in the metric $\mathbf{W} = \int \boldsymbol{\phi}(t)\boldsymbol{\phi}'(t)dt$ defined by the $\mathbf{a}$-basis functions and with data $\mathbf{X} = \int \mathbf{x}\boldsymbol{\phi}'(t)dt$; that is, expressions (3) and (4) become

$$\begin{aligned} \mathbf{F}(\mathbf{a}) &= \mathbf{X}\mathbf{A}'(\mathbf{A}\mathbf{W}\mathbf{A} + \lambda_1\mathbf{R}_1)^{-1} \\ \mathbf{F}(\mathbf{a}) &= [\mathbf{I}_N \otimes (\mathbf{C}'\mathbf{W}\mathbf{C}) + \lambda_1\mathbf{I}_N \otimes \mathbf{R}_1 + \lambda_2\mathbf{R}_2 \otimes \mathbf{I}_K + \lambda_{12}\mathbf{R}_{12}]^{-1}\mathrm{vec}(\mathbf{C}'\mathbf{X}'). \quad (12) \end{aligned}$$

The objective function H is now $H(\mathbf{a}|\mathbf{x}) = \|\mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{a}\|^2 + P_H$ where the norm in the first term is now defined by summing over i and integrating over t.

3.1 Principal components of the Berkeley growth data accelerations

The Berkeley growth data discussed in Ramsay and Silverman (2005) provide a number of interesting functional data analysis problems. The functional PCA is illustrated here by the analysis over the age range of three to eighteen years of 54 second height derivative curves for females, calculated from carefully tuned strictly monotone smooths of the original discrete data, and displayed in Figure 2.

In order to reduce computation time, each curve was represented by re-fitting these height accelerations evaluated over 501 equally spaced points using 53 B-spline basis functions with equally spaced knots. The factor loading functions $\mathbf{a}$ were represented by 23 B-spline basis functions with equally spaced knots. The age range was also started at three years rather than one year in order to eliminate the increased instability of acceleration estimates in infancy. Features to be noted are the wide variation in the age of the middle of the pubertal growth spurt, marked by the crossing of zero with negative slope at around 12 years; and also the single large acceleration variation between three and seven years.

The analyses were first carried out without any regularization for either $\mathbf{F}(\mathbf{a})$ or $\mathbf{a}$; and for $K = 1, \ldots, 4$. The squared multiple correlations $R^2 = (\mathtt{SSY} - \mathtt{SSE})/\mathtt{SSY}$,

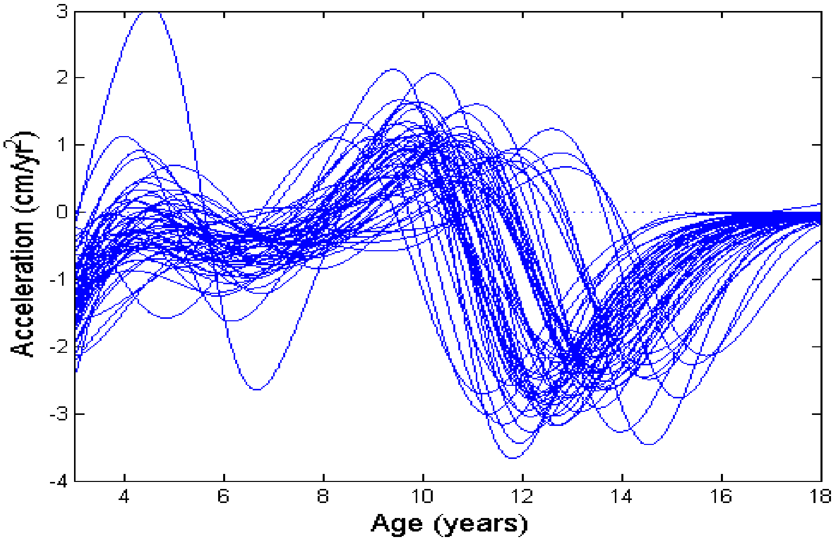

Figure 2: The second derivatives of 54 female height curves estimated from the Berkeley growth data, but restricted to begin at three years rather than the original one year.

where $\mathtt{SSY} = \sum_i \int x_i^2(t)\,dt$ and $SSE = \sum_i \int [x_i(t) - \mathbf{f}_i\mathbf{a}(t)]^2\,dt$, for these analyses were 0.683, 0.887, 0.941 and 0.965, respectively; and these values agreed with the corresponding functional principal components values described in Ramsay and Silverman (2005) to four decimal places. By inspecting the root mean squared residual functions $\mathtt{RMSE}(t) = \sqrt{N^{-1}\sum_i [x_i(t) - \hat{x}_i(t)]^2}$ for each number of factors, shown in Figure 3, we can see what each increase in dimensionality of the subspace contributes to the fit. Steady improvement occurs over $K = 2, 3, 4$ at age 4, 10.5 and 16, but rather less improvement is provided for the stable later childhood ages 6 to 9 years. Much of the improvement after age 10 is no doubt due to the improved capacity of higher dimensional fits to account for the variable age of puberty.

We then focussed on the two-dimensional PCA in order to consider some regularization. Figure 4 displays the two principal component functions $a_1(t)$ and $a_(2)$ that arise from the unregularized PCA initialized by the eigenfunctions of the covariance operator computed from the functional data. As expected, these initial components are already optimal, and no further improvement was achieved by optimizing H. The second component indicated by a dashed line resembles closely the typical shape of an acceleration function, but only accounts for 20% of the variation. The dominant component plotted as a solid curve accounts for 67% of the variance, and acts when added to the second component to shift this component forward or backward in time. In short, the dominant component of variation in these curves is essentially phase variation, percentage that it represents is close to that given in Kneip and Ramsay (2008).

The factor scores associated with the unregularized PCA are shown in Figure 5. We see that these scores tend to be approximately distributed along the segment of a circle with center at the origin. Figure 6 plots the distances of the scores from

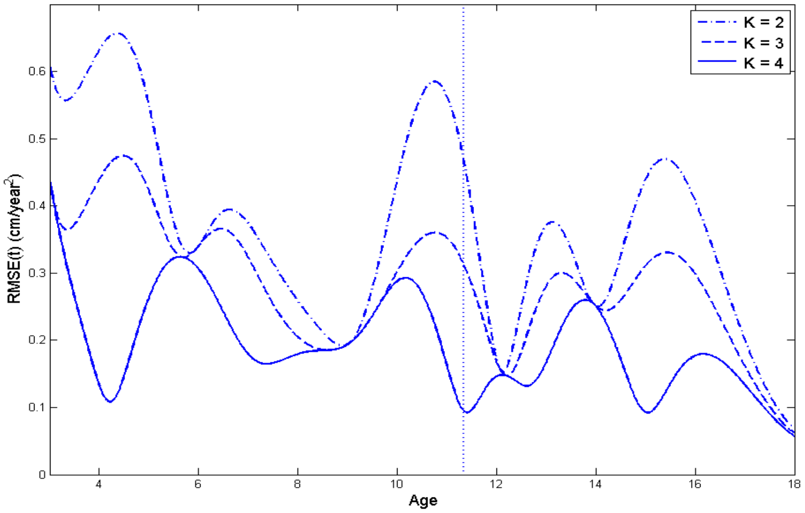

Figure 3: The root mean squared residual for unregistered functional principal component fits to the 54 Berkeley growth acceleration curves for females. The number of factors ranged from 2 to 4. The vertical dotted line indicates the average age of middle of the pubertal growth spurt in this sample.

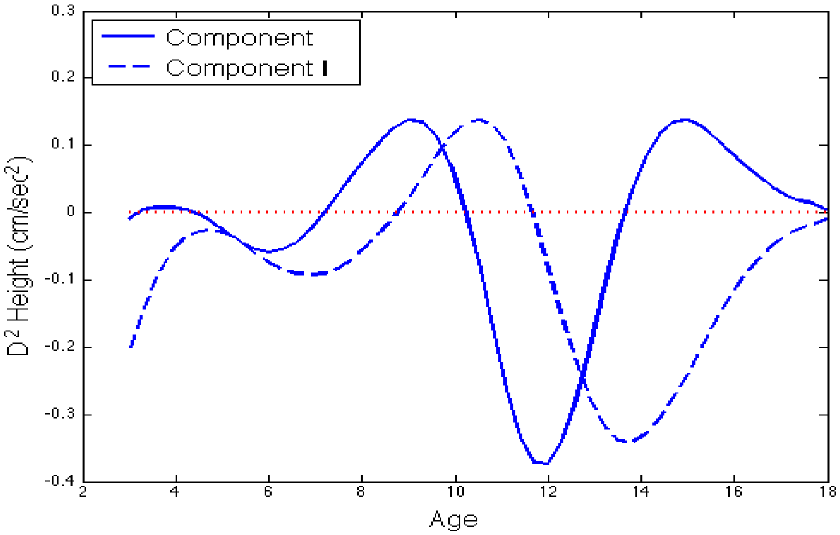

Figure 4: The first two principal component functions for an unregularized functional PCA of the growth acceleration functions in Figure 2. The solid and dashed lines corrrespond to principal components representing 67% and 20% of the variation in the original data.

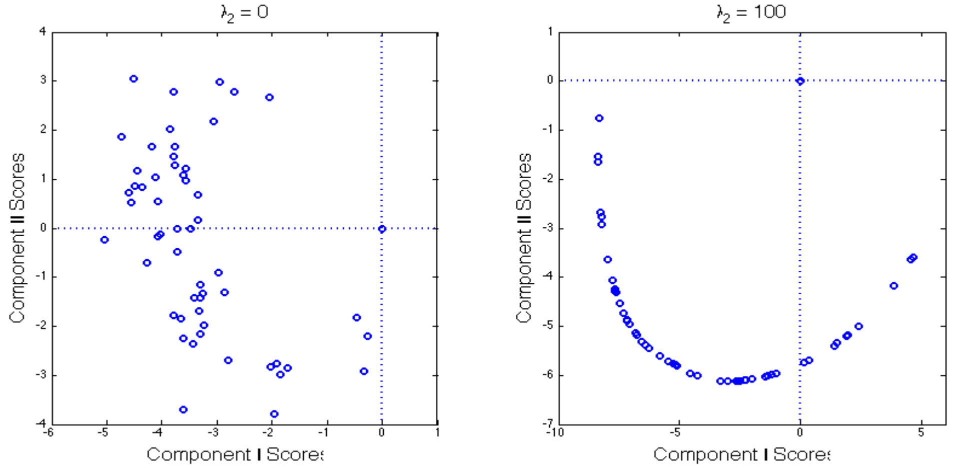

Figure 5: The left panel plots as dots the factor scores derived from an unregularized PCA of the Berkeley growth acceleration curves. The circles in the right panel define a curved trajectory arising from a heavily regularized PCA result from smoothing the PCA results toward an arc defined by the smooth trend in Figure 6.

this origin against the angle of their relations to the origin. We see that there is an evident descending trend. This arises because the distances from the origin represent the intensities of the growth spurts, and the angles are associated with how late the pubertal spurt occurs. Early spurts are intense so as to compensate for a few years to lost growth, while late spurts are mild because more prior growth has already occurred. The Figure also shows a smoothing spline approximation of this trend, and the arc plotted in the left panel Figure 5 is constructed from this smooth.

We now consider a regularization of the PCA. We use the second term in (12) to smooth the columns of $\mathbf{F}$ toward the columns of the smooth approximation resulting from converting polar coordinate angle/distance pairs into Cartesian coordinates. The roughness penalty matrix $\mathbf{R}_2$ projects the columns of $\mathbf{F}$ on to the complement of the space spanned by the columns of this smooth approximation to $\mathbf{F}$. The right panel of Figure 5 displays the result of a rather heavy-handed smoothing using $\lambda_2 = 100$, and we see that the scores now are distributed almost exactly on the approximation. It is interesting to consider how much this affects the fit to the data; the error sum of squares increases from 127.2 for the unregularized analysis to 144.7 for this heavy regularized. The associated squared multiple correlation is $R^2 = 0.12$. If consider that the right panel is associated with a pure phase variation model, then this again suggests that only about 12% of the variation in the data is due to amplitude variation in the dashed curved display in Figure 4. Of course, more principal components increase this percentage, but only slightly it turns out.

3.2 Partial least squares for weather data

The weather data described in Ramsay and Silverman (2005) are the context for an interesting comparison of regression on principal components versus the functional version of partial least squares. In general temperature is primarily of interest as an

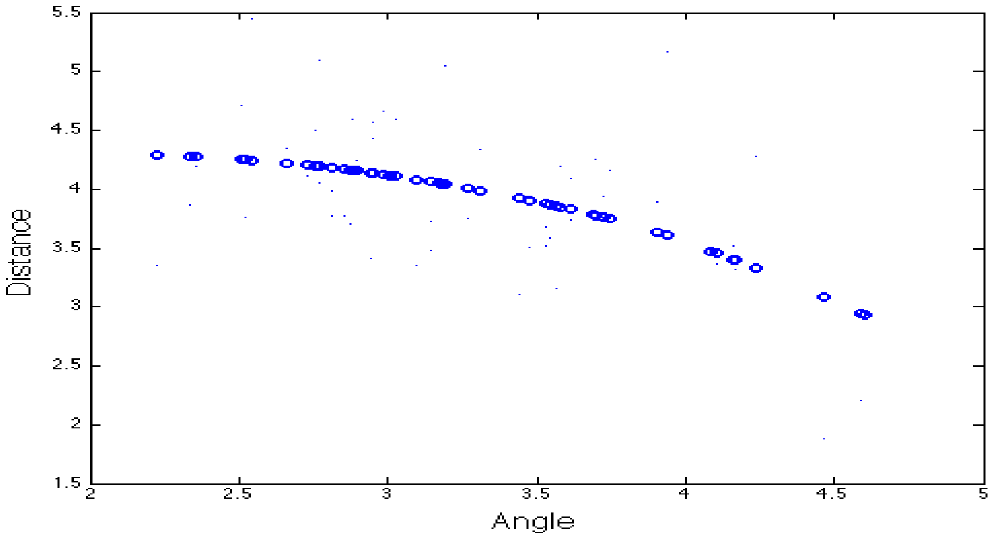

Figure 6: The relation of distances from the origin (0,0) of the unregularized factor scores to their corresponding angles from the positive horizontal. The smooth curve is a function estimated by a smoothing spline with 56 equally spaced knots and a smoothing parameter value of 1.0.

obvious explanatory variable for a variety of processes that are of more direct interest. Among these, precipitation is the most important for crop scientists and economies like Canada's that export a large amount of food products. Here we predict the mean daily precipitation for 35 weather stations, taken over the entire year and over 34 years of observation.

Since the winter months exhibit much more variability, we run the year from July 1 to June 30. An low-resolution expansion of the daily temperature data averaged over 34 years is shown in Figure 10. We attempt to predict mean precipitation in tenths of a millimetre from two functional components which are used to take linear combinations of these 35 curves. The components PCA and PLS are shown in Figure , and the component scores are displayed in Figure

The PCA portion of the component plot shows no surprises; the dominant component is mean seasonal variation and the subdominant component is roughly the annual mean temperature. These two components account for 98.9% of the variation in the temperature curves. The corresponding component scores have a mild dispersion around a curved trajectory, with the arctic stations being on the lower right and pacific stations on the upper left. But the PLS components seem to measure something quite different, and the corresponding component scores are virtually linear in their distribution. Both components have a strong five-cycle periodic shape, with the main difference being in the variation of amplitudes of these cycles across the year.

Figure **??** displays the respective fits to the external mean precipitation variable. The PCA components have a squared multiple correlation of $R^2 = 0.41$, while the PLS components to much better with $R^2 = 0.78$. Both correlations would be slightly better with the elimination of Prince Rupert, the wettest place in Canada. We now

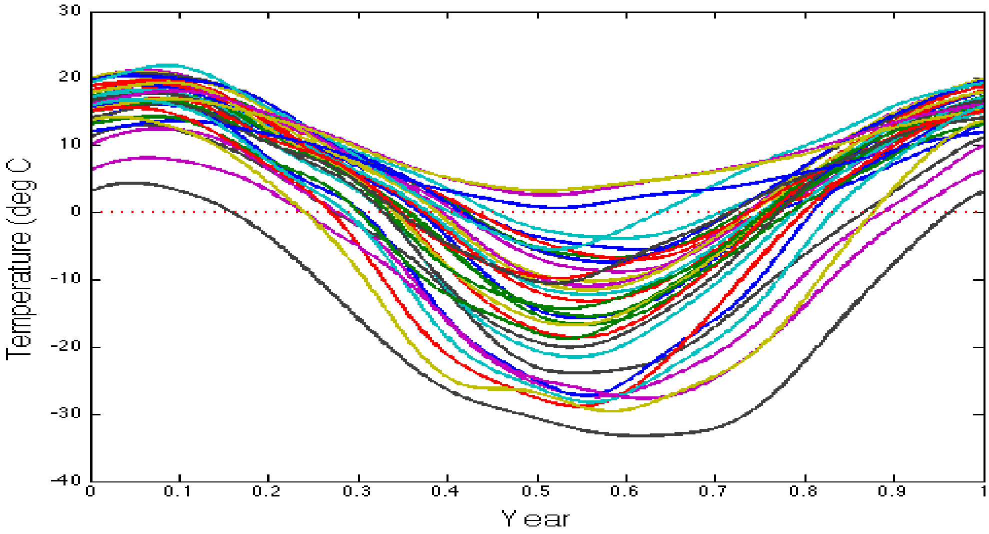

Figure 7: Temperature curves for 35 Canadian weather stations over the year beginning July 1 and ending June 30.

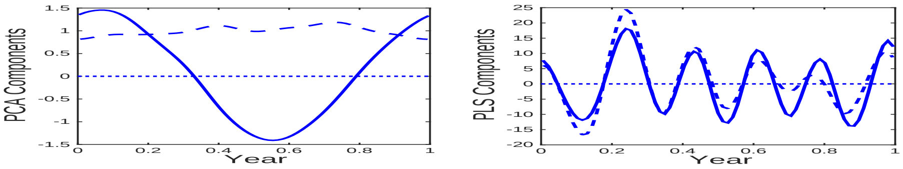

Figure 8: The left panel shows the first two functional principal components for the temperature curves in Figure and the right panel shows the two functional components optimizing the fit to the 35 mean precipitations. In each panel the solid line is the first and dominant component and the dashed line is the subdominant component.

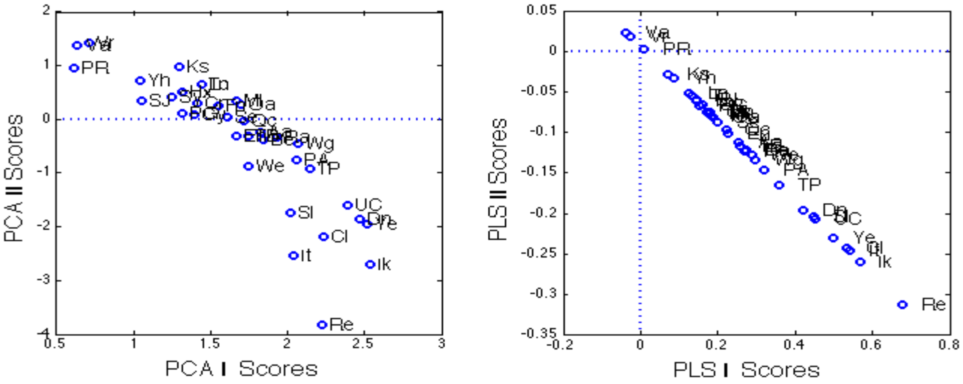

Figure 9: The left panel shows the first two sets of 35 functional principal component scores for the temperature curves in Figure and the right panel shows the functional component scores optimizing the fit to the 35 mean precipitations.

see five clusters in the PLS fit, corresponding to, going from left to right, arctic, subarctic, continental, eastern and Pacific stations, and we see now that the five-cycle pattern in the right panel of the components plot is designed to select out stations in this order. The time of the year in which this pattern crosses zero with a positive slope can be seen a time of higher slope than is usual for each curve.

Canonical correlation analysis applied to the two score matrices after subtracting their respective column means yields squared correlations of 1.0 and 0.175, indicating that there is a direction in the PCA plane that is unaltered in the PLS counterpart, relative to which the correlation between the respective orthogonal directions is close to zero.

4 Functional PCA with registration

Suppose now that we are in the functional principal component context with N data functions x_i; and, as in Section 3, the K principal components functions a_k define the approximation $\hat{x_i} = \mathbf{f}_i(\mathbf{a})\mathbf{a}$. But now we consider that there is phase variation from one observed function to another, so that instead of measuring time or whatever else t represents on a fixed scale, we need to apply a smooth strictly increasing transformation $\tau_i(t)$ of standard time prior to computing the value of the functions a_k and the factor scores $\mathbf{f}_i(\mathbf{a})$ that they define. This case is discussed in detail in Kneip and Ramsay (2008).

A family of a smooth strictly monotonic functions τ that is defined in terms of a

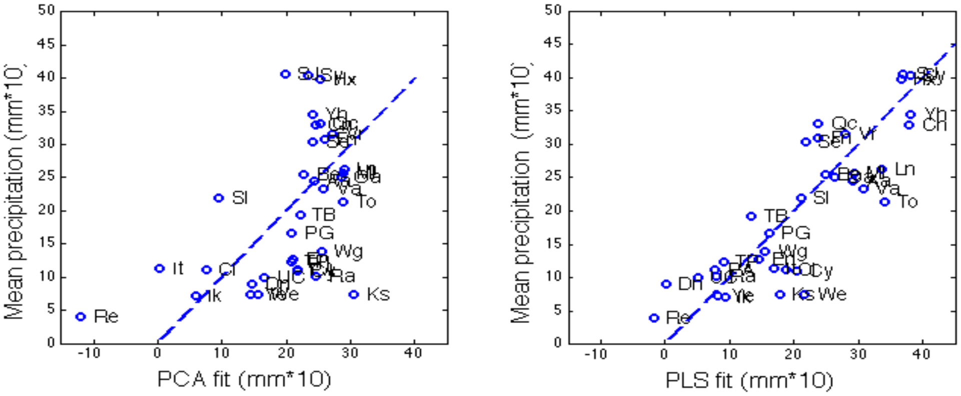

Figure 10: The left panel shows the approximation of the mean precipitation scores provided by the functional principal component of the temperature curves in Figure and the right panel shows the PLS fit to the 35 mean precipitations. Prince Rupert an outlier, and plots above the vertical scale of the plots.

basis function expansion is

$$\tau(t|\mathbf{d}_i') = T\frac{\int_0^t \exp[\mathbf{d}_i'\boldsymbol{\psi}(z)]\,dz}{\int_0^T \exp[\mathbf{d}_i'\boldsymbol{\psi}(z)]\,dz} \tag{13}$$

where $t \in [0, T]$, $\boldsymbol{\psi}$ is a vector of M basis functions and $\mathbf{d}_i$ is a vector of coefficients of length L for the ith case defining the basis function expansion in the log of the integrand. Let L by N matrix $\mathbf{D}$ contain the $\mathbf{d}_i$'s in its columns.

The principal component coefficient function for approximating the function value $x_i(t)$ are now $\boldsymbol{\alpha}[\tau(t|\mathbf{d}_i)]$. Notice that, according to this notation, each time the coefficients $\mathbf{C}$ are changed, we must revaluate the coefficients $\mathbf{d}_i$, so that the time-warping coefficients in N by M matrix $\mathbf{D}$ are themselves functions $\mathbf{D}(\mathbf{C})$ of structural parameter parameters $\mathbf{C}$, and hence nuisance parameters. In turn, we also still retain the relation $\mathbf{F}(\mathbf{a})$, which now therefore depends on the value of $\mathbf{D}$ as well. That is, we have a three-level cascade of parameters, indicated by $\mathbf{F}[\mathbf{D}(\mathbf{C})]$.

The function $\mathbf{F}(\mathbf{C})$ is no longer linear, as it was in (12, since it is mediated by what is usually a nonlinear time warping function, such as (13). In the case of the least squares criterion for H, this implies that each time $\mathbf{C}$ is modified in an outer optimization of criterion $H(\mathbf{C})$, an inner optimization $G(\mathbf{D}|\mathbf{C})$ is required, which in turn makes use of the linear relation $\mathbf{F}(\mathbf{a})$ in (4).

It is known (Kneip and Ramsay, 2008) that the inner optimization criterion $G(\mathbf{D}|\mathbf{C})$ ought not to be a least squares measure, since this can produce undesirable estimates of phase variation. Instead, it has been found quite satisfactory to minimize the

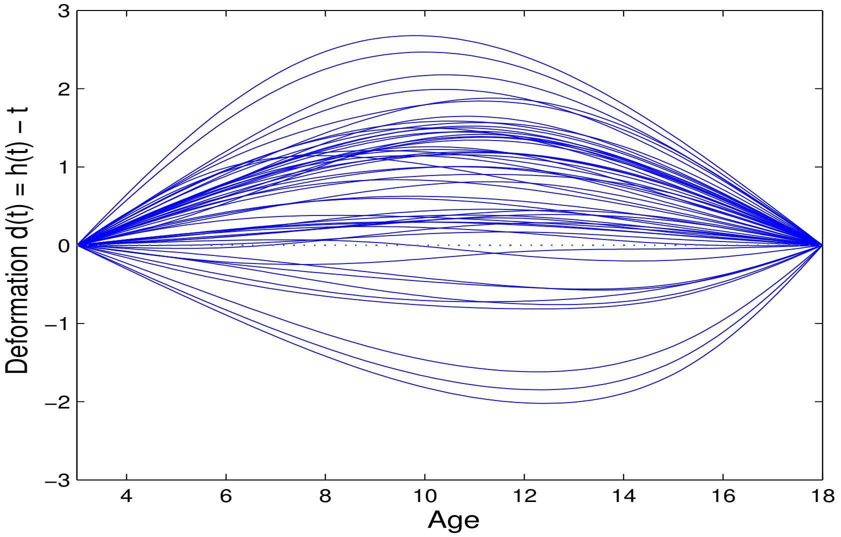

Figure 11: The estimated deformation functions $\delta(t) = \tau(t) - t$

smalles–eigenvalue–of–crossproduct criterion (SEofC)

$$G(\mathbf{D}|\mathbf{C}) = \sum_i^N \texttt{MINEIG} \left[\begin{array}{ll} \int \{x_i(t)\}^2 \, dt & \int x_i(t) \, \hat{x}_i[\tau(t|\mathbf{d}_i)] \, dt \\ \int x_i(t) \, \hat{x}_i[\tau(t|\mathbf{d}_i)] \, dt & \int \{\hat{x}_i[\tau(t|\mathbf{d}_i)]\}^2 \, dt \end{array} \right] + \lambda_1 \mathbf{d}_i' \mathbf{V} \mathbf{d}_i. \tag{14}$$

The second term in this criterion permits control over the smoothness of the warping functions. Note that here and elsewhere in the computation inner products of the form $\langle f, g \rangle_\tau = \int (f \circ h)(t)(g \circ h)(t) \, dt$. See the Appendix for comments on how these may be accurately approximated by the use of a pair of ordinary differential equations.

The principal components analysis of the Berkeley growth accelerations was repeated for $K = 1, 2, 3$, but this time registering the principal component fit to the functional data. Seven B-spline basis functions with equally spaced knots were used for the ψ_ℓ's, and the integrated squared first derivative roughness penalty was multiplied by a roughness parameter $\lambda_1 = 1$. The estimated deformation functions $\delta(t) = \tau(t) - t$ for $K = 2$ are displayed in Figure 11, and we see that the deformations tend to be simple in shape and maximal excursions of about two and half years. A satellite PCA of the deformation functions reveals only a single large principal component.

The squared multiple correlations were 0.924, 0.969 and 0.979, respectively; and these values are each comparable to the corresponding unregistered PCA using two more factors. Figure 12 now shows relatively simple increments to fit as K passes from one to three; The second dimension introduces a factor that accommodates the single large acceleration excursion in early childhood in Figure 2, and the third factor does a nice job of capturing additional variation in the pubertal growth spurt years that is not due to the timing of the pubertal growth spurt. In other words, this analysis neatly separates amplitude variation (three dimensional) from phase variation (primarily one-dimensional).

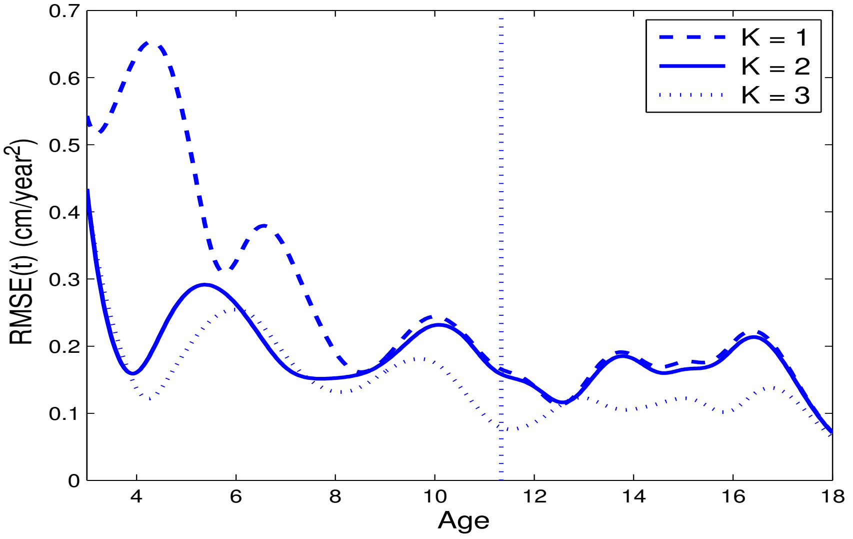

Figure 12: The root mean squared residual for registered functional principal component fits to the 54 Berkeley growth acceleration curves for females. The number of factors ranged from 2 to 4. The vertical dotted line indicates the average age of middle of the pubertal growth spurt in this sample.

Bibliography

Bates, D. M. and Watts, D. B. (1988) *Nonlinear Regression Analysis and Its Applications*, New York: Wiley.

Cao, J. and Ramsay, J. O. (2007) Parameter cascades and profiling in functional data analysis. *Computational Statistics,* 22, 335-351.

Cao, J. and Ramsay, J. O. (2008) Generalized profiling estimation for global and adaptive penalized spline smoothing. *Computational Statistics and Data Analysis,* 53, 2550-2563.

Cao, J. and Ramsay, J. O. (2010) Linear mixed-effects modeling by parameter cascading. *Journal of the American Statistical Association,* 105, 365-374.

de Jong, S. and Kiers, H. (1992) Principal covariates regression: Part I. Theory. *Chemometrics and Intelligent Laboratory Systems,* 14, 155-164.

Kneip, A. and Ramsay, J. O. (2008) Combining registration and fitting for functional linear models. *Journal of the American Statistical Association,* 103, 1155-1165.

Neyman, J. and Scott, E. L. (1948) Consistent estimates based on partially consistent observations. *Econometrika,* 16, 1-32.

Ramsay, J. O. & Silverman, B. W. (2005) *Functional Data Analysis,* New York: Springer.

Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J. (2007) Parameter estimation for differential equations: A generalized profiling approach (with discussion). *Journal of the Royal Statistical Society, Series B,* 69, 741-796.