

Chapter 2

Background

Analyzing the variation in common shape curves and more specifically thermal performance curves is the subject of this dissertation and related work in evolutionary biology, FDA, and in shape statistics on this topic is discussed in this chapter. Shape invariant models used in FDA is discussed in Section 2.1. Previous work on linear and nonlinear methods for estimating the variation, and decomposition and quantification of the variation is discussed in Section 2.2. The main ideas that contrast our TMV method to previous work are that in TMV the variation is decomposed into **predetermined** modes of interest, some of which **nonlinear** and the method to **quantify** nonlinear modes in TMV takes into account the **geometry** of the space of variation.

2.1 Shape invariant model

The model used in TMV is a particular case of the shape invariant model which was suggested for curves in [Lawton et al. (1972)]. In the Shape Invariant Model (SIM) each curve f_i is modelled as a parametric transformation of the common shape f by the height parameter h_i ,

the location parameters m_i and amplitude parameters a_i and b_i

$$f_i(t) = a_i f(b_i(t - m_i)) + h_i$$

Several papers in FDA used similar models to SIM to model variations in a set of curves with common shape but the focus was more on estimating the common shape or realigning the curves with respect to the nonlinear variations. A semiparametric model to realign the curves was proposed in [Hardle and Marron (1990)]. In [Kneip and Gasser (1988)] and [Kneip et al. (2000)] a structural average estimating the common shape was derived by aligning chosen landmarks of the curves and estimating the alignment function. In [Wang and Gasser (1997)], and [Wang and Gasser (1999)] time warping techniques were used to align the curves, which does not require a choice of landmarks. When the common shape is an image or an object instead of a curve, additional transformation such as rotations were discussed in [Bookstein (1991)], [Small (1996)], and [Dryden and Mardia (1998)]. The focus in this dissertation is not on estimating the template shape or estimating the variance-covariance in the data but decomposing the variation in the data into modes of interest and quantifying each mode.

2.2 Analyzing the Variation

A common approach to analyze the genetic variation is to first estimate the genetic covariance matrix G (or its continuous counterpart, the covariance function) and then decompose this estimated genetic matrix using Principal Component Analysis (PCA). The genetic covariance matrix G is generally estimated from a mixed effect model fit to the data using a restricted maximum likelihood estimation. For examples of the model and applications, see

[Meyer (1991)], [Kirkpatrick, Hill, and Thompson (1994)], [Meyer (1998)], [Meyer (2001)]. In the caterpillar example presented in Chapter 1, this approach would model the temperature effect as a fixed environmental effect, the family effect as a random genetic effect and the interaction family*environment as a random effect, see [Kingsolver et al. (2004)]. Restricted maximum likelihood method estimates the coefficients in the matrix G by maximizing the likelihood under some constraints on the components of G . These constraints incorporate genetic information, such as genealogy. Once the genetic covariance matrix G is estimated, it is decomposed using Principal Component Analysis. Our method TMV does not make any distribution assumption, it does not estimate the parameters by maximizing the likelihood but rather minimizing the L_2 error. Moreover, as we see in Subsection 2.2.1, the PCA decomposes the variation along linear modes and some of the modes of interest to biologists are nonlinear. In this dissertation, the focus is not on estimating the genetic variation but on decomposing the genetic variation. Principal Component Analysis (PCA) is discussed in Subsection 2.2.1. In Subsection 2.2.2, several proposed alternative methods for the nonlinear case are contrasted to TMV .

2.2.1 Linear decomposition: Principal Component Analysis

Show graphic for PCA for a linear case, and a PCA for nonlinear case.

A classical method to analyze variation in multivariate data is Principal Component Analysis (PCA). PCA finds the linear orthogonal directions of variation of greatest variability in the data. It is used sometimes to reduce the dimensionality of the data since the number of these linear directions is usually smaller than the original dimension of the data. In the discrete case, PCA has a simple linear algebra derivation, it decomposes the variance matrix in an orthonormal space of eigenvectors called principal components directions and the standard-

ized eigenvalues associated to the eigenvectors give the percentage of the variation explained by each direction, see [Anderson (1971)]. To use the structure and smoothness of functional data, an extension of PCA is used in FDA where the variance covariance function is decomposed in an orthonormal space of eigenfunctions. Details, discussion and some variations of these extensions can be found in Chapter 6 to Chapter 8 of [Ramsay and Silverman (1997)]. In quantitative biology, considering quantitative data as infinite dimensional traits or continuous curves and finding eigenfunctions and eigenvalues of the variance-covariance matrix was first suggested by [Kirkpatrick and Heckman (1989)], and this method has been widely used since then in quantitative genetics. Some examples of applications of this analysis in quantitative genetics to mice growth rate is in [Kirkpatrick et al. (1990)], and to thermal performance curves in [Gilchrist (1996)], [Kingsolver et al. (2001)], and [Kingsolver et al. (2004)]. An extension of functional principal component decomposition to include effect of covariate variables is presented in [Chiou, Muller and Wang (2003)] and [He, Muller, and Wang (2003)] where it is applied to egg laying curves for 1000 female Mediterranean fruit-flies.

A weakness of PCA in the discrete or the continuous case is that it does not effectively characterize nonlinear variations. This fact is widely recognized in the FDA literature and is the main reason why realigning functional data with respect to nonlinear directions is discussed in several papers cited in Section 2.1, because after realignment the data could be analyzed using PCA. In the parametric Principal Component Analysis (PCA) presented in [Silverman (1995)], the realignment and PCA decomposition are done both at the same time. However, quantifying nonlinear variation was not discussed in this paper as nonlinear variations were considered as noise. A second problem with PCA is that each principal component is data driven and it does not always have a biological interpretation. In our analysis of the

data in Chapter 3, the directions of variation are predetermined and of principal interest to evolutionary biologists and some of them are nonlinear.

2.2.2 Nonlinear case

There are multiple methods which account for nonlinear variation in multivariate data. Most of these methods are used in analysis of high dimensional data, such as images, for dimensionality reduction. We focus in this subsection on methods of similar geometric interpretation as TMV, such as Principal Curves, Principal Surfaces, Self Organizing Maps, and Auto-Associative models. We briefly present each method and then compare them to TMV.

Principal Curves

Principal curves were introduced in [Hastie and Stuetzle (1989)] and the algorithm was widely used¹ for multiple data to account for nonlinearity of the space of variation. For multivariate data with d variables, a principal curve is defined as any function that goes through the *middle* of the point cloud in the d dimensional space, i.e each point on a principal curve is the average of the points that project in that curve, and this property is called *self-consistency*. More precisely, a differentiable curve $\alpha : I \rightarrow \mathbb{R}^d$ is a principal curve if it satisfies the self-consistency property, $E(X|\alpha(s)) = \alpha(s)$ a.s. Principal Curves are found by iteration starting with the linear principal component and taking a projection step, an expectation step, and an updating step at each iteration. In the particular case of linear variations in the data, each linear principal component satisfies the self-consistency property which makes it a principal curve. However, unlike in the linear case, there is no inherent "ordering" of the Principal Curves. The paper by [Delicado (2001)] presents a new definition of first principal curve based

¹The article [Hastie and Stuetzle (1989)] was cited over 140 times. Most of the citations were in the computer science literature for dimensionality reduction.

on the notion of principal oriented points. The total variance that this first principal curve maximizes is based on a conditional variance given that the p -dimensional random variable lies in a hyperplane defined by a point and an orthogonal direction. This characterization allows to define locally a second principal curve, which together with the first principal curve allows to define a principal surface.

Principal Manifolds

The self-similarity property defined in the previous subsection was used to characterize *Principal manifolds*, see [Tarpey and Flury (1996)] and [Chang and Ghosh (2001)], which are higher dimensional structures going through the *middle* of the data.

Self Organizing Maps

Self-organizing maps (SOM) is a nonparametric latent variable model with a topological constraint, see [Kohonen (1995)]. Common topologies are lines, squares, or hexagonal grids. SOM mapping is similar to a discretized self-similarity principle for a Principal manifold. So, SOM serves of an approximation to the principal surface, which converges to it for a large enough number of nodes. SOM is a data driven method to reduce dimensionality, and the reduced modes are not always interpretable.

Auto-Associative Neural Network

Auto-Associative neural network model is presented as a nonlinear generalization of PCA in [Dong and McAvoy (1996)] with a statistical flavor in [Girard and Iovleff (2002)]. In this last paper, the center of variation in the manifold is taken as the expected value, which does not always lie in the manifold. Moreover, the variation along each direction is quantified by the classical ratio of sums of squares using the Euclidean distance. In TMV, the center of

variation is a point in the manifold. In addition, the quantification of the variation take into account the geometry of the space of variation.

All the nonlinear methods we described in this subsection are non-parametric data-driven methods used mainly for dimensionality reduction, and the principal curves and surfaces are not always interpretable. Although there is an attempt in [Delicado (2001)] to decompose nonlinear manifolds onto lower nonlinear principal curves, this paper does not present a way to decompose the variation in this manifold along the principal curves. In TMV, the variation in the data is decomposed along curves of variation that are of main interest.

2.2.3 Quantifying the variation

As we see in Chapter 4 the space of variation of interest in the caterpillar data is a non-linear manifold. The Fréchet mean set and the Fréchet variance, defined in Section 5.1, are the measure of the center and the associated variance of the distribution of a point cloud along a metric manifold [Fréchet (1948)]. Conditions on the manifold or choice of an appropriate distance to insure uniqueness of the Fréchet mean and consistency of the Fréchet mean estimates are discussed in several recent papers in the area of statistical shape analysis [Bhattacharya and Patrangenaru (2003)], [Le (2001)], and [Le and Kume (2000)]. We define in Chapter 5, distances on a manifold that allow for the decomposition of the variation in modes of interest. In addition, we define in this chapter a ratio sum of squares to quantify each direction.