

# Chapter 1

## Introduction

Functional data, such as curves or images, are increasingly collected in a number of fields, and especially in biology. A common feature in the collected sample of curves or images is to have a similar shape, modulo variance type of shifting and scaling. Vertical shift and horizontal shift are two examples of directions of variation, or modes of variation, in functional data, and they are of interest in many biological applications. Non-linear modes, such as the vertical shift, are challenging to characterize using usual statistical methods. The method we present in this dissertation, Template Modes of Variation (TMV), allows the analysis of variation of any set of curves, or images, of common shape. In this analysis, we decompose the variation in the data into modes of interest, linear or non-linear. We also quantify each mode by taking into consideration the curved geometry of the space of variation of interest. We present in this Chapter, an example of biological curves, representative of a large class of data collected in biology. This example is an important expository thread in this dissertation. We use it to illustrate functional data of common shape, to motivate our analysis and to show the results of our methodology. In this chapter, some examples of functional data are given in Section 1.1. We discuss the common template shape feature in the data and its implication in modelling in Section 1.2. We show in Section 1.3 the difficulties that a non-linear mode, such

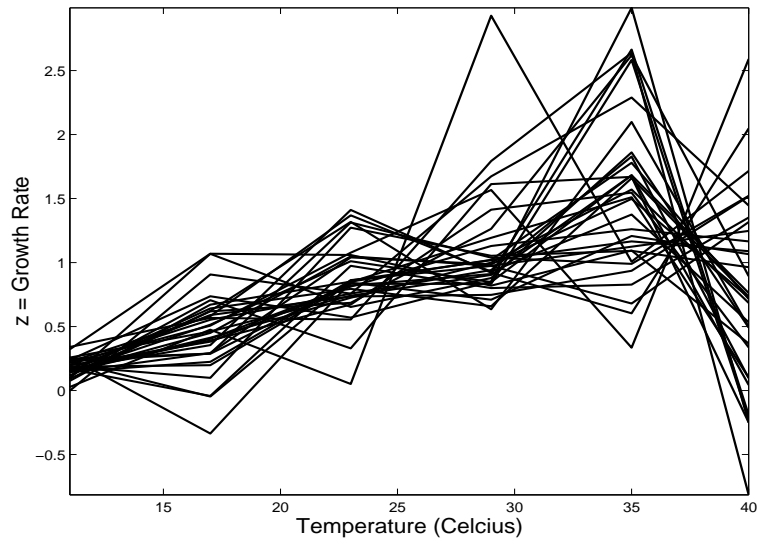


Figure 1.1: Caterpillars data: growth rate  $z$  as a function of temperature  $t$  for 32 families.

as the horizontal shift, poses in usual statistical methods. To overcome these shortcoming, we introduce our new methodology in Section 1.4 and refer to other chapters in this dissertation.

## 1.1 Functional data

Samples of images, or curves, are a common type of biological or physical data sets. Some examples of curves, presented in [Ramsay and Silverman (1997)], are curves of height over time for different boys, curves of temperature over time at different weather stations, and handwriting of a word by different individuals. Other examples of curves in biology are Continuous Reaction Norms(CRN)s, such as growth rate as a function of temperature for different caterpillar's families, shown in Figure 1.1, velocity as a function of temperature for different wasps families, [Gilchrist (1996)], wheel running distance as a function of age for different mice, [Kingsolver et al. (2001)], and weight as a function of age for beef cattle, see [Meyer (2001)].

Figure 1.1 represents curves of growth rate at different temperatures for different caterpillar families <sup>1</sup>. In the caterpillar data, the growth rate  $z$  was measured for each individual at 6 distinct temperatures  $t_1, \dots, t_6$  ranging from  $11^\circ C$  to  $40^\circ C$ . In Figure 1.1, each curve represents the mean growth rate for a family at different temperatures, and the set of 32 curves represents the growth rate curves for a population of 32 different families.

Even though the measured data is discrete of finite dimension i.e  $d = 6$ , the relationship between the growth rate and the temperature is considered continuous by biologists. In all the examples cited in this section, the underlying process generating the curves is continuous or smooth, and this property is common in curves or images data sets. Because of this continuity property, these data sets are also referred to as *functional data*. In the longitudinal data approach, see [Diggle et al. (1994)], or multivariate analysis approach, see [Anderson (1971)], the underlying smoothness of these data sets is not taken into account. The statistical entities of interest in these approaches are the finite dimensional data  $d$ -vectors, which would be for the caterpillar example 32 finite dimensional 6-vectors of growth rate. In contrast, the Functional Data Analysis (FDA) framework, described in [Ramsay and Silverman (1997)], takes advantage of the underlying smoothness of the data. It considers the continuous functions, rather than the corresponding discrete measurements, as the statistical entities of interest. In this framework, the set of 32 curves in Figure 1.1 is considered as a set of discretized measurements (with errors) of underlying curves  $f_1, \dots, f_{32}$ . The model is written as

$$z_{i,j} = f_i(t_j) + \epsilon_{i,j},$$

where  $z_{i,j}$  is the growth rate of  $i$ th family at the  $j$ th temperature,  $f_i$  is the underlying

---

<sup>1</sup>A family here means individuals of similar genotypes. In this example it is a set of offsprings with same parents

continuous curve for family  $i$ , and  $\epsilon_{i,j}$  is the additive error. In the FDA framework, the statistical entities of interest are the  $f_i$ 's.

## 1.2 Template shape and variation

We see in Figures 1.1, and 1.2 an important feature of the caterpillar data, the curves have a common shape, each curve increases slowly, tends to reach one maximum, and decreases rapidly. A similar shape is common in functional data sets, it reflects a common underlying biological or physical process in the population. For example, a similar shape in the height curves, in [Ramsay and Silverman (1997)], reflects a common pattern of growth in boys, and a similar shape in the temperature curves reflects a common seasonal pattern in weather stations. Similarly, a similar shape in the caterpillar data reflects a common environmental effect of temperature on the growth rate. As in the caterpillar data, we find, in general, a common shape feature in Continuous Reactions Norms in biology. So, the assumption of a common shape curve  $f$  is a realistic assumption in TMV. This assumption is helpful in modelling also, since we can think of the variation in functional data as variation around this template shape, in several possible directions.

Understanding the variation in the curves gives an insight on the underlying patterns of variability in the population. For example, the variation in the boys' height curves shows the population variability in growth patterns. The variation in the weather data shows the spatial variability of seasonal variations. Similarly, the family variation in the caterpillar curves shows the population variability of the common temperature effect pattern.

Sometimes scientists are interested in particular directions, or modes, of variation in the data. In our example, biologists identified different genetic variations or genetic-tradeoffs of

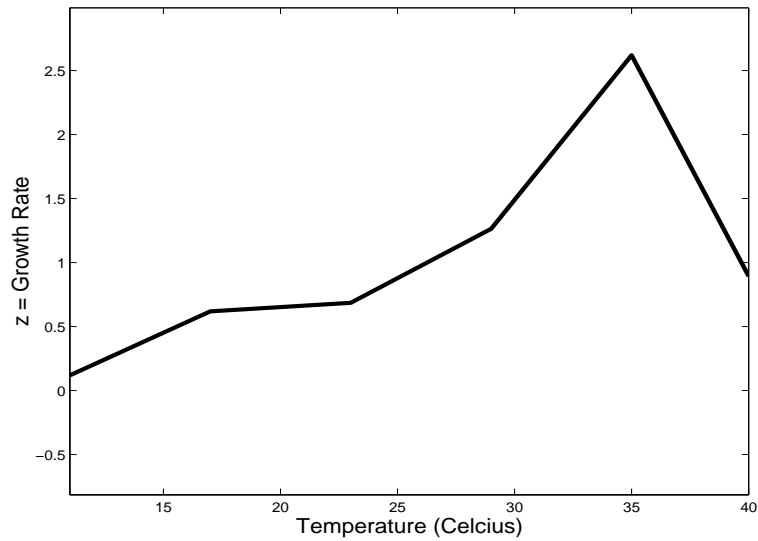


Figure 1.2: Growth rate  $z$  as a function of temperature  $t$  for a family.

interest, see [Kingsolver et al. (2001)]. In the analysis of variation that we present in this dissertation, the variation in the curves is decomposed into predetermined modes of interest. The focus in this dissertation is not on the identification of the common shape, but on the analysis of variation of the curves.

### 1.3 Modes of variation, linear versus nonlinear

Multiple FDA methods, such as functional Principal Component Analysis (PCA), analyze modes of variation in the functional data. However, these methods are not effective in analyzing *non-linear* modes. For example, two particular modes of variations of interest to evolutionary biologists in CRNs are the vertical shift of curves, shown in Figure 1.3, and horizontal shift of curves, shown in Figure 1.4. The horizontal shift of curves is also called *registration* in the FDA literature and is often observed in functional data, see [Ramsay and Silverman (1997)].

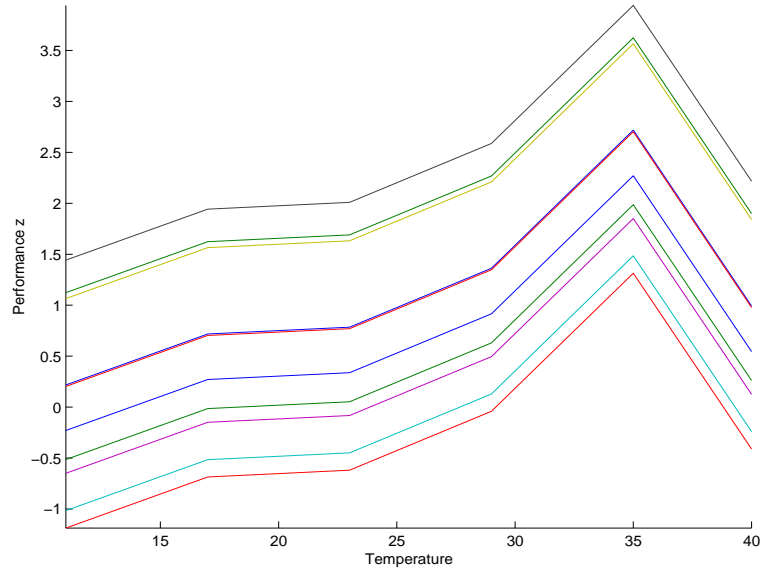


Figure 1.3: Vertical Shift of curves.

The vertical shift is a **linear** variation expressed as

$$f_i(t) = f(t) + h_i \quad (1.1)$$

where  $h_i$  is the difference in height of the  $i$ th curve from the template shape  $f$ . The horizontal shift is a **nonlinear** variation expressed as

$$f_i(t) = f(t - m_i) \quad (1.2)$$

where  $m_i$  is the location displacement between the  $i$ th curve and the template shape  $f$ .

To see the limitations of PCA with non-linear modes, we contrast the results of a PCA on data with a linear variation, the vertical shift, to the results of a PCA on data with a non-linear variation, the horizontal shift. An application of PCA to those two variations shows that the first principal component direction completely captures the vertical shift but fails to

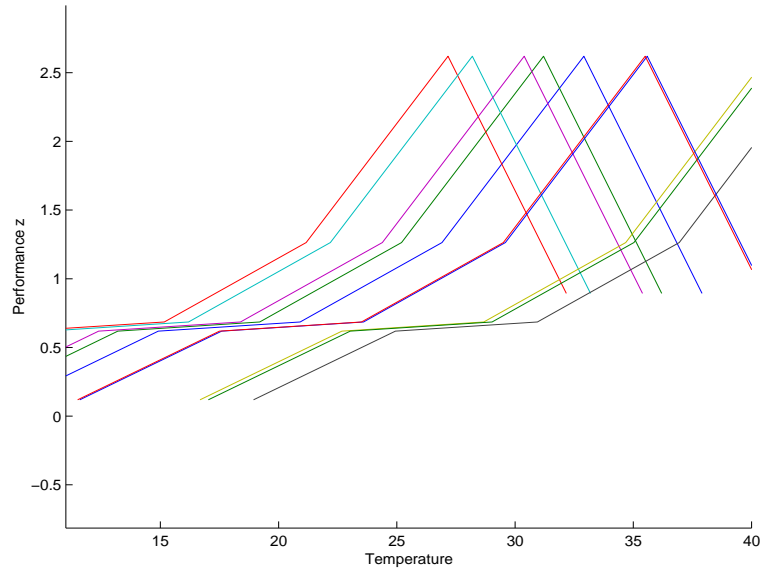


Figure 1.4: Horizontal Shift of curves.

capture most of the horizontal shift. Since the vertical shift variation is linear, the mean curve is a good estimate of the common shape or template shape as shown on the right quadrant of Figure 1.5, and the variation is easily identified by the first principal component direction PC1 as shown in Figure 1.6. However, since the horizontal shift variation is nonlinear, the mean curve is a poor estimate of the common shape as shown on the right quadrant of Figure 1.7 and the variation is not identified by PCA as shown in Figure 1.8. In our method, we do not use the sample mean curve as a measure of center of variation, but we rather use a curve which falls in the *middle* of the space of variation. We also define a new Ratio of Sum of Squares ( $\widetilde{RSS}$ ), which generalizes the linear Ratio of Sum of Squares that quantifies linear modes, to the quantification of non-linear modes. This new ratio takes into account the curved geometry of the space of variation.

When there is more than one mode of variation of interest, for example simultaneous vertical shift and horizontal shift, PCA also fails to decompose the variation into these two

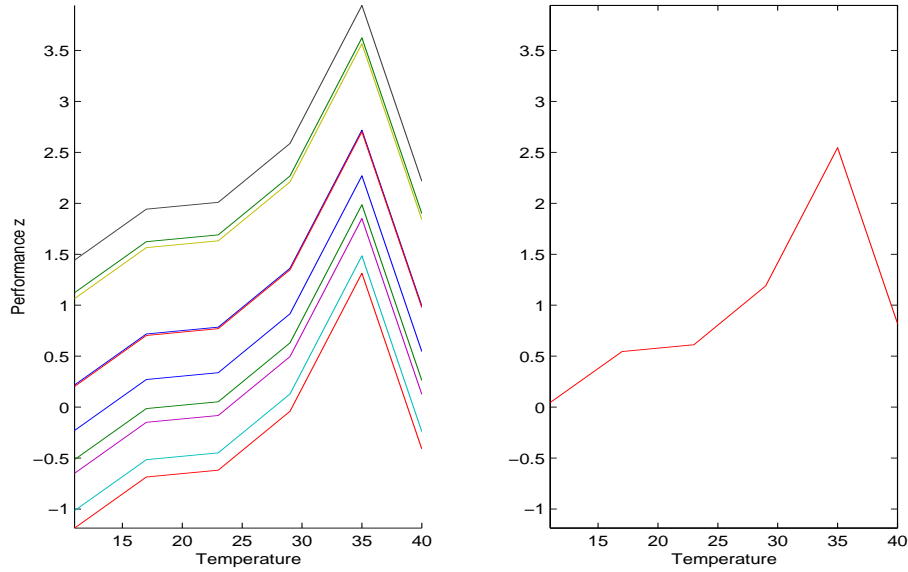


Figure 1.5: *Left* : Curves vertically shifted. *Right* : Mean curve. The mean curves captures the common shape.

modes of interest. By considering the curvature of the space of variation and defining a new metric in it, we decompose the total variation in the curves into these modes of interest.

We view these two variations of the curves as examples of a general model where  $f_i$  is a parametric transformation of the common shape  $f$ , where the parameters  $\theta$  reflect the modes of variations of interest. This feature is summarized in the following shape invariant model

$$f_i = R(\theta_i, f)$$

where  $f$  is the common shape,  $\theta_i$  is the parameter of variation for family  $i$ , and  $f_i$  is the curve for family  $i$ . When there is only one mode of variation  $\theta_i$  is a scalar, when there are more than one simultaneous variations,  $\theta_i$  is a vector.

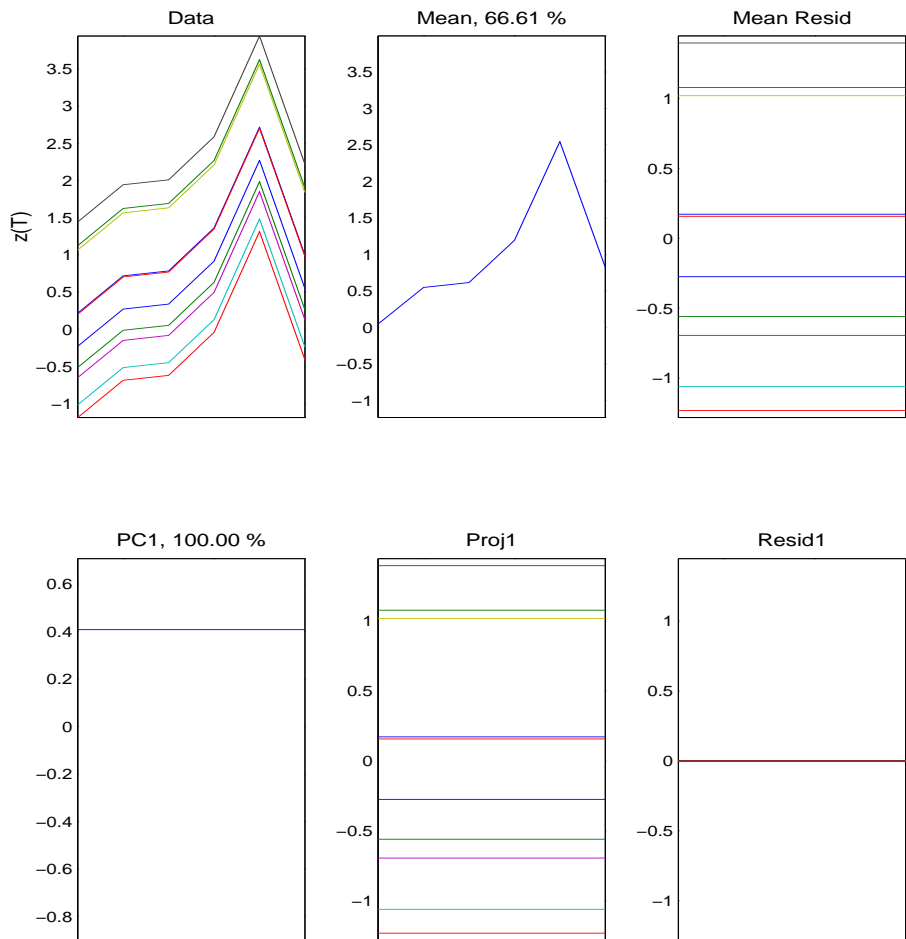


Figure 1.6: Results of PCA on the curves with vertical shift. Because this variation is linear and one dimensional, we can see that the first principal component explains 100% of this variation. This principal component of coordinates  $(1, 1, \dots, 1)$  is the vector that represents the vertical shift.

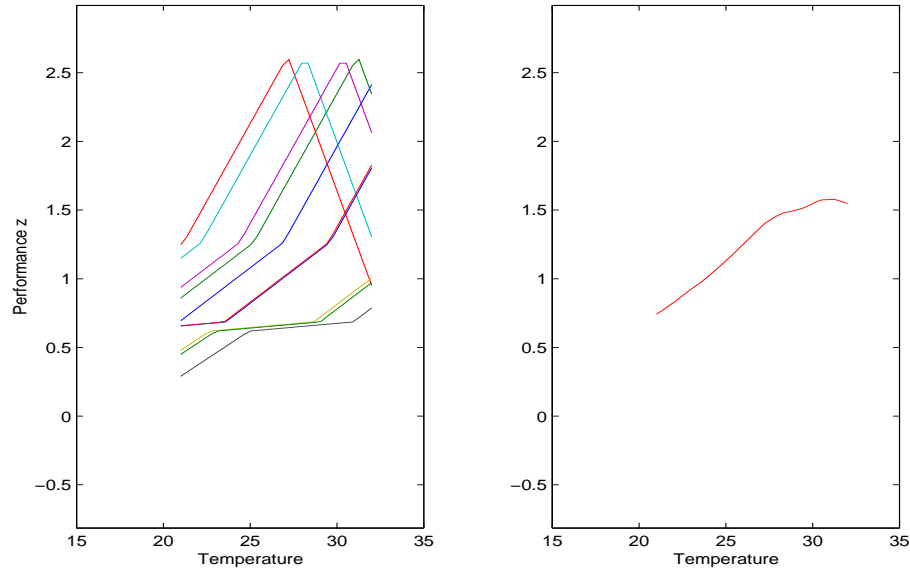


Figure 1.7: *left* : Curves horizontally shifted. The data have been truncated so that the displayed part is the common range of curves. *Right* : Mean curve. The mean curve does not capture the common shape.

## 1.4 Analysis of nonlinear variation, statistics in manifolds

We further contrast our method to other linear and nonlinear methods in FDA and image analysis or shape statistics in Chapter 2. We successfully apply our method to two different examples of CRNs, and discuss the important implications of our decomposition and quantification in Biology. We illustrate in Chapter 4, with a toy data set, the geometry of the space of variation when there are non-linear modes. We also intuitively define and justify our metric and decomposition in this chapter. We show that when the data varies along a linear mode, such as the vertical shift, the space of variation is a line. However, when the data varies along a nonlinear mode, such as the horizontal shift, the space of variation is a curve or a one dimensional nonlinear manifold. As the number of modes of variation of the curves increases, the dimensionality of the manifold of variation increases. Because of the nonlinearity of some modes of interest in the caterpillar data, the corresponding space

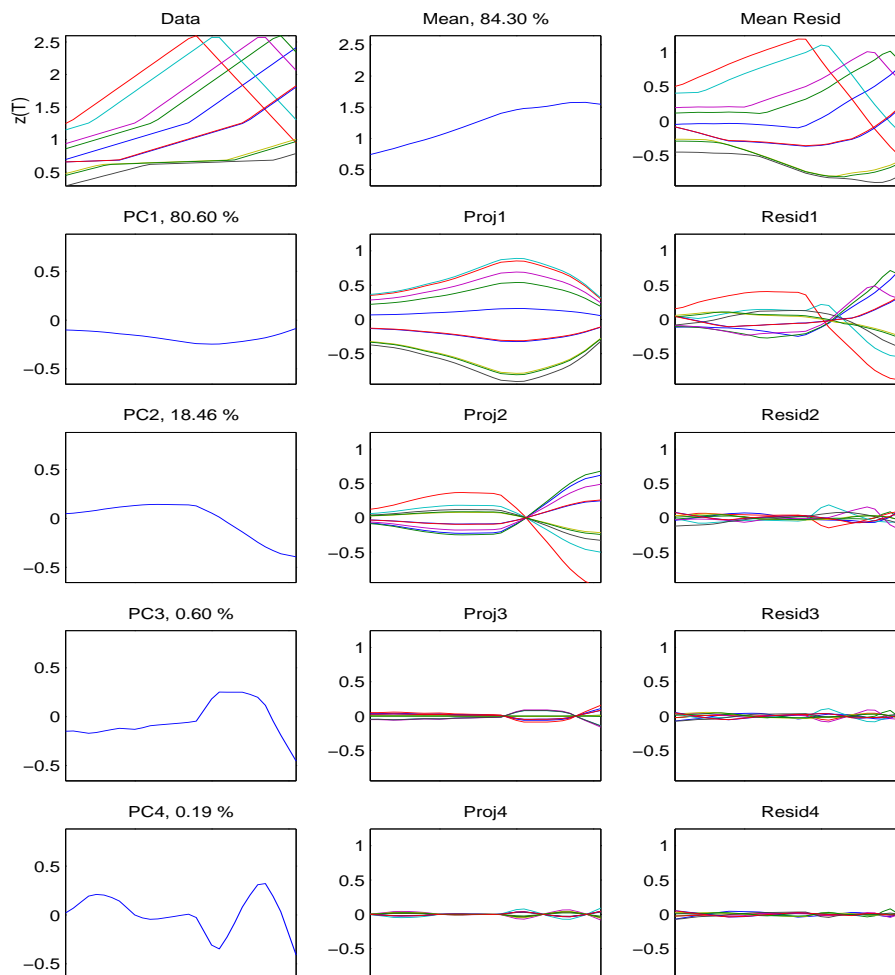


Figure 1.8: Results of PCA on the curves with horizontal shift. We have four principal components directions, the first two explain 98% of the variation in the data.

of variation is not a linear space but rather a curved space or a *manifold*. Using differential geometry and shape statistics tools, we formally define in Chapter 5, the decomposition and quantification in a general model. We find that the problem of decomposing modes of variations of the curves, and quantifying each mode, becomes a problem of defining appropriate metrics in the manifold. Such metrics would take into account the geometry of the space of variation. Proofs of all these results, on the toy example, and for the general model, as well as the description of the algorithm applied to the data sets are found in the Appendix.