

Quantifying Nonlinear Modes of Variation of Curves

R. Izem

*University of North Carolina, Chapel Hill, Department of Statistics.
Chapel Hill NC 27599-3260, USA
rizem@email.unc.edu*

J.S. Marron

*University of North Carolina, Chapel Hill, Department of Statistics.
Chapel Hill NC 27599-3260, USA
marron@email.unc.edu*

The goal of this paper is to quantify specified modes of variation of a set of curves with common shape. If a mode of variation of interest is linear, it is classically quantified by a ratio of sums of squares. However, this ratio is not an accurate measure of variation for nonlinear modes. A new measure of variation is proposed to quantify linear and nonlinear modes of variation. A toy parabola example is used to illustrate the quantification of a linear mode of variation: vertical shift of curves, and a nonlinear mode of variation: horizontal shift of curves.

Curves with common shape

A set of curves of common shape is a common type of physical or biological functional data set. Examples are curves of growth rate over time for different individuals, handwriting of a word for different individuals, and curves of temperatures over time for different weather stations [Ramsay and Silverman (1997)]. The data that motivated the following analysis are continuous reaction norm curves in evolutionary biology measuring the change of a physical characteristic over a continuous aspect of the environment for different genotypes [Kingsolver J.G., Gomulkiewicz R. , Carter P.A. (2001)]. In all the above examples, each curve is a variation of a common shape curve. The interest in this paper is to quantify specified modes of variations of the curves. The measure of quantification will be illustrated for a linear mode: the vertical shift, and a nonlinear mode: the horizontal shift. These two particular modes of variations are of biological interest to evolutionary biologists because they correspond to two different directions of evolution of the genetic variation in the population under selection [Kingsolver J.G., Gomulkiewicz R. , Carter P.A. (2001)]. The horizontal shift is also of interest in registration of curves problems[Ramsay and Silverman (1997)]. The approach proposed in this paper is different from a principal component analysis (PCA) of the data. PCA finds linear variations of the data which are data driven, and it is often hard for the biologist to find a biological interpretation of a particular mode of variation. In the following analysis, the specified modes of variation are the modes of interest and they can be nonlinear. It is assumed that the common curve shape $f(t)$ is known or can be estimated from the data.

Toy example

The template curve $f(t)$ is a downward parabolic curve, it is sampled at three distinct points (t_1, t_2, t_3) . As illustrated in Figure 1, sampling at three points will allow a representation of a curve as a point in three dimensions. Let $z_{i,j}$ be our observed value for individual i at t_j . The first set of curves is modelled as a curve shape $f(t)$, a vertical mode of variation parameterized by h , and additive noise denoted $\epsilon_{i,j}$

$$z_{i,j} = f(t_j) + h_i + \epsilon_{i,j}, \text{ for } 1 \leq i \leq n \text{ and } 1 \leq j \leq 3$$

Similarly, the second set of curves is modelled by a common curve shape $f(t)$, a horizontal

mode of variation parameterized by m , and additive noise denoted $\epsilon_{i,j}$

$$z_{i,j} = f(t_j - m_i) + \epsilon_{i,j}, \text{ for } 1 \leq i \leq n \text{ and } 1 \leq j \leq 3$$

Those two models are particular examples of the following general model

$$(1) \quad z_{i,j} = R(\theta_i, t_j) + \epsilon_{i,j}, \text{ for } 1 \leq i \leq n \text{ and } 1 \leq j \leq d$$

where θ is the parameter of variation, $R(\theta, t)$ is the regression function which is sampled at d points (t_1, \dots, t_d) , and R is a known function.

Linear and nonlinear modes

As shown in the two left plots of Figure 2 when a vertically shifted family of parabolas is represented as a point cloud, all the points lie in a line. Thus, this variation is linear and one dimensional. This property holds more generally for all template shapes and all dimensions d . When the mode of variation is a horizontal shift of curves, as shown in the two right plots of Figure 2, we see that the data points fall in a parabolic curve. More generally, for a set of curves with any one dimensional nonlinear variation and any common template shape in dimension d , the points will fall in a one dimensional manifold or curve. However the shape of the one dimensional manifold will depend on the template shape. This paper shows how to properly account for this curvature in the definition of the center and the spread of the projections and the quantification of nonlinear variations.

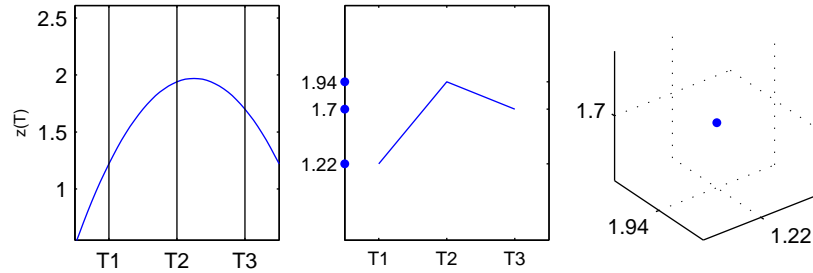


Figure 1. (1) Toy parabola shown in the curve space. (2) Toy parabola reduced to $(f(t_1), f(t_2), f(t_3))$. (3) Toy parabola shown in the point cloud space.

Quantification of linear and nonlinear variations

A linear mode of variation is quantified by the ratio sum of squares RSS

$$RSS = \frac{SSM}{SST}, \text{ where } SST = SSM + SSE$$

SSM is the sum of squares explained by the model and SSE is the sum of squared errors. In particular, for the vertical shift.

$$SSM = d \sum_i (h_i - \bar{h})^2, \text{ and } SSE = \sum_{i,j} (z_{i,j} - f(t_j) - h_i)^2$$

SSM quantifies the variations of the projections along the line. This variation is standardized in RSS by SST to make the quantifications of modes comparable to each other. Using the notation in Equation 1, SSM and SSE are defined in general by

$$SSM = \sum_{i,j} (R(\theta_i, t_j) - \bar{z}_{.j})^2, \text{ and } SSE = \sum_{i,j} (z_{i,j} - R(\theta_i, t_j))^2$$

SSM is a meaningful measure of variation for a linear mode but the toy example shows that SSM is not an accurate measure of variation for a nonlinear mode. As shown in Figure 3, in the nonlinear case the sample mean of the data is not at the center of the projections along the manifold. The geodesic sample mean of the projections \tilde{R} , or the sample mean of the projections along the manifold, is a better characterization of the center. As shown in Figure 4, a better measure of spread for a nonlinear mode should use the geometric distance d_g which is the distance along the manifold, instead of the Euclidean distance. By using \tilde{R} and d_g we propose to quantify a nonlinear mode of variation by \widetilde{RSS} defined as,

$$(2) \quad \widetilde{RSS} = \frac{\widetilde{SSM}}{\widetilde{SST}}$$

$$(3) \quad \widetilde{SST} = \widetilde{SSM} + SSE, \text{ where } \widetilde{SSM} = \sum_i d_g(R(\theta_i, \underline{t}) - \tilde{R})^2$$

When the mode of variation is linear, \widetilde{RSS} equals the classical RSS . \widetilde{RSS} is a standardized measure of the spread along a line for a linear case and along the manifold in the nonlinear case.

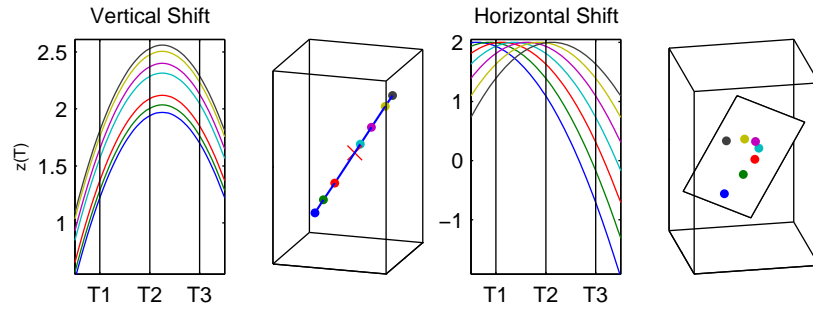


Figure 2. (1) *Parabolas vertically shifted.* (2) *Visualization of the variation in 3d.* (3) *Parabolas horizontally shifted.* (4) *Visualization of the variation in 3d*

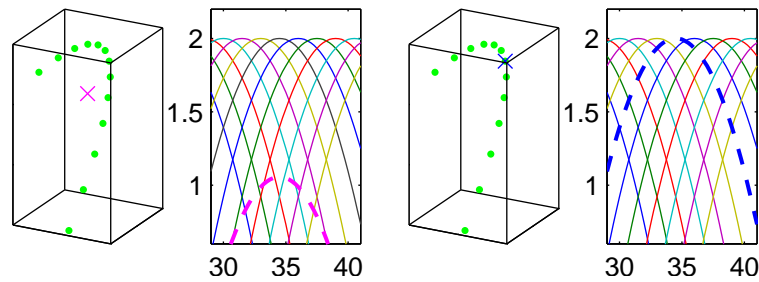


Figure 3. (1) *The data (.) and the pointwise mean (X).* (2) *Curve representation of the data (solid line) and the pointwise mean (dashed line).* (3) *The data (.) and the proposed mean (X).* (4) *The curve representation of the projections (solid line) and the geodesic mean (dashed line) of the data with respect to the horizontal shift.*

Discussion

\widetilde{RSS} as defined in Equation 2 quantifies any one parameter mode of variation of any template shape, it is more meaningful than RSS to quantify variations. To use this measure, it

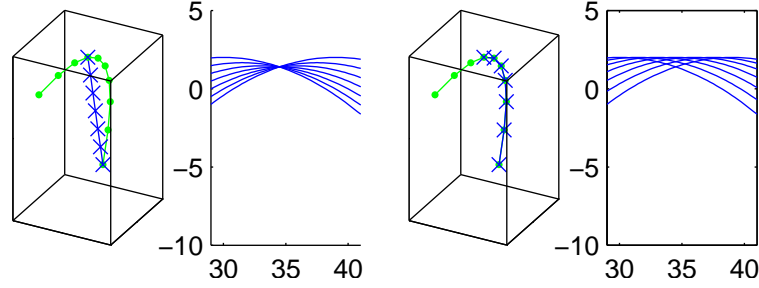


Figure 4. (1) The projections (.) and points (X) in the Euclidean distance line. (2) Curve representation of points in the line. As we march along the line, the width, height and location of the maximum vary. (3) The projections (.) and points (X) along the curve. (4) Curve representation of the points in the curve, the location of the maximum is the only change.

is necessary first to find the geodesic sample mean \tilde{R} or the mean along the manifold of variation and second to compute the geometric distance or distance along the manifold between a point in the manifold and the geodesic sample mean $d_g(R(\theta_i, t), \tilde{R})$. The geodesic sample mean \tilde{R} is the point in the manifold such that

$$\tilde{R} = \underset{x}{\operatorname{Argmin}} \sum_{i=1}^n (d_g(R(\theta_i, t), x))^2$$

which is approximated by a minimization algorithm. The geometric distance admits simple discrete approximation. Results of this method on thermal performance curves in evolutionary biology, not yet published, could be found in <http://www.unc.edu/~rizem>.

\widetilde{RSS} quantifies a one parameter mode of variation at a time. So, if two modes of variations of the curves were of interest the measure \widetilde{RSS}_1 for the first mode and \widetilde{RSS}_2 for the second mode would not be comparable. The reason is that first the sum of SSM and SEE would be different for the two modes and that secondly the two variations might not be orthogonal. A method of simultaneous quantification of the modes of variation is currently being developed.

REFERENCES

Kingsolver J.G., Gomulkiewicz R. , and Carter P.A. (2001) Variation, selection and evolution of function-valued traits. *Genetica*, **112-113**, 87-104.

Ramsay J. O. and Silverman B. W. *Functional data analysis*, Springer: Springer Series in Statistics.

RÉSUMÉ

On propose dans cet article une mesure pour quantifier un mode de variation linéaire ou nonlinéaire d'un ensemble de courbes de forme commune. La mesure classique pour quantifier un mode de variation linéaire est le quotient de sommes de carrés dénoté par RSS . Si le mode de variations des courbes est nonlinéaire, le quotient RSS perd sa signification comme mesure de variation. Une nouvelle mesure pour quantifier un mode de variation linéaire ou nonlinéaire est proposée et dénotée par \widetilde{RSS} . On utilise dans cet article un exemple de courbes paraboliques pour illustrer la quantification d'un mode de variation linéaire: translation verticale des courbes, et d'un mode de variation nonlinéaire, translation horizontale des courbes.