

## RESEARCH STATEMENT

Characterizing nonlinear variation in high dimensional data is a challenging problem and it gives the opportunity for statisticians to develop new methodologies to address it. A set of curves of common shape is an example of high dimensional data and is a common type of physical or biological functional data set. Examples are curves of growth rate over time for different individuals, different handwriting of the same word by an individual, and curves of temperatures over time for different weather stations [5]. My dissertation work was motivated by an analysis of variation in a set of curves in evolutionary biology: growth rate curves for different caterpillar families [2]. Understanding the structure of the genetic variation in these curves allows to infer the evolution of population characteristics' from one generation to the next. In particular, one linear and two nonlinear genetic modes of variation were of interest to biologists because they would induce different evolutionary responses. Since existing methodologies fail to address these interests, I developed a new method to decompose and quantify predetermined linear and nonlinear modes of variation in a set of curves of common shape. This method takes into account the geometry of the manifold, induced by the modes of variation, to define an appropriate Ratio of Sums of Squares (*RSS*) to quantify each mode [1]. To conduct this multidisciplinary research, I was funded for two years by a bioinformatics fellowship. In my dissertation, I collaborated with Biology Professor J.G Kingsolver to understand the data and the biological problems at stake. Under the supervision of Professor J.S Marron, I then developed a statistical methodology to address these biological questions and successfully applied it to the data. Whereas the motivation and one application of the methodology is in biology, the theoretical framework of the problem is in functional data analysis and differential geometry and the resulting methodology has broad applications. The differential geometry framework is often used in shape analysis problems to address nonlinear spaces of variation for reduction of dimensionality, discrimination or classification.

**Functional data, finite versus infinite dimensional** The functional data I analyze in my dissertation is a set of thermal performance curves showing the change of growth rate of caterpillars as a function of temperature. This data is a particular example of *continuous reaction norm curves* in evolutionary biology. Although the growth rate of each caterpillar was measured at only six different temperatures, the underlying relationship between the growth rate and the temperature is considered continuous (i.e infinite dimensional) and could be represented by a continuous curve. In the functional data analysis approach that I take in my dissertation, the statistical entities are the underlying growth rate curves rather than the discrete measurements. Because the set of curves are variations of a common shape curve, those variations could be modelled using a Shape Invariant Model (SIM)[3].

**Linear and nonlinear modes of variation** Three modes of variation corresponding to three different tradeoffs were of particular interest to evolutionary biologists because each mode would induce a different evolutionary response of the population under selection [2]. The first mode is vertical shift variation of curves or the *faster-slower* tradeoff. The second mode is the horizontal shift mode of variation of curves or the *hotter-colder* tradeoff. The horizontal shift is also of interest in functional data in registration of curves problems [5]. The third mode is the thinning and widening mode of variation of curves or the *generalist-specialist* tradeoff. While the vertical shift mode is linear, the horizontal shift and generalist-specialist modes are nonlinear. The classical method to analyze variation in a set of curves is Principal Component Analysis (PCA). However, PCA finds *linear* modes of variation that are data driven and it is often hard to find a biological interpretation for each mode. Nonlinear methods of analyzing the variance such as principal curves find also direction that are data driven and do not take into account the geometry of the space of variation to quantify each mode. In the method I present in my dissertation, Template Mode of Variation (TMV), the modes of variation are predetermined modes of interest which could be nonlinear. TMV decomposes the variation in a set of curves into predetermined modes assuming that the common shape is known or can be estimated from the data. Moreover, TMV quantifies each mode of interest by an appropriate (*RSS*) taking into account the geometry of the manifold of variation. When the space is linear and the modes of variations are orthogonal, the *RSS* in TMV is equivalent to the *RSS* defined in PCA.

**Quantifying the variation** Because some modes of variation of interest are nonlinear, the space of variation of the data is a nonlinear space. For the biological data in particular, I used a 3-parameter (SIM) to model

the three modes of variation of interest and I found that the space of variation is a manifold of dimension three in the larger euclidean space defined by the data. So, even if the data is high dimensional or infinite dimensional (i.e when growth rate is measured at a large or infinite number of temperatures), the effective space of variation is only of dimension three. The key idea to decompose and quantify modes of variations in a nonlinear manifold is to define the appropriate distance  $d_g$  along the manifold  $M$  that allows the decomposition of the total variation and quantification of each mode of interest. This distance is the arc distance when the manifold is one dimensional (curve or a line)[1]. Once this distance is defined, the center (resp. the spread) of variation is the Fréchet mean (resp. Fréchet variance)[4] defined for the metric manifold  $(M, d_g)$  with respect to a measure along the manifold. In practise, the Fréchet mean and variance are estimated from the data, and the  $RSS$  quantifying a mode is the ratio of the variance along this mode on the total variation in the data. In the last chapter of my dissertation, I define an appropriate  $d_g$  and discuss existence and uniqueness of the Fréchet mean, consistency and statistical properties of its estimates from the data under certain constraints on the common shape, the specified modes of variation and the distribution of the parameters of variation.

**Results and implications in evolutionary biology** I found from analyzing the data that the 3-parameter model explains most of the variation in the data (75% of the total variation in the data) which is very large considering that the three modes of variation have a biological interpretation. From the total variation, only about 7% is explained by the faster-slower tradeoff or vertical modes of variation of curves. This result was surprising to biologists and meant that selecting individuals with best performance in a certain range of environments would not result in the next generation on individuals with best performance over all environments. I am taking part in a further investigation in evolutionary biology to check if the low variation along the faster-slower mode is a common trend in continuous reaction norm curves.

**Future work** An immediate goal is to extend the theoretical findings for a broad range of modes of variation and to test sensitivity of the decomposition to the choice of the template shape. After finishing my dissertation I plan to submit the theoretical results of this work for publication and to make the matlab code for this methodology freely available for biologists and statisticians. I plan to continue my collaboration with evolutionary biologists, but as this method has potentially much broader application, I plan also to apply it to data in other fields such as medical imaging. I have a general interest in high dimensional data such as curves, or images, and in developing methods to understand the variation in this data either for reduction of dimensionality, discrimination or classification purposes. In addition to the theoretical strength in mathematics and statistics I acquired in the Department of Statistics and Operations Research in the University of North Carolina at Chapel Hill, I designed my curriculum to be exposed to statistical methods applied in biology and genetics. For instance, I took seminars in protein folding in the Department of Operations Research, statistical methods applied to phylogenetic reconstruction in the Department of Biostatistics, and analysis of gene expression array at the Institute of Statistics and Decision Sciences at Duke university. I also attended workshops and seminars in bioinformatics. Reduction of dimensionality as well as discrimination and classification methods are of high importance in those fields. During my collaboration with biologists I realized also how important it is to develop user friendly software along with creative visual display of the results to explain the method and the results. I also have an interest in improving graphical display of existing methods and creating graphical representations of new statistical methods. In my internship at the National Academies of Sciences I was introduced to policy and its interactions with science. Policy decisions are often based on data analysis and affect our everyday lives as citizens and as researchers. I have an interest in participating as a researcher in projects relating to policy.

## References

- [1] Izem R., Marron J.S (2003). Quantifying nonlinear modes of variation of curves. *Proceeding of the International Statistical Institute 54th session*
- [2] Kingsolver J.G., Gomulkiewicz R. , Carter P.A. (2001). Variation, selection and evolution of function-valued traits. *Genetica*, **112-113**, 87-104.
- [3] Lawton, W.H, Sylvestre, E.A and Maggo, M.S (1972) Self modeling regression. *Technometrics*, **14**, 513-532.
- [4] Le, H.L. Kume, A. (2000) Estimating Frechet means in Bookstein's shape space, *Advances in applied probability*, **32**, 101-113.
- [5] Ramsay J. O. , and Silverman B. W. , *Functional data analysis*, Springer: Springer Series in Statistics.