Studies in Classification, Data Analysis, and Knowledge Organization

Salvatore Ingrassia
Roberto Rocci
Maurizio Vichi *Editors*

New Perspectives in Statistical Modeling and Data Analysis

Springer

Salvatore Ingrassia • Roberto Rocci
Maurizio Vichi
Editors

New Perspectives in Statistical Modeling and Data Analysis

Proceedings of the 7th Conference
of the Classification
and Data Analysis Group
of the Italian Statistical Society,
Catania, September 9-11, 2009

Springer

Editors
Prof. Salvatore Ingrassia
Università di Catania
Dipartimento Impresa
Culture e Società
Corso Italia 55
95129 Catania
Italy
s.ingrassia@unict.it

Prof. Roberto Rocci
Università di Roma "Tor Vergata"
Dipartimento SEFEMEQ
Via Columbia 2
00133 Roma
Italy
roberto.rocci@uniroma2.it

Prof. Maurizio Vichi
Università di Roma "La Sapienza"
Dipartimento di Statistica
Probabilità e Statistiche Applicate
Piazzale Aldo Moro 5
00185 Roma
Italy
maurizio.vichi@uniroma1.it

ISSN 1431-8814
ISBN 978-3-642-11362-8 e-ISBN 978-3-642-11363-5
DOI 10.1007/978-3-642-11363-5
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011930788

Cover design: deblik, Berlin

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Dealing with FDA Estimation Methods

Tonio Di Battista, Stefano A. Gattone, and Angela De Sanctis

Abstract In many different research fields, such as medicine, physics, economics, etc., the evaluation of real phenomena observed at each statistical unit is described by a curve or an assigned function. In this framework, a suitable statistical approach is Functional Data Analysis based on the use of basis functions. An alternative method, using Functional Analysis tools, is considered in order to estimate functional statistics. Assuming a parametric family of functional data, the problem of computing summary statistics of the same parametric form when the set of all functions having that parametric form does not constitute a linear space is investigated. The central idea is to make statistics on the parameters instead of on the functions themselves.

1 Introduction

Recently, Functional data analysis (FDA) has become an interesting research topic for statisticians. See for example Ferraty and Vieu (2006) and Ramsay and Silverman (2007) and reference therein. In many different fields, data come to us through a process or a model defined by a curve or a function. For example, in psychophisiological research, in order to study the electro dermal activity of an individual, the Galvanic Skin Response (GSR signal) can be recorded and represented by a continuous trajectory which can be studied by means of the tools of FDA (Di Battista et al. 2007). We want to deal with circumstances where functional data are at hand and the function is known in its closed form. In particular, we consider a parametric family of functional data focusing on parameters estimation of the function. For example, Cobb-Douglas production functions are frequently used in economics in order to study the relationship between input factors and the level of production. This family of functions takes on the form $y = f(K, L) = L^{\alpha} K^{\beta}$, where L is one factor of production (often labour) and K is a second factor of production (often capital) and α and β are positive parameters with $\alpha + \beta = 1$. In biology, growth functions are used to describe growth processes (Vieira and Hoffmann 1977). For example, the logistic growth function $Z = a/[1 + \exp\{-(b + ct)\}]$ where a, b and c are parameters,

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_40,

$a > 0$ and $c > 0$, and the Gompertz growth function $Z = \exp(a - bc^t)$ where a, b and c are parameters, $b > 0$ and $0 < c < 1$. The aims of FDA are fundamentally the same as those of any area of statistics, i.e., to investigate essential aspects such as the mean and the variability function of the functional data. Moreover, one could be interested in studying the rate of change or derivatives of the curves. However, since functional data are often observed as a sequence of point data, then the function denoted by $y = x(t)$ reduces to a record of discrete observations that we can label by the n pairs (t_j, y_j) where y_j is the value of the function computed at the point t_j. A first step in FDA is to convert the values $y_{i1}, y_{i2}, \ldots y_{in}$ for each unit $i = 1, 2, \ldots, m$ to a functional form computable at any desired point t. To this purpose, the use of basis functions ensures a good fit in a large spectrum of cases. The statistics are simply those evaluated at the functions pointwise across replications.

It is well known that the sample mean $\bar{x}(t) = \frac{1}{m}\sum_{i=1}^{m} x_i(t)$ is a good estimate of the mean if the functional data are assumed to belong to L^2. If we do not need of a scalar product and then of an orthogonality notion, we can consider every L^p space, $p > 1$, with the usual norm (Rudin 2006). In general the functional data constitute a space which is not a linear subspace of L^p. For example, let $y_1 = A_1 L^{\alpha_1}$ and $y_2 = A_2 L^{\alpha_2}$ be Cobb-Douglas functions in which for simplicity the production factors A_1 and A_2 are assumed constant. The mean function is $\bar{y} = \frac{A_1 L^{\alpha_1} + A_2 L^{\alpha_2}}{2}$ which is not a Cobb-Douglas function and its parameter does not represent the well known labour elasticity which is crucial to evaluate the effect of labour on the production factor. In general, the results of this approach may not belong to a function with the same closed form of the converted data so that erroneous interpretations of the final functional statistic could be given.

In this communication we want to emphasize a new approach which is focused on the true functional form generating the data. First of all, we introduce a suitable interpolation method (Sung Joon 2005) that allows us to estimate the function that is suspected to produce the functional datum for each replication unit. Starting from the functional data we propose an explicit estimation method. The objective is to obtain functional statistics that belong to the family of functions or curves suspected to generate the phenomenon under study. In the case of a parametric family of functional data, we use the parameter space in order to transport the mean of the parameters to the functional space. Assuming a monotonic dependence from parameters we can obtain suitable properties for the functional mean. At illustrative purpose, two small simulation studies are presented in order to explore the behaviour of the approach proposed.

2 Orthogonal Fitting Curve and Function

Generally, functional data are recorded discretely as a vector of points for each replication unit. Thus, as a first step we need to convert the data points to a curve or a function. Methods such as OLS and/or GLS do not ensure the interpolation of a wide class of curves or functions. A more general method is given by the

Least Squares Orthogonal Distance Fitting of Curves (ODF) (Sung Joon 2005). The goal of the ODF is the determination of the model parameters which minimize the square sum of the minimum distances between the given points $\{\mathbf{Y}_j\}_{j=1}^n$ and the closed functional form belonging to the family of curves or function $\{f(\boldsymbol{\theta}, t)\}$ with $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_p\}$. In ODF the corresponding points $\left\{\mathbf{Y}_j^*\right\}_{j=1}^n$ on a fitted curve are constrained to being membership points of a curve/surface in space. So, given the explicit form $f(\boldsymbol{\theta}; t)$ such that $\mathbf{Y}^* - f(\boldsymbol{\theta}; t) = 0$, the problem leads to minimize a given cost function. Two performances indices are introduced which represent in two different ways the square sum of the weighted distances between the given points and the functional form $f(\boldsymbol{\theta}; t)$: the performances index $\sigma_0^2 = \|P(\mathbf{Y} - \mathbf{Y}^*)\|^2 = (\mathbf{Y} - \mathbf{Y}^*)^T \mathbf{P}^T \mathbf{P}(\mathbf{Y} - \mathbf{Y}^*)$ in coordinates based view or $\sigma_0^2 = \|\mathbf{Pd}\|^2 = \mathbf{d}^T \mathbf{P}^T \mathbf{Pd}$ in distance based view, where $\mathbf{P}^T\mathbf{P}$ is a weighting matrix or error covariance matrix (positive definite), $\mathbf{Y}^* = \left\{\mathbf{Y}_j^*\right\}_{j=1}^n$ is a coordinate column vector of the minimum distance points on the functional form from each given point $\{\mathbf{Y}_j\}_{j=1}^n$, $\mathbf{d} = (d_1, d_2, \ldots, d_n)^T$ is the distance column vector with $d_j = \left\|\mathbf{Y}_j - \mathbf{Y}_j^*\right\| = \sqrt{\left(\mathbf{Y}_j - \mathbf{Y}_j^*\right)^T \left(\mathbf{Y}_j - \mathbf{Y}_j^*\right)}$. Using the Gauss Newton method, it is possible to estimate the model parameters $\boldsymbol{\theta}$ and the minimum distance points $\left\{\mathbf{Y}_j^*\right\}_{j=1}^n$ with a variable separation method in a nested iteration scheme as follows

$$\min_{\theta \in R^p} \min_{\left\{\mathbf{Y}_j^*\right\}_{j=1}^n \in Z} \sigma_0^2 \left(\{\mathbf{Y}_i^* (\boldsymbol{\theta})\}_{j=1}^n\right) \tag{1}$$

whith $Z = \left\{\mathbf{Y} \in R^n : \mathbf{Y} - f(\boldsymbol{\theta}; t) = 0, \boldsymbol{\theta} \in R^p, t \in R^k\right\}$.

3 Direct FDA Estimation Methods

Let S be a family of functions with p real parameters that is $S = \{f_\theta\}$ with $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_p) \in \Theta$. In an economic setting, S could be the family of Cobb-Douglas production functions, i.e., $f_{\alpha,\beta}(K, L) = K^\alpha L^\beta$ with $\alpha > 0$, $\beta > 0$ and $\alpha + \beta = 1$. Starting from m functional data belonging to S, $f_{\theta_1}, f_{\theta_2}, \ldots, f_{\theta_m}$, the objective is to find an element of S said functional statistic denoted with $f_{\hat{\theta}} = H\left(f_{\theta_1}, f_{\theta_2}, \ldots, f_{\theta_m}\right)$.

3.1 The Functional Mean

In the following we assume that functional data constitute a subspace S of some L^p space, $p > 0$, with the usual norm (Rudin 2006). We consider first the functional mean of the functions $f_{\theta_1}, f_{\theta_2}, \ldots, f_{\theta_m}$. When S is a vectorial subspace, then we can express the functional mean as the sample mean $f_{\hat{\theta}} = \frac{f_{\theta_1} + f_{\theta_2} + \ldots + f_{\theta_m}}{m}$.

Because S is closed with respect to linear combinations, we have that $f_{\hat{\theta}} \in S$. In this setting a straightforward property is that the integral of the functional mean is the mean of the integrals of each functional datum. For example, let S be the family of functions of the following form $f_{\alpha} = \alpha g(x)$, then

$$f_{\hat{\alpha}} = \frac{\sum_{i=1}^{m} f_{\alpha_i}(x)}{m} = \frac{\sum_{i=1}^{m} \alpha_i g(x)}{m} = \frac{\sum_{i=1}^{m} \alpha_i}{m} g(x). \tag{2}$$

This proves that $f_{\hat{\alpha}}(x)$ is an element of S and its parameter is the mean of the parameters $\alpha_1, \alpha_2, \ldots, \alpha_m$. At the same time it is easy to prove that if S is not a vectorial space then this functional statistic doesn't necessarily lead to an element belonging to S. We go along in two ways. The first one is to verify if there is an element in S that has got as integral the mean of the integrals of the functional data. For instance, let S be the family of functions $f_{\alpha}(x) = x^{\alpha}$, with $0 < \alpha < 1$ and domain the closed interval $[0, 1]$. If $m = 2$, then $\int_0^1 x^{\alpha} dx = \frac{\int_0^1 x^{\alpha_1} dx + \int_0^1 x^{\alpha_2} dx}{2}$, that is $\frac{1}{\alpha+1} = \frac{\frac{1}{\alpha_1+1} + \frac{1}{\alpha_2+1}}{2}$ which admits a unique solution. For example if we have got two functions with parameters $\alpha_1 = \frac{1}{2}$ and $\alpha_2 = \frac{1}{3}$ then $\hat{\alpha} = \frac{7}{17}$. Unfortunately, in general the solution may not exist in the real field and/or it is not unique and it would be necessary to introduce some constraints on the parameters not easy to interpret.

A second way to solve the problem without ambiguity is the following. We assume that every functional datum f_{θ} is univocally determined by the parameter θ or equivalently there is a biunivocal correspondence between S and the parameter space Θ. Then, a functional statistic for the space of the functional data can be obtained through a statistic in the parameter space. In the case of a parametric family of functional data, we use the parameter space in order to transport the statistics in Θ to S. Let the functional data be $f_{\theta_1}, f_{\theta_2}, \ldots, f_{\theta_m}$, then a functional statistic for the set of the functional data is given by a suitable statistic of the parameters $\theta_1, \theta_2, \ldots, \theta_m$ say $\hat{\theta} = K(\theta_1, \theta_2, \ldots, \theta_m)$. The functional statistic will be the element of S that has got as parameter the statistic $\hat{\theta}$, following the scheme:

$$\begin{array}{ll} \theta_i & \leftarrow f_{\theta_i} \\ \downarrow & \qquad\qquad i = 1, 2, \ldots, m \\ \hat{\theta} = K(\theta_i) & \rightarrow f_{\hat{\theta}}. \end{array} \tag{3}$$

A possible way of defining the function K is the analogy criterion. If we want to estimate the functional mean or median then the function K would be the mean or the median of the parameters. Obviously, other ways of defining the function K are possible. The advantage in this case is that we can require for the functional mean and variability the same properties of the mean and variance of the parameters. In particular, for the functional mean, we can assume that the functions are linked to each parameter by a monotonic dependence. For example, if we have only a parameter α, we can suppose $\alpha_1 \leq \alpha_2 \Rightarrow f_{\alpha_1}(x) \leq f_{\alpha_2}(x)$ or $f_{\alpha_1}(x) \geq f_{\alpha_2}(x)\ \forall x$. In such a case, for the mean parameter $\hat{\alpha}$, we obtain $f_{\alpha_1}(x) \leq f_{\hat{\alpha}}(x) \leq f_{\alpha_2}(x)\ \forall x$. Moreover this property ensures also that $\int f_{\alpha_1}(x)dx \leq \int f_{\hat{\alpha}}(x)dx \leq \int f_{\alpha_2}(x)dx$. It is easy to verify that monotonic decreasing dependence is verified by the family $S = f_\alpha(x) = x^\alpha$ with $0 < \alpha < 1$ and $x \in [0, 1]$.

3.2 Functional Variability

In order to study the functional variability we first introduce the functional quantity $v_i^r(t) = \left| f_{\theta_i}(t) - f_{\hat{\theta}}(t) \right|^r$ which is the r-th order algebraic deviation between the functional observed data f_{θ_i} and the functional statistics $f_{\hat{\theta}}$. Then the functional variability can be measured pointwise by the r-th order functional moment

$$V^r(t) = \frac{1}{m} \sum_{i=1}^{m} v_i^r(t). \tag{4}$$

The function $V^r(t)$ has the following properties:

- if $f_{\theta_i}(t) = f_{\hat{\theta}}(t)$ for $i = 1, 2, \ldots, m$ and $\forall t$, than $V^r(t) = 0$;
- defining the L^p norm of a function as $\| f_\theta(t) \|_{L^p} = \int |f_\theta(t)|^p \, dt$ then we have that

$$\left\{ \| f_{\theta_i} - f_{\hat{\theta}} \|_{L^p} \to 0 \right\} \Rightarrow \left\{ f_{\theta_i} \overrightarrow{a.e.} f_{\hat{\theta}} \Leftrightarrow v_i^r \overrightarrow{a.e.} 0 \; \forall i = 1, 2, \ldots, m \Leftrightarrow V^r \overrightarrow{a.e.} 0 \right\}.$$

We remark that, if the function f_θ in S is expandable in Taylor's series, that is

$$f_\theta(t) = \sum_{k=0}^{\infty} \frac{f_\theta^k(a)}{k!} (t - a)^k \tag{5}$$

where a is a fixed point of an open domain and $f_\theta^k(a)$ is the k-th derivative of the function f_θ computed at point a, an approximation of the functional variability can be obtained by Taylor's polynomials s_{θ_i} of f_{θ_i} and $s_{\hat{\theta}_i}$ of $f_{\hat{\theta}_i}$ respectively:

$$\frac{1}{m} \sum_{i=1}^{m} \left| s_{\theta_i}(t) - s_{\hat{\theta}_i}(t) \right|^p. \tag{6}$$

This fact is useful from a computation point of view. In order to give some insights to the approach proposed in the next section two small simulation studies are proposed.

4 A Simulation Study

We conduct two small simulation studies in order to evaluate the estimation method proposed for the functional statistic $f_{\hat{\theta}} = H\left(f_{\theta_1}, f_{\theta_2}, \ldots, f_{\theta_m}\right)$ equal to the functional mean.

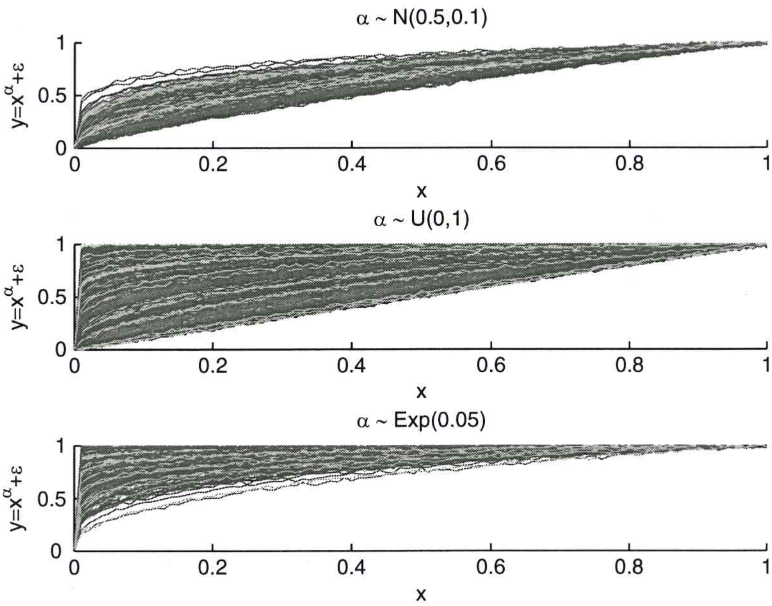

Fig. 1 Functional populations $S = \{f_\theta\} = x^\alpha + \epsilon$ with three different space parameter θ and $\epsilon \sim N(0, 0.01)$

4.1 Power Functions

We suppose that the observations are contaminated with some error so that the resulting family $S = \{f_\theta\}$ of functions is defined as $S = \{x^\alpha\}+\epsilon$ with $\theta = \alpha \in R^1$ with $0 < \alpha < 1$ and $0 \leq x \leq 1$. We simulate different populations by assigning to α different distributions such as the truncated Normal, the Uniform and the truncated Exponential with different parameters and to ϵ a white noise with standard error equal to 0.01. At illustrative purpose in Fig. 1 there are three populations for $\alpha \sim N(\mu = 0.5, \sigma = 0.1)$, $\alpha \sim U(0, 1)$ and $\alpha \sim Exp(0.05)$. Values of α outside the interval $(0, 1)$ were discarded.

In order to evaluate the estimation method proposed in Sect. 3, we sample from each population $J = 5{,}000$ samples for various sample sizes m. As the functions are observed with error we first need to apply the ODF method of Sect. 2 to estimate the function parameter α for each function. Once for each sample the estimates $\theta_1, \theta_2, \ldots, \theta_m$ are available, the scheme detailed in (3) can be applied in order to obtain the functional mean statistic of the sample. In Fig. 2 we show the results for a sample size of $m = 10$. In particular, for each population, the functional mean statistic together with the estimated standard error are plotted.

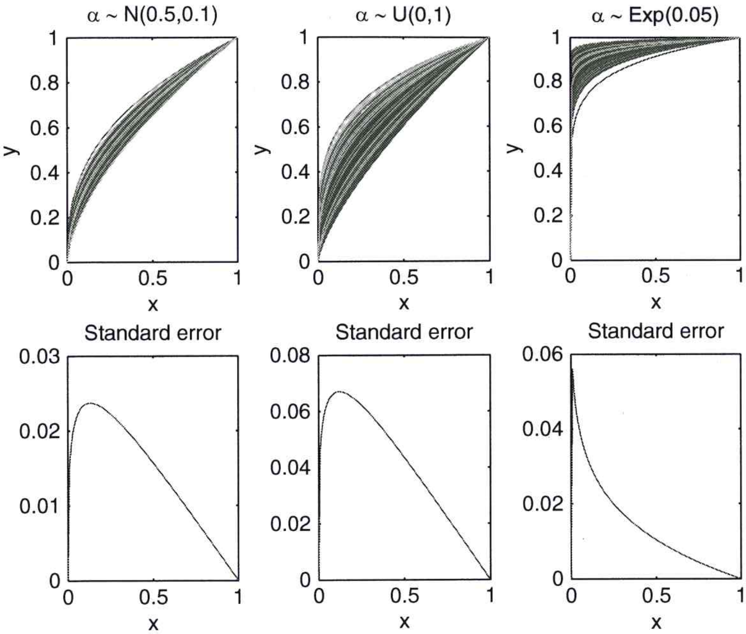

Fig. 2 J = 5,000 Functional mean statistics for a sample size $m = 10$

4.2 Functional Diversity Profiles

At illustrative purpose, we present an ecological application of the estimation method proposed. Suppose to have a biological population made up of p species where we are able to observe the relative abundance vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ in which the generic θ_j represents the relative abundance of the j-th species. One of the most remarkable aspects in environmental studies is the evaluation of ecological diversity. The most frequently used diversity indexes may be expressed as a function $f_{\boldsymbol{\theta}}$ of the relative abundance vector. Patil and Taillie (1982) proposed to measure diversity by means of the β-diversity profiles defined as

$$\Delta = f_{\boldsymbol{\theta}}(\beta) = \frac{1 - \sum_{j=1}^{p} \theta_j^{\beta+1}}{\beta}. \tag{7}$$

β-diversity profiles are non-negative and convex curves. In order to apply functional linear models on diversity profiles, Gattone and Di Battista (2009) applied a transformation which can be constrained to be non-negative and convex. In the FDA context, it is convenient considering the β-diversity profile as a parametric function computable for any desired argument value of $\beta \in [-1, 1] \setminus \{0\}$. The space parameter is multivariate and given by $\boldsymbol{\theta}$. In order to evaluate the estimation method proposed in Sect. 3, we simulate different biological populations by assigning to each component of $\boldsymbol{\theta}$ different distributions such as the Uniform, the Poisson and

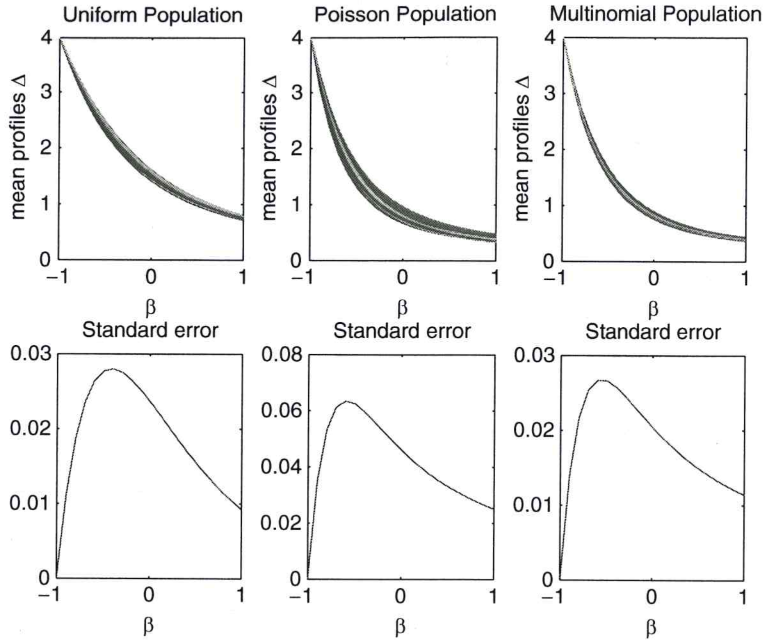

Fig. 3 J = 5,000. Functional mean diversity profiles $\Delta = f_{\hat{\theta}} = \frac{1-\sum_{j=1}^{p}\theta_j^{\beta+1}}{\beta}$ and standard error for a sample size $m = 5$

the multinomial distribution. From each population we sample $J = 5,000$ samples with different sample sizes. The function Δ in (7) is observed without error so that we do not need to apply the ODF method of Sect. 2. For each sample of size m we can evaluate the estimates $\hat{\boldsymbol{\theta}}$ from the observed $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_m$ and the scheme detailed in (3) can be applied in order to obtain the functional mean statistic $\Delta = f_{\hat{\theta}}$. In Fig. 3 we show the results for three populations with $p = 5$ species with different level of diversity. From each population we randomly choose samples of size $m = 5$. The parameters of the Poisson and the Multinomial distributions are $\lambda = 100 * [0.55, 0.19, 0.13, 0.07, 0.06]$ and $[0.55, 0.19, 0.13, 0.07, 0.06]$, respectively. For each population, the functional mean statistic together with the estimated standard error are plotted. As desired, all the functional statistics result to be non-negative and convex. Furthermore, even though monotonic dependence from the parameters is not verified with diversity profiles, the functional mean satisfies the internality property in all the simulation runs.

References

Di Battista, T. Gattone S.A., & Valentini, P. (2007). Functional Data Analysis of GSR signal, Proceedings *S.Co. 2007: Complex Models and Computational Intensive Methods for Estimation and Prediction*, CLEUP Editor, Venice, 169–174.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: Theory and practice*. New York: Springer-Verlag.

Gattone, S. A., & Di Battista, T. (2009). A functional approach to diversity profiles. *Journal of the Royal Statistical Society, Series C, 58*, 267–284.

Patil, G. P., & Taillie, C. (1982). Diversity as a concept and its measurements. *Journal of the American Statistical Association, 77*, 548–561.

Ramsay, J. O., & Silverman, B. W. (2007). *Functional data analysis*. New York: Springer.

Rudin, W. (2006). *Real and complex analysis*. McGraw-Hill, New York.

Sung Joon, A. (2005). *Least squares orthogonal distance fitting of curves and surfaces in space*. New York: Springer.

Vieira, S., & Hoffmann, R. (1977). Comparison of the logistic and the Gompertz growth functions considering additive and multiplicative error terms. *Applied Statistics, 26*, 143–148.

Studies in Classification, Data Analysis, and Knowledge Organization

Salvatore Ingrassia · Roberto Rocci · Maurizio Vichi *Editors*

New Perspectives in Statistical Modeling and Data Analysis

This volume provides recent research results in data analysis, classification and multivariate statistics and highlights perspectives for new scientific developments within these areas. Particular attention is devoted to methodological issues in clustering, statistical modeling and data mining. The volume also contains significant contributions to a wide range of applications such as finance, marketing, and social sciences. The papers in this volume were first presented at the 7th Conference of the Classification and Data Analysis Group (ClaDAG) of the Italian Statistical Society, held at the University of Catania, Italy.

Statistics
ISSN 1431-8814

ISBN 978-3-642-11362-8

9 783642 113628

▶ springer.com