# SIS | 2022

## 51st Scientific Meeting of the Italian Statistical Society

Caserta, 22-24 June

Università degli Studi della Campania *Luigi Vanvitelli*

SIS Società Italiana di Statistica

www.unicampania.it

# Book of the Short Papers

Editors: Antonio Balzanella, Matilde Bini, Carlo Cavicchia, Rosanna Verde

Matilde Bini (Chair of the Program Committee) - *Università Europea di Roma*
Rosanna Verde (Chair of the Local Organizing Committee) - *Università della Campania "Luigi Vanvitelli"*

PROGRAM COMMITTEE

Matilde Bini (Chair), Giovanna Boccuzzo, Antonio Canale, Maurizio Carpita, Carlo Cavicchia, Claudio Conversano, Fabio Crescenzi, Domenico De Stefano, Lara Fontanella, Ornella Giambalvo, Gabriella Grassia - Università degli Studi di Napoli Federico II, Tiziana Laureti, Caterina Liberati, Lucio Masserini, Cira Perna, Pier Francesco Perri, Elena Pirani, Gennaro Punzo, Emanuele Raffinetti, Matteo Ruggiero, Salvatore Strozza, Rosanna Verde, Donatella Vicari.

LOCAL ORGANIZING COMMITTEE

Rosanna Verde (Chair), Antonio Balzanella, Ida Camminatiello, Lelio Campanile, Stefania Capecchi, Andrea Diana, Michele Gallo, Giuseppe Giordano, Ferdinando Grillo, Mauro Iacono, Antonio Irpino, Rosaria Lombardo, Michele Mastroianni, Fabrizio Maturo, Fiammetta Marulli, Paolo Mazzocchi, Marco Menale, Giuseppe Pandolfi, Antonella Rocca, Elvira Romano, Biagio Simonetti.

ORGANIZERS OF SPECIALIZED, SOLICITED, AND GUEST SESSIONS

Arianna Agosto, Raffaele Argiento, Massimo Aria, Rossella Berni, Rosalia Castellano, Marta Catalano, Paola Cerchiello, Francesco Maria Chelli, Enrico Ciavolino, Pier Luigi Conti, Lisa Crosato, Marusca De Castris, Giovanni De Luca, Enrico Di Bella, Daniele Durante, Maria Rosaria Ferrante, Francesca Fortuna, Giuseppe Gabrielli, Stefania Galimberti, Francesca Giambona, Francesca Greselin, Elena Grimaccia, Raffaele Guetto, Rosalba Ignaccolo, Giovanna Jona Lasinio, Eugenio Lippiello, Rosaria Lombardo, Marica Manisera, Daniela Marella, Michelangelo Misuraca, Alessia Naccarato, Alessio Pollice, Giancarlo Ragozini, Giuseppe Luca Romagnoli, Alessandra Righi, Cecilia Tomassini, Arjuna Tuzzi, Simone Vantini, Agnese Vitali, Giorgia Zaccaria.

ADDITIONAL COLLABORATORS TO THE REVIEWING ACTIVITIES

Ilaria Lucrezia Amerise, Ilaria Benedetti, Andrea Bucci, Annalisa Busetta, Francesca Condino, Anthony Cossari, Paolo Carmelo Cozzucoli, Simone Di Zio, Paolo Giudici, Antonio Irpino, Fabrizio Maturo, Elvira Romano, Annalina Sarra, Alessandro Spelta, Manuela Stranges, Pasquale Valentini, Giorgia Zaccaria.

PUBLISHED BY PEARSON

WWW.PEARSON.COM

*ISBN 9788891932310*

# Modes of variation for Lorenz Curves

## *Modi di variazione per curve di Lorenz*

Enea G. Bongiorno and Aldo Goia

**Abstract** This work illustrates how to perform functional principal component analysis and to compute the modes of variations for a sample of Lorenz curves. In particular, to coherently implement functional principal component analysis in a proper manner, Lorenz curves are suitably transformed. The procedure is applied at the income Lorenz curves for the Italian regions in the years 2000, 2006 and 2010.
**Abstract** *Questo lavoro illustra come implementare l'analisi delle componenti principali funzionali e come calcolare i modi di variazione per un campione di curve di Lorenz. In particolare, al fine di implementare in maniera coerente l'analisi delle componenti principali funzionali, le curve di Lorenz sono trasformate opportunamente. La procedura è applicata alle curve di Lorenz del reddito per le regioni italiane negli anni 2000, 2006 e 2010.*

**Key words:** Lorenz curves, Modes of variation, income distributions

## 1 Introduction

In some applications, ranging from Economics to Biology, from Chemistry to Environmetrics, it is interesting to consider the notion of concentration, that is the attitude of a non–negative r.v. $X$ to redistribute its total mass over the individuals within the population. This concept allows to represent and distinguish situations ranging from the maximum concentration setting (when one individual holds the total mass) to the equidistribution one (when each individual hold the same mass).

Enea G. Bongiorno
Università del Piemonte Orientale, Dipartimento di Studi per l'Economia e l'Impresa, via Perrone, 18, 28100, Novara, Italia e-mail: enea.bongiorno@uniupo.it

Aldo Goia
Università del Piemonte Orientale, Dipartimento di Studi per l'Economia e l'Impresa, via Perrone, 18, 28100, Novara, Italia e-mail: aldo.goia@uniupo.it

A formal way to depict the concentration of a probability law is given by the Lorenz Curve (LC) [5] that is defined by

$$\begin{aligned} L\colon [0,1] &\to [0,1] \\ p\ \ &\mapsto L(p) = \textstyle\int_0^p Q(t)dt/\mu, \end{aligned}$$

where $\mu = \mathbb{E}[X]$, $Q(p) = \inf\{x : F(x) \geq p\}$ is the quantile function of $X$ defined for any $p \in (0,1)$ and with $F$ being the cdf of $X$. For a LC one has $L(0) = 0$, $L(1) = 1$, $L(p) \leq p$ and $L$ is continuous, increasing and convex on $[0,1]$. As an instance, consider the empirical LCs (i.e. based on the empirical versions of mean and quantile function) of household income of the 20 regions of Italy for the years 2000, 2006, 2010 estimated from the Bank of Italy Survey on Household Income and Wealth, see Fig. 1. Since $L(p)$ is the percentage of the income $X$ held by the $p100\%$ "poorest" part of the population, each curve represents how the income concentrates within a region population in a given year.

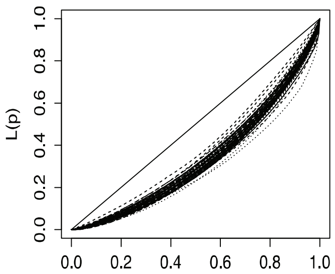

**Fig. 1** Each curve illustrates the concentration of family income in a given year (2000, 2006, 2010) and region for a total of 60 empirical LCs.

These curves can be seen as a sample of a random element taking values in $\mathscr{L}or$, the family of continuous, increasing and convex functions from $[0,1]$ to itself passing through the origin and $(1,1)$. In this view, one can explore data by borrowing techniques from functional data analysis (FDA): a recent branch of statistics that studies those phenomena whose observations are (discretized) curves; see e.g. [3, 4, 6]. Altough a standard FDA approach for LCs is possible, in general, it is not advisable. In fact, LCs are special functional data not directly observed but estimated from a sample of a real random variable: this leads to a double stochasticity issue that could impact over usual FDA techniques. Moreover, given the constrained nature of the Lorenz curve process, $\mathscr{L}or$ is not a structured space (for instance Hilbert) and then classical methods should be used with caution.

The aim of this work is to explore the variability of the described data by means of the "modes of variation". For a given functional process, its $j$-th mode of variation is the mean function perturbed by $\pm k\sqrt{\eta_j}v_j$ where, $k > 0$ and $\{\eta_j, v_j\}$ are the $j$-th eigenelements of the covariance operator of the process. As a consequence, modes of variation are usually computed after the functional principal component analysis

(FPCA), but, given the above remarks on LCs, a naive application of FPCA leads to modes of variation not belonging to $\mathscr{L}or$ and then to incoherent interpretations. To tackle such issue, a preliminary transformation of data is necessary.
The remain part of this work is divided in two sections: Sect. 2 describes the embedding proposed by [1] and the procedure to compute the modes of variations whereas Sect. 3 illustrates some shortcomings arising with a naive FPCA and applies the method presented in Sect. 2 to the Bank of Italy dataset (see Fig. 1).

## 2 Embedding and FPCA

Consider

$$\mathscr{L}or = \{L \in C^2_{[0,1]} : L(0) = 0, L(1) = 1, L' > 0, L'' > 0\},$$

where $L'$ and $L''$ denote the first and second derivative of $L$ respectively.
The following map

$$\psi(L) = -\ln\left(L''\right) + \int_0^1 \ln(L''(p))dp, \qquad \forall L \in \mathscr{L}or$$

is a bijection from $\mathscr{L}or$ into the separable Hilbert space $\mathscr{L}^2_c = \{g \in \mathscr{L}^2_{[0,1]} : \int g = 0\}$ and its inverse, for any $g \in \mathscr{L}^2_c$, is given by

$$\psi^{-1}(g)[p] = p + (p-1)\int_0^p z\exp\left(-g(z)\right)/\kappa_g dz + p\int_p^1 (z-1)\exp\left(-g(z)\right)/\kappa_g dz$$

where $\kappa_g = \int_0^1 \int_0^p \exp\{-g(z)\}dzdp$ is a scale technical factor. Hence, thanks to $\psi$, $\mathscr{L}or$ can be endowed with a Hilbert structure inherited by $\mathscr{L}^2_c$. This allows to properly perform FPCA and to compute modes of variations in $\mathscr{L}^2_c$ as usual. Moreover, $\psi^{-1}$ can be used to map the obtained results back in $\mathscr{L}or$.
In particular, given a sample of empirical LCs $\{\widehat{L}_i(p), i = 1, \dots, n\}$ each one estimated from a sample drawn from a random variable $X_i$, the following procedure can be implemented.

**An embedding approach for Lorenz FPCA**

1. Get $\widetilde{L}''_i(p)$ from $\widehat{L}_i(p)$ by using a suitable smoother (e.g. local polynomial).
2. Embed the LC in the Hilbert space $\mathscr{L}^2_c$ by means of $\psi$:

$$\psi(\widehat{L}) = -\ln(\widetilde{L}''_i) + \int_0^1 \ln(\widetilde{L}''_i(p))dp.$$

3. Implement the FPCA in $\mathscr{L}^2_c$ by computing the empirical
   - mean $\widehat{\mu}$, covariance operator $\widehat{\Sigma}$ and its eigenelements $\{\widehat{\lambda}_j, \widehat{\xi}_j\}$;

- $j$-th mode of variation of $\psi(\widehat{L})$ that is

$$\widehat{m}_{j,k} = \widehat{\mu} \pm k\sqrt{\widehat{\lambda}_j}\widehat{\xi}_j,$$

for any $k > 0$ and $j \in \{1, \dots, n\}$.

4. Pull $\widehat{m}_{j,k}$ back into $\mathscr{L}or$ by using $\psi^{-1}$, to get $\widehat{M}_{j,k} = \psi^{-1}(\widehat{m}_{j,k})$ the $j$-th mode of variation in $\mathscr{L}or$.

The described procedure is statistically consistent since, under mild regularity conditions on the cdf $F$ and as $n \to \infty$, $\widehat{M}_j(k)$ converges in probability to $M_j(k)$ the theoretical $j$–th modes of variation when LCs are integrally observed.

## 3 Application

In this section the proposed approach is applied to the Bank of Italy dataset (see Fig. 1). To better understand why an embedding approach is advantageous to study the modes of variation instead of a direct one approach, the FPCA is firstly performed on the original dataset of empirical LCs: the corresponding first three modes of variations for different $k$ are plotted in Fig. 2. From the latter, it emerges that the direct approach provides coherent interpretations only for small values of $k$ since for large values of $k$ the modes of variations are no longer LCs. Fig. 3 depicts the modes of variations computed via the embedding approach for different $k$. As expected, since they are elements of $\mathscr{L}or$, it is possible to understand how the first three PCs impacts on the mean and how they explain the variability of LCs.

Another interesting point is the analysis of the information brought by the factor plane. Since the phenomenon under study is rather complex, some synthetic indexes, such as the Gini one, are often used to help the researchers. The PCs allow to explain the basic dynamics that regulate the composition of the LCs and therefore to go beyond the analysis of a single index. To do this, consider the track-plots that allow to appreciate the dynamics over time of the LCs with respect to the first two PCs; see Fig. 4. Note that, even if the Gini index for one specific region can assume similar values in distinct years, it can be placed in different quadrants of the factorial plane over the time suggesting the presence of latent structures that can not be detected by the synthetic index alone.

**Acknowledgements** The authors are members of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). The financial support of Università del Piemonte Orientale is acknowledged by the authors.

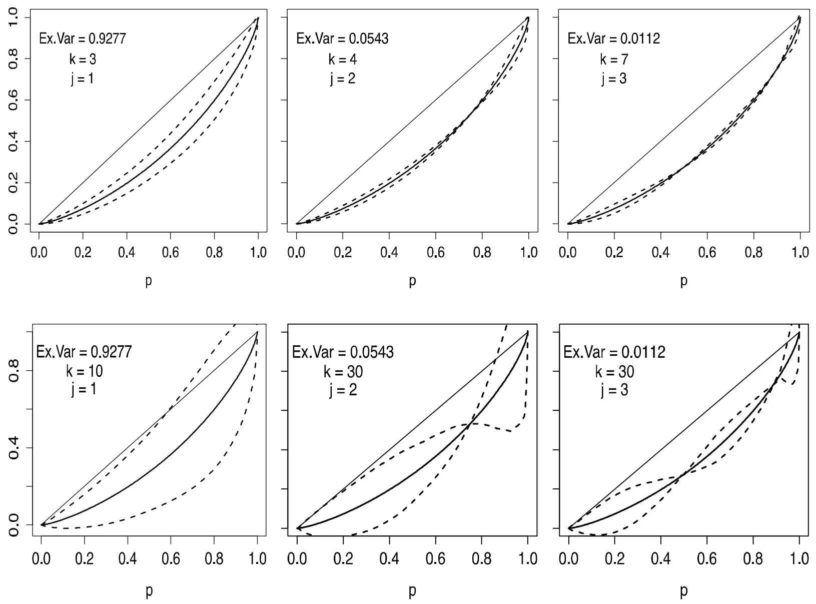

**Fig. 2** Fraction of explained variance of the $j$-th PC, mean curve (solid line) and modes of variation for $j = 1, 2, 3$ and different $k$ (dashed lines) for the sample of original LCs.

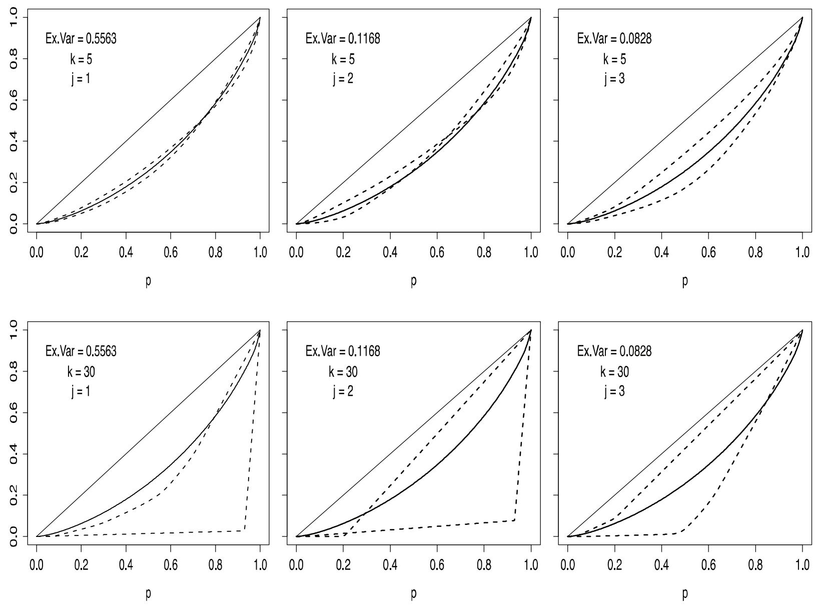

**Fig. 3** Fraction of explained variance of the $j$-th PC, mean curve $\widehat{M}_j(0)$ (solid line) and modes of variation $\widehat{M}_j(k)$ for different values of $j$ and $k$ (dashed lines) for the sample of LCs in Fig. 1.

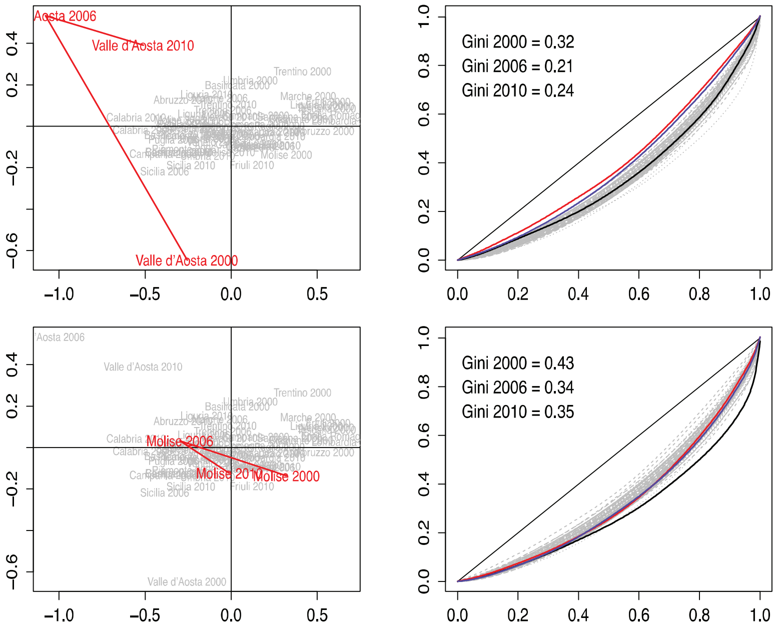

**Fig. 4** (Left) Track-plots in the factorial plane of the first two PCs. (Right) Assosiated LCs and Gini indexes.

## References

1. Bongiorno, E.G., Goia, A.: Describing the Concentration of Income Populations by Functional Principal Component Analysis on Lorenz curves. J. Multivariate Anal., **170**, 10–24 (2019)
2. Bosq, D.: Linear Processes in Function Spaces: Theory and Applications. Lectures Notes in Statistics, 149, Springer–Verlag, Berlin (2000)
3. Ferraty, F., Vieu, P.: Nonparametric functional data analysis. Theory and practice. Springer Series Stat. (2006)
4. Kokoszka, P., Reimherr, M.: Introduction to functional data analysis. Chapman and Hall/CRC (2017)
5. Lorenz, M.O.: Methods of measuring the concentration of wealth. Amer. Statistical Assn. J., **9** (70), 209–219, (1905)
6. Ramsay, J.O., Silverman, B.W.: Functional data analysis, 2nd ed., New York: Springer (2005)