47th Scientific Meeting
of the
Italian Statistical Society

PROCEEDINGS

Editors: S. Cabras, T. Di Battista and W. Racugno

CUEC EDITRICE

ISBN: 978-88-8467-874-4
47th SIS Scientific Meeting of the Italian Statistica Society
Cagliari, June 10-14, 2014.

CUEC Cooperativa Universitaria Editrice Cagliaritana
via Is Mirrionis 1, 09123 Cagliari
Tel. e Fax 070271573
www.cuec.eu
info@cuec.eu

ISBN 978-88-84-67-874-4
9788884678744

Program Committee

The Program Committee of the 47^{th} SIS Scientific Meeting is composed by:

Tonio Di Battista (*Chair*)	Università di Pescara
Marco Alfò	Sapienza Università di Roma
Giulio Barcaroli	ISTAT
Gianni Betti	Università di Siena
Daniela Cocchi	Università di Bologna
Stefano Gattone	Università di Roma "Tor Vergata"
Andrea Giommi	Università di Firenze
Salvatore Ingrassia	Università di Catania
Michele La Rocca	Università di Salerno
Letizia Mencarini	Università di Torino
Sonia Petrone	Università Commerciale Bocconi
Alessandra Petrucci	Università di Firenze
Donato Posa	Università di Lecce
Walter Racugno	Università di Cagliari
Giovanna Ranalli	Università di Perugia
Marco Riani	Università di Parma
Roberto Rocci	Università di Roma "Tor Vergata"
Arjuna Tuzzi	Università di Padova

Preface

The edition of this volume gave us the opportunity to perceive that, together with many well-known Italian statisticians belonging to the national and international community, many young researchers are emerging. They presented, at the *47th Scientific Meeting of the Italian Statistical Society*, their remarkable contributions, both from the methodological and the applicative point of views.

Although some papers may not look, in their actual form, fully mature from the scientific and communicative point of views, we decided - in agreement with the referees - to publish them since promising and full of ideas. In this respect, the contributions published in this volume provide a comprehensive overview of the current Italian scientific researches in theoretical and applied statistics.

This volume also contains several contributions presented by foreign researchers, highlighting the fact that the Italian Statistical Society has an attractive role in the international scientific community.

Finally, we would like to emphasize that, even from the abstracts of the contributions, the wideness of the collaborations between the statisticians and the experts from other fields emerges. This denotes that, also in Italy, statistical methods are spreading in the different fields of the scientific researches.

Stefano Cabras
Tonio Di Battista
Walter Racugno

Cagliari, June 11, 2014

Table of Contents

List of papers

List of abstracts

You can access to the full paper by clicking on their link.

CP16 - Environmental And Poverty Data Analysis

- Lucia Paci, Daniela Cocchi, Alan Gelfand. **Quantifying uncertainty associated with a numerical model output** (Full paper: 2824.pdf)
- Giuliana Passamani, Paola Masotti. **Smoothed Common Trend in Multivariate Time Series Air Pollution Data** (Full paper: 2952.doc)
- Silvia Perra, Stefano Cabras, Alberto Serci, Alessandra Mura, Antonella De Arca, Stefano Renoldi, Antonello Podda. **IDMS: The Sardinian Index of Multiple Deprivation** (Full paper: 2860.doc)
- Alessio Pollice, Vito Muggeo, Federico Torretta, Rocco Bochicchio, Mariana Amato. **Growth curves of sorghum roots via quantile regression with P-splines** (Full paper: 2956.pdf)
- Abel Rodriguez, Fernado Quintana. **On species sampling sequences induced by residual allocation models** (Full paper: 2885.pdf)

CP17 - Topics In Regression Models

- Matilde Bini, Vieri Del Panta, Margherita Velucchi. **Mixtures of Logit Regressions Detection with Forward Search** (Full paper: 2982.pdf)
- Silvia Columbu, Matteo Bottai. **Conditional concordance of the signs of the residuals of quantiles regressions of multivariate outcomes** (Full paper: 2965.pdf)
- Roberto Fontana, Fabio Rapallo, Maria Piera Rogantin. **Indicator functions and saturated fractions for factorial designs: a case study** (Full paper: 2888.pdf)
- Pia Clara Pafundi, Gianmarco Vacca. **Complex Redundancy Analysis models with covariate effect: a simulation study** (Full paper: 2869.doc)
- Mariangela Sciandra. **Variable selection in mixed models: a graphical approach** (Full paper: 2931.pdf)

CP18 - Bayesian Methods And Models

- Julyan Arbel, Kerrie Mengersen, Judith Rousseau. **On diversity under a Bayesian nonparametric dependent model** (Full paper: 2936.pdf)
- Antonio Canale, Bruno Scarpa. **Skew-normal nonparametric mixture models** (Full paper: 2844.pdf)
- Giulia Roli, Meri Raggi. **Bayesian hierarchical models for misaligned data: a simulation study** (Full paper: 2927.pdf)
- Catia Scricciolo, Sophie Donnet, Vincent Rivoirard, Judith Rousseau. **Posterior contraction rates for empirical Bayes procedures with applications** (Full paper: 2918.pdf)

POSTER SESSION

- Antonio Arcos, María Del Mar Rueda, David Molina, Maria Giovanna Ranalli. **Frames2: an R package for estimation in dual frames** (Full paper: 2993.pdf)
- Simona Arcuti, Alessio Pollice, Crescenza Calculli, Nunziata Ribecco, Angelo Tursi. **Spatial smoothing on complex regions: a case study on the median length of deep water rose shrimps in the North-Western Ionian Sea** (Full paper: 2921.pdf)

- Filippa Bono, Marcella Giacomarra. **The effect of support schemes on Photovoltaic installed capacity in Europe: a WDEA-STATIS analysis** (Full paper: 2988.docx)
- Maria Caterina Bramati, Flaminia Musella. **Bayesian Network structural learning in multivariate time series** (Full paper: 2856.pdf)
- Maurizio Brizzi, Alessia Orrù. **Gender Differentiation of Human Longevity in Sardinian Provinces** (Full paper: 2943.doc)
- Simone Del Sarto, M. Giovanna Ranalli, Elena Stanghellini, Davide Cappelletti, Beatrice Moroni, Stefano Crocchianti, Silvia Castellini. **Modelling the effect of vehicular–traffic and meteorology on fine particle concentration using Additive Mixed Models: the case of the town of Perugia** (Full paper: 3148.pdf)
- Rosa Falotico, Paolo Mariani. **Outsourcing in the Italian NHS: a statistical measure of mismatch between private supply and public demand** (Full paper: 2929.doc)
- Francesca Fortuna, Fabrizio Maturo. **Functional analysis of variance for parametric functional data** (Full paper: 3150.pdf)
- Massimiliano Giacalone, Paolo Carmelo Cozzucoli. **A performance comparison of the Lp-norm methods in multicollinearity situations, supposing a generalized normal distribution errors** (Full paper: 2969.doc)
- Clara Grazian, Christian Robert. **Jeffreys Priors for Mixture Models** (Full paper: 2891.pdf)
- Fabrizio Maturo, Francesca Fortuna. **Bell shaped fuzzy numbers associated with the normal curve** (Full paper: 3149.pdf)
- Daria Mendola, Annalisa Busetta, Philippe Van Kerm, Anna Maria Milito. **Material deprivation among foreigners in Italy** (Full paper: 2890.doc)
- Eugenia Nissi, Annalina Sarra. **Local Spatial Analysis of Cardiovascular Diseases in Canadian Health Regions** (Full paper: 3151.pdf)
- Zoe Nivolianitou, Myrto Konstandinidou, Chrys Caroni, Irini Kefalogianni. **Analysis of the causal factors in incidents of the Greek petrochemical industry** (Full paper: 2866.doc)
- Anna Pinto, Giulia Mascarello, Nicoletta Parise, Silvia Bonaldo, Stefania Crovato, Licia Ravarotto. **A classification of Italian consumers based on a proposed measure of their attitudes towards food risks** (Full paper: 2896.pdf)
- Giovanni Pistone. **A version of the geometry of the multivariate Gaussian model, with applications** (Full paper: 2855.pdf)
- Pasquale Recchia, Ernesto Toma. **Family size and educational outcomes: empirical evidence through multilevel approach** (Full paper: 2884.docx)
- Maria Piera Rogantin, Giovanni Pistone. **Fractionalization and Polarization** (Full paper: 2910.pdf)
- Elvira Romano, Maria Dolores Ruiz-Medina, Rosa M. M. Espejo. **A spatial functional approach for curve classification** (Full paper: 3110.pdf)
- Giorgio Russolillo, Laura Trinchera. **An extension of Non-Metric approach to inwards directed PLS Path Models** (Full paper: 3036.pdf)
- Aldo Solari, Jelle Goeman. **Adapting Benjamini-Hochberg by Simes Inequality** (Full paper: 3083.pdf)
- Susanna Zaccarin, Domenico De Stefano, Vittorio Fuccella, Maria Vitale. **Co-authorship Patterns of Italian Statisticians by Combining Different Data Sources** (Full paper: 2941.pdf)

A performance comparison of the Lp-norm methods in multicollinearity situations, supposing a generalized normal distribution errors

Un confronto tra le prestazioni dei metodi di norma Lp in casi di multicollinearità, ipotizzando una distribuzione di errori normale generalizzata

Massimiliano Giacalone. Paolo Carmelo Cozzucoli

Abstract The normal distributions of order p, known also as exponential power distributions, are used in the description of non normal random errors in the general situation of L_p-norm estimation. L_1-norm estimators (least absolute deviation); L_2-norm estimators (least squares) and other L_p-norm proposal estimators are reviewed and applied to estimate the parameters of a particular linear regression model. In the paper are also reported few results concerning three alternative p-estimation methods to evaluate their performance for a linear model.

Abstract Le distribuzioni di ordine p, note anche come distribuzioni di potenza esponenziali, sono utilizzate nella descrizione degli errori casuali non normali nella situazione generale della stima L_p-norm. Stimatori di norma L_1 (deviazione dal minimo assoluto); stimatori di norma L_2 (minimi quadrati) e altri stimatori proposti del tipo Lp-norm sono rivisitati e applicati per stimare i parametri di un particolare modello di regressione lineare. Nel lavoro vengono, inoltre, riportati alcuni risultati riguardanti tre alternativi metodi di stima al fine di valutare le loro prestazioni per un modello lineare.

Key words: Normal Distributions of order *p*, exponential power distributions, L_p-norm estimators, Monte Carlo studies, p estimation methods.

---

[1] Massimiliano Giacalone, lecturer with annual appointment, University of Bologna "Alma Mater Studiorum"; Forli' Campus; email: massimilia.giacalone@unibo.it

Paolo Carmelo Cozzucoli, University of Calabria, Department of Economics, Statistics and Finance; email: cozzucoli@unical.it

1 The Exponential Power Function and the L_p-norm estimators

The E.P.F is a family of probability functions proposed by Subbotin in 1923 and studied by Vianelli (1963), Lunetta (1966), Mineo A (1989). The density function is:

$$f_p(z)=\frac{1}{2p^{1/p}\sigma_p\Gamma\left(1+1/p\right)}\exp\left[-\frac{1}{p}\left|\frac{z-M_p}{\sigma_p}\right|^p\right] \qquad (1)$$

where $M_p=E(z)$ is the *location parameter*, $\sigma_p=\left(E\left[\left|z-M_p\right|^p\right]\right)^{1/p}$ is the *scale parameter*, and $p>0$ is the *shape parameter*.

Considering the Pearson kurtosis index β_2 we distinguish:

1. $0<p<1$ double exponential distributions, $\beta_2>6$;
2. $1<p<2$: leptokurtic distributions, $3<\beta_2<6$;
3. $p>2$: platikurtic distributions, $1.8<\beta_2<3$.

   For particular values of p we have:

   a) the **Laplace distribution** ($p=1$, $\beta_2=6$);

   b) the **Gaussian distribution,** ($p=2$, $\beta_2=3$);

   c) and the **Uniform distribution** ($p\to\infty$, $\beta_2\to 1.8$).

This family of curves, taken up by several authors, is often cited in the literature as exponential power distribution, but with different parameter settings, for example that of Box, Tiao (1973), to study the robustness of Bayesian inference, in which p is related to a parameter β of non-normality by the relation: $p=2/(1+\beta)$.
In the form used by us, μ= E[Z] is the true value of a quantity, whose observed values z are affected by errors with dispersion σ_p=(E| Z - μ |^p)$^{1/p}$, average deviation absolute order p. The curves are unimodal, symmetric and, for p>1, bell-shaped. As special cases, have the Laplace distribution (p=1), the normal (p=2) and the uniform (p $\to\infty$).The shape parameter p is connected with the kurtosis, and the values of the index of kurtosis β_2 vary from β_2 for $p\to 0^+$ to 1,8 for uniform distribution ($p\to\infty$).

1.1 L$_p$-*norm estimators for linear regression models*

Let us consider a sample of *n* observed data (x_i, y_i), a general linear regression model is:

$$y_i = g\left(\underline{x}_i, \underline{\theta}\right) + e_i \tag{2}$$

The L_p -norm estimators minimize the sum of the **p**-th power of the absolute deviations of the observed points from the regression function:

$$S_p(\theta) = \sum_{i=1}^{n} \left| y_i - g(x_i, \theta) \right|^p \qquad 1 \le p \le \infty \tag{3}$$

Under regularity assumptions the log-likelihood related to the sample is given by:

$$L\left(\theta, \sigma_p, p\right) = -n\log\left[2p^{1/p}\sigma_p\Gamma(1 + 1/p)\right] - \left[\left(p\sigma_p\right)^{-1}\sum\left|y_i - g\left(\underline{x}_i, \underline{\theta}\right)\right|^p\right] \tag{4}$$

where we consider $z = y_i$ and $\mu_p = g(\underline{x}_{i,}\underline{\theta})$

$$\delta L / \delta\theta_j = \sum_{i=1}^{n} \left| y_i - g\left(\underline{x}_i, \underline{\theta}\right) \right|^{p-1} \text{ sign } \left(y_i - g\left(\underline{x}_i, \underline{\theta}\right)\right) \frac{\delta g}{\delta\theta_j} = 0$$

$$\sum \left| y_i - g\left(\underline{x}_i, \underline{\theta}\right) \right|^p = \min \quad \text{with } p \ge 1 \tag{5}$$

The optimal exponent p for the L_p-norm estimators of the regression parameters is the shape parameter p of the E.P.F.

If p is unknown we have two related problems to consider:

1) The estimation of the *exponent* p on the sample data

2) The choice of the minimization algorithm to obtain the *regression parameters* estimation (see Gonin & Money, 1989).

2 Multicollinearity in linear regression models

The purpose of a regression model is to find out to what extent the outcome (dependent variable) can be predicted by the independent variables. The strength of the prediction is indicated by R^2, also known as variance explained or strength of determination.

The absence of multi-collinearity is essential to a multiple regression model. In regression when several predictors (regressors) are highly correlated, this problem is called multi-collinearity or collinearity. When things are related, we say they are linearly dependent on each other because you can nicely fit a straight regression line to pass through many data points of those variables. Collinearity simply means co-dependence. It is problematic when one's purpose is explanation rather than mere prediction. Collinearity makes it more difficult to achieve significance of the collinear parameters. But if such estimates are statistically significant, they are as reliable as any other variables in a model. And even if they are not significant, the sum of the coefficient is likely to be reliable. In this case, increasing the sample size is a viable remedy for collinearity when prediction instead of explanation is the goal (Leahy, 2001). However, if the goal is explanation, measures other than increasing the sample size are needed as the Ridge Regression (Morris, 1982; Pagel Lunneberg 1985) or the Partial Least Squares (Cassel, et al 1999).

Variation Inflation is the consequence of multi-collinearity. We may say multi-collinearity is the symptom while variance inflation is the disease. In a regression model we expect a high variance explained (R-square). The higher the variance explained is, the better the model is. However, if collinearity exists, probably the variance, standard error, parameter estimates are all inflated.

The same information might be expressed in terms of the so-called Variance Inflation Factors:

$$\mathbf{VIF = 1/(1-R^2)}$$

As the name indicates, the V.I.F. measures the factor by which the parameter's variance (in an orthogonal regression; hence without multicollinearity) is multiplied (c.q. inflated).

In lecterature we find many examples of how the condition index and the Variance Inflation Factors can be used to test for possible multicollinearity. The VIF test is then applied to our example-equation in the following simulation studies.

3 The simulation plan and final remarks

We consider 500 samples of sizes n=50, 100, 200 generated from E.P.F., and 6 values of *p*, ranging from 1.0 to 3.5 with step 0.5. The algorithm for generating the ε_i (for $p \geq 1$) from an E.P.F. is suggested by Chiodi (1986).

The values of y_i are given by the multiple regression model :

$$y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 \qquad (6)$$

with $\mathbf{X_1}$ $\mathbf{X_2}$ identically and independent variables, distributed from a Gaussian standardized distribution and $\mathbf{X_3}$ linear combination of $\mathbf{X_1}$ and $\mathbf{X_2}$.

$$\mathbf{X_3 = X_1 + X_2 + Z} \qquad \text{with} \quad \mathbf{Z - N(0, \sigma_z)} \tag{7}$$

Then we can write the related variance and covariance matrix:

	X_1	X_2	X_3
X_1	1	0	1
X_2	0	1	1
X_3	1	1	$2+\sigma^2_z$

It is easy to see that: $\mathbf{E\,(X_3^2) = E\,(X_1^2) + E\,(X_2^2) + \sigma^2_z = 2 + \sigma^2_z}$.

and the correspondent correlation matrix is equal to:

1	0	$1/(2+\sigma^2_z)$
0	1	$1/(2+\sigma^2_z)$
$1/(2+\sigma^2_z)$	$1/(2+\sigma^2_z)$	1

Where $\mathbf{r_{13}} = \mathrm{cov}\,(X_1, X_3) / (\mathrm{var}(X_1)\,\mathrm{var}(X_3)) = \mathbf{1 / (1^{*}(2+\sigma^2_z)) = r_{23}}$

And $\mathbf{R^2_{3.12} = 1 - detA/detA_{33} = 2/\,(2+\sigma^2_z)}$

We can observe that in our simulation model the rate of multicollinearity is proportional (in inverse way) to σ^2_z.

Considering the parameter values $\boldsymbol{\beta_0 = 1}$, $\boldsymbol{\beta_1 = 2}$, $\boldsymbol{\beta_2 = 3}$, $\boldsymbol{\beta_3 = 4}$, we make a comparative analysis applying the three following estimation methods on the same samples :

1. Least squares estimators (L_2).
2. L_p -norm estimators with theoretical p of the E.P.F. (L_p).
3. L_p -norm estimators with $p=1$ (Least absolute deviation) (L_1)

The simulation studies show how in some cases (L_p) achieves more efficient estimates for the regression parameters when compared with least squares procedure. In particular, using the L_p and the L_1 method better performances in terms of the variances of the parameter estimates, are always obtained in the case of nonnormal symmetric distributions respect to the least squares situation also considering a model with collinear regressors.

References

Agro' G. (1995), Maximum likelihood estimation for the exponential power function parameters. Comm. Statist - Simula. Computa.,24, 523,536.

Box G.E.P., Tiao G.C. (1973) *Bayesian inference in statistical analysis*. Addison-Wesley Ed.; Reading, Massachusetts.

Chiodi M. (1986), "Procedures for Generating Pseudo-Random Numbers from a Normal Distribution of Order p" Riv. Stat. Applicata, 19, 7-26.

Chiodi M. (2000) Le curve normali di ordine p nell'ambito delle distribuzioni di errori accidentali: una rassegna dei risultati e problemi aperti per il caso univariato e per quello multivariato. Atti della XXXX Riunione Scientifica SIS. Ed. C.I.S.U. Roma.

Giacalone M. (1997) Lp-norm estimators for nonlinear regression models. Student, vol 2, 119-130.

Gonin R. and Money A.H. (1989) "Nonlinear Lp-Norm Estimation" Marcel Dekker Inc. New York - Basel.

Leahy, K. (2001). Multicollinearity: When the solution is the problem. in Olivia Parr Rud (Ed.). Data Mining Cookbook (pp. 106 - 108). John Wiley & Sons, Inc,

Lunetta G. (1966) Di alcune distribuzioni deducibili da una generalizzazione dello schema della curva normale. Annali della Facoltà Economia e Commercio di Palermo, 20,1, 117-143.

Mineo A. M. (1994) Un nuovo metodo di stima di p per una corretta valutazione dei parametri di intensità e di scala di una curva normale di ordine p. Atti della XXXVII Riunione Scientifica della SIS. Ed. C.I.S.U. Roma; 1, 147-154.

Mineo, A.M. (2006), Le Curve Normali di Ordine p come Distribuzione Generale degli Errori Accidentali: Uno Strumento Software per il loro Utilizzo. In "Metodi, Modelli e Tecnologie dell'Informazione a supporto delle Decisioni".Ed. P.Amenta - L. D'Ambra- M. Squillante - A. Ventre, F. Angeli, 463-471.

Subbotin M. T.(1923) On the law of frequency of errors, Matem. Sbornik, 31, 296-301.