Statistics and Operations Research Transactions,

vol. 42, n. 1 (2018)

SORT 42 (1) January-June 2018, 3-26 DOI: 10.2436/20.8080.02.66

Using a Bayesian change-point statistical model with autoregressive terms to study the monthly number of dispensed asthma medications by public health services

J.A M. de Queiroz[1,2], D.C. Aragon[3], L.M. de Mello[4], I.T.S. Previdelli[2] and E.Z. Martinez[4,*]

Abstract

In this paper, it is proposed a Bayesian analysis of a time series in the presence of a random change-point and autoregressive terms. The development of this model was motivated by a data set related to the monthly number of asthma medications dispensed by the public health services of Ribeirão Preto, Southeast Brazil, from 1999 to 2011. A pronounced increase trend has been observed from 1999 to a specific change-point, with a posterior decrease until the end of the series. In order to obtain estimates for the parameters of interest, a Bayesian Markov Chain Monte Carlo (MCMC) simulation procedure using the Gibbs sampler algorithm was developed. The Bayesian model with autoregressive terms of order 1 fits well to the data, allowing to estimate the change-point at July 2007, and probably reflecting the results of the new health policies and previously adopted programs directed toward patients with asthma. The results imply that the present model is useful to analyse the monthly number of dispensed asthma medications and it can be used to describe a broad range of epidemiological time series data where a change-point is present.

MSC: 62P10, 62F15, 62M10, 37M10.

Keywords: Time series, regression models, Bayesian methods, change-point model, epidemiological data.

*e-mail: edson@fmrp.usp.br

[1] Instituto Federal de Educação, Ciência e Tecnologia do Paraná (IFPR), Jacarezinho, Brazil

[2] Master Program in Biostatistics, Department of Statistics, State University of Maringá (UEM), Maringá, PR, Brazil

[3] Department of Pediatrics, Ribeirão Preto Medical School, University of São Paulo (USP), Ribeirão Preto, SP, Brazil

[4] Department of Social Medicine, Ribeirão Preto Medical School, University of São Paulo (USP), Ribeirão Preto, SP, Brazil.

Received: April 2017

Accepted: January 2018

1. Introduction

In many situations, epidemiological data come in the form of time series. Disease notifications, hospitalizations due to a specific disease and mortality rates over a given time interval are examples of variables which can be studied as time series. Statistical models are useful to describe patterns of these series, such as temporal trends and seasonal fluctuations. These models can be also used to predict future observations after observing a series of longitudinal data, thus supplying information to aid in the surveillance and management of events of public health interest.

Change-point models (Jensen and Lautkebohmert, 2007; Lee, 2010) have been increasingly used in a broad spectrum of applications, such as in econometrics (Hackl, 2012), medicine (Ghosh and Vaida, 2007) and environmental studies (Achcar et al., 2010; Achcar, Rodrigues and Tzintzun, 2011). These models are statistical tools used in practical problems where a random variable indexed by time has modified their behaviour at one or more time instants. Thus, these models are useful when the interest of the analyst lies in determining whether the observed time series is homogeneous over the time interval. As an example, Achcar et al. (2008) considered a change-point analysis for the incidence of tuberculosis cases in New York City from 1970 to 2000, when the number of cases of the disease presented three trends. In the first period of time, the trend of declining incidence was probably associated with good control programs. In the second period, there were increasing incidence rates, and in the third period there was a new trend of declining rates. Modern Bayesian methods of inference by using Markov Chain Monte Carlo (MCMC) techniques have been used to fit time series data in the presence of one or more change-points (Achcar and Loibel, 1998; Barry and Hartigan, 1993; Carlin, Gelfand and Smith, 1992; Dey and Purkayastha, 1997; Lavielle and Lebarbier, 2001), including multiple change-point models where the number of change-points is unknown (Chib, 1998; Fearnhead, 2006).

The present article introduces a single Bayesian model for change-point detection including autoregressive terms to be applied to the monthly number of asthma medications dispensed by the public health services of Ribeirão Preto, Southeastern Brazil. Climatic variables are included as independent variables.

2. Methods

2.1. Field of study and dataset

The present study is part of a larger research on dispensation of medications to treat pulmonary diseases in the public health services of Ribeirão Preto, a city located in the northwest region of the State of São Paulo, Brazil. Ribeirão Preto is ranked the eighth largest city in the State of São Paulo, with about 600 thousand inhabitants (IBGE census data, available from www.censo2010.ibge.gov.br/sinopse/). The city belongs to the health coverage area of the XIII Regional Health Department of the Health Secretariat

of the State of São Paulo, being considered a regional health care centre reference for interventions of medium and high complexity and attending more than 1.2 million people, of which approximately 62% depend exclusively on the Brazilian National Health System (SUS) (Bittar, Mendes and Magalhães, 2011). The public healthcare network in Ribeirão Preto is composed by municipal, state and philanthropic services, involving 36 pharmacies providing pharmaceutical care according to the National Drug Policy guidelines (GM Ordinance number 3916 of November 30th, 1998) and currently offering to the population over 260 medications indicated for the treatment of various diseases, including asthma.

Table 1: *Monthly data on the number of dispensed medications (salbutamol sulfate tablets of 2 mg) in Ribeirão Preto, Brazil, from February 1999 to December 2011.*

Year	Jan.	Feb.	Mar.	Apr.	May.	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
1999		2,376	2,900	1,699	2,066	4,329	5,486	5,651	7,732	5,505	4,267	3,843
2000	6,855	7,809	7,659	6,386	6,822	8,297	5,936	8,255	6,171	7,119	4,980	7,523
2001	9,563	9,269	9,605	10,150	11,867	10,482	11,718	12,412	9,183	13,667	12,046	9,150
2002	10,641	11,975	10,651	6,089	13,843	15,336	16,418	15,401	12,518	12,960	11,295	13,100
2003	11,756	5,043	2,057	9,131	13,654	12,785	15,071	10,549	11,633	9,085	12,884	11,218
2004	10,477	12,671	18,303	17,445	15,606	15,011	19,448	17,124	15,132	13,218	17,054	14,596
2005	14,433	12,569	17,053	16,110	18,346	19,218	18,847	19,209	15,435	18,274	17,313	18,392
2006	13,411	12,675	18,597	16,258	20,357	20,457	16,339	18,552	16,910	20,617	19,634	23,567
2007	21,981	22,981	25,914	21,607	30,083	19,008	23,103	21,893	16,974	20,066	17,606	15,846
2008	18,134	18,578	18,306	17,982	21,032	19,222	19,274	15,841	13,864	14,600	13,431	12,865
2009	11,722	10,862	14,184	13,414	15,257	16,914	13,906	14,752	14,762	14,305	12,590	14,843
2010	11,876	12,284	14,468	13,505	13,765	11,929	4,313	10,475	11,644	11,837	9,949	10,040
2011	9,328	9,095	8,998	7,987	8,161	9,278	7,343	7,672	6,082	5,678	6,141	4,951

Data on the number of dispensed medications were provided by the HygiaWeb Information System, a health information system which has been used by the Municipal Health Secretariat of Ribeirão Preto since 1992. This system enables to record information on health services in the entire municipal public healthcare network. In 1998, the implementation of a medication management module in the HygiaWeb System enabled the recording of data on dispensation of medications and pharmaceutical care. Therefore, it has been possible to retrieve secondary information about the dispensation of the main medications for asthma treatment since 1999, covering the whole city. For the purposes of the present study, data on only one drug used to alleviate the symptoms of asthma have been considered for developing the statistical model, namely, salbutamol sulfate tablets of 2 mg. Full data on the number of dispensed medications, from February 1999 to December 2011, are listed in Table 1. In addition, data on temperature and precipitation in the city of Ribeirão Preto were obtained from the Integrated Agrometeorological Information Center of the Agronomic Institute (CIIAGRO, Centro Integrado de Informações Agrometeorológicas do Instituto Agronômico).

The local Research Ethics Committee has approved the present study (CEP/CSE/FMRP/USP, protocol number 453) and the permission to access and use the records from

the HygiaWeb System was granted by the local representative of the Health Department (document 248/11-GS RAS/ras).

2.2. Statistical model

Let y_t be the number of dispensed medications at the month t, $t = 1,\ldots,n$, where n is the number of months in the time series. The proposed model is given in a general form by

$$y_t = \alpha + g_1 I_{[1,\theta]}(t) + g_2 I_{(\theta,n]}(t) + S_t + \varepsilon_t,$$

where

$$g_k(t) = \beta_k (t-\theta) + \sum_{r=1}^{R} \psi_{kr} (x_{rt} - \overline{x}_r) + \sum_{j=1}^{p} \gamma_{kj} (y_{t-j} - \overline{y}), \;\; k = 1,2,$$

α is an intercept term, $I_{\{A\}}(t)$ denotes an indicator function such that $I_{\{A\}}(t) = 1$ if $t \in \{A\}$, and 0 otherwise, θ is the change-point to be estimated such that θ is an integer number in the interval $[1,n]$, $x_{1t}, x_{2t}, \ldots, x_{Rt}$ are observations of R covariates at the month t, $\overline{x}_r$ denotes the mean of $x_{r1},\ldots,x_{rn}$, $r = 1,\ldots,R$, ψ_{1r} and ψ_{2r} are the effects of the covariate x_{rt} on y_t before and after θ, respectively, $\overline{y}$ denotes the mean of $y_1,\ldots,y_n$, the terms $\gamma_{11},\ldots,\gamma_{1p},\gamma_{21},\ldots,\gamma_{2p}$, are autoregressive parameters of order p to be estimated and the random error terms are represented by ε_t. In addition,

$$S_t = \eta_1 \sin\left(\frac{2\pi t}{12}\right) + \eta_2 \cos\left(\frac{2\pi t}{12}\right)$$

is a monthly periodic function for estimating seasonal patterns, where η_1 and η_2 are real numbers. By using this model, it is assumed that the terms ε_t, $t = 1,\ldots,n$, are independent and follow the normal distribution with mean 0 and variance depending on the change-point θ, or say,

$$\varepsilon_t \sim N\left(0, \sigma_1^2 I_{[1,\theta]}(t) + \sigma_2^2 I_{(\theta,n]}(t)\right).$$

Thus, σ_1^2 and σ_2^2 are the variances of ε_t before and after the change-point, respectively. This model formulation corresponds to the following likelihood function:

$$f(\mathbf{y}|\theta,\boldsymbol{\xi}) = \prod_{t=1}^{n} (2\pi\lambda_t)^{-\frac{1}{2}} \exp\left[-\sum_{t=1}^{n} \frac{(y_t - \mu_t)^2}{2\lambda_t}\right],$$

where

$$\lambda_t = \sigma_1^2 I_{[1,\theta]}(t) + \sigma_2^2 I_{(\theta,n]}(t), \tag{1}$$

$$\mu_t = \alpha + g_1 I_{[1,\theta]}(t) + g_2 I_{(\theta,n]}(t) + S_t, \tag{2}$$

$\mathbf{y} = (y_1, y_2, \ldots, y_n)^\intercal$ and $\boldsymbol{\xi} = (\alpha, \beta_1, \beta_2, \gamma_{1,1}, \ldots, \gamma_{1,p}, \gamma_{2,1}, \ldots, \gamma_{2,p}, \psi_{1,1}, \ldots, \psi_{1,R}, \psi_{2,1}, \ldots, \psi_{2,R}, \sigma_1^2, \sigma_2^2, \eta_1, \eta_2)^\intercal$ is the vector of parameters. By definition, $f(\mathbf{y}|\theta,\boldsymbol{\xi})$ denotes the joint probability density function of the sample $Y = (Y_1, Y_2, \ldots, Y_n)$. In the Bayesian analysis, it is assumed that the parameters of the vector $\boldsymbol{\xi}$ and θ have distributions based on previous knowledge (the prior distributions), which are updated by using the data (represented by $f(\mathbf{y}|\theta,\boldsymbol{\xi})$) to produce the posterior distributions. This is formalized by the Bayes' theorem, given by $f(\theta,\boldsymbol{\xi}|\mathbf{y}) \propto f(\mathbf{y}|\theta,\boldsymbol{\xi})\, p(\theta,\boldsymbol{\xi})$, where $p(\theta,\boldsymbol{\xi})$ is the joint prior distribution and $f(\theta,\boldsymbol{\xi}|\mathbf{y})$ is the joint posterior distribution. The prior distributions can be "non-informative", with little effect on the posterior distribution. Thus, the following prior distributions for the parameters of the vector $\boldsymbol{\xi}$ are considered: $\alpha \sim N(0, c_1)$, $\beta_1 \sim N(0, c_2)$, $\beta_2 \sim N(0, c_3)$, $\gamma_{1,j} \sim N(0, c_{4,j})$, $\gamma_{2,j} \sim N(0, c_{5,j})$, $j = 1, \ldots, p$, $\eta_1 \sim N(0, c_6)$, $\eta_2 \sim N(0, c_7)$, $\psi_{1,r} \sim N(0, c_{8,r})$, $\psi_{2,r} \sim N(0, c_{9,r})$, $r = 1, \ldots, R$, $\sigma_1^2 \sim IG(c_{10}, c_{11})$ and $\sigma_2^2 \sim IG(c_{12}, c_{13})$, where $c_1, \ldots, c_{13}$ are known values for the hyperparameters of the prior distributions, $N(0, c)$ denotes a normal distribution with mean 0 and variance c, and $IG(h_1, h_2)$ denotes an inverse gamma distribution with mean $h_2/(h_1 - 1)$ and variance $h_2^2/[(h_1 - 1)^2(h_1 - 2)]$. Large values of $c_1, \ldots, c_{13}$ yield non-informative prior distributions for their respective parameters. It is further assumed prior independence among these parameters. In addition, it is assumed a categorical prior distribution for the change-point θ such that the prior probabilities of the values $1, 2, \ldots, n$ are assumed to be equal to $1/n$.

Alternatively, it can be considered that the terms ε_t follow a non-standardized Student's t-distribution with υ degrees of freedom, a location parameter μ_t, a scale parameter λ_t and variance $\lambda_t^2 \upsilon (\upsilon - 2)^{-1}$ for $\upsilon > 2$. In this case, the model formulation corresponds to the following likelihood function:

$$f(\mathbf{y}|\theta, \upsilon, \boldsymbol{\xi}) = \prod_{t=1}^{n} \left\{ \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\Gamma\left(\frac{\upsilon}{2}\right) \lambda_t \sqrt{\pi \upsilon}} \left[1 + \frac{1}{\upsilon} \left(\frac{y_t - \mu_t}{\lambda_t} \right)^2 \right]^{-\frac{\upsilon}{2}} \right\},$$

where $\Gamma(\cdot)$ is the gamma function, and λ_t and μ_t are given by (1) and (2), respectively. For the Bayesian analysis, one can consider the same prior distributions assumed for the previous model and a continuous uniform prior distribution for υ, or say, $\upsilon \sim U(2, c_\upsilon)$, where c_υ is a known hyperparameter ($c_\upsilon > 2$). In order to perform a brief sensitivity analysis, we have also considered fixed values for υ.

A Bayesian Markov Chain Monte Carlo (MCMC) procedure using the Gibbs sampler algorithm (Casella and George, 1992) was used to estimate the posterior distributions of the parameters of interest and variance components (Carlin and Louis, 1996). The Gibbs sampler algorithm was run for 510,000 iterations and sampled in every 10th simulation. To eliminate the effect of the initial values, the first 10,000 iterations were discarded as a "burn-in-sample". In this way, 50,000 final Gibbs samples were used for inferences. The 95% credible intervals ($95\%CI$) were obtained from the 2.5% and 97.5% percentiles of the posterior samples of the parameters. The 95% credible inter-

vals are the Bayesian equivalent of the traditional 95% confidence intervals, expressing the central 95% of the range of values that are credible for the respective estimated parameter. Usual diagnostic methods were employed to check the convergence of the MCMC calculations (Carlin and Louis, 1996). After the model fitting, the assumption of independence between the successive random error terms ε_t was graphically verified by plotting their respective autocorrelation and partial autocorrelation functions in relation to different lags. The estimation was performed by using the MCMC algorithm implemented in the freely available OpenBUGS software (Lunn et al., 2000). The OpenBUGS code used for this analysis is given in the Appendix A.

2.3. Model specifications

In the absence of covariates, three different models were fitted to the data as described below.

- **Model 1:** In this model, the autoregressive terms $\gamma_{1,1},\ldots,\gamma_{1,p},\gamma_{2,1},\ldots,\gamma_{2,p}$ were discarded, and consequently, $g_k = \beta_k(t-\theta)$, for $k = 1,2$. This model does not consider the presence of covariates, nor the monthly periodic function S_t.
- **Model 2:** This model is similar to Model 1 but it considers the autoregressive terms $\gamma_{1,1},\ldots,\gamma_{1,p},\gamma_{2,1},\ldots,\gamma_{2,p}$.
- **Model 3:** This model is similar to Model 2 but it considers the monthly periodic function S_t.

Models 1 to 3 were fitted based on the assumption that the residuals ε_t follow a normal distribution or a Student's t-distribution. In addition, Model 4 is defined as follows:

- **Model 4:** This model is similar to Model 3, but it includes an independent variable. The following variables were considered: average monthly temperature (^{o}C), maximum and minimum monthly temperature (^{o}C) and average monthly precipitation (mm). Due to its highly skewed distribution, a log transformation was applied to the measures of average monthly precipitation. These variables were selected due to their known effects on the asthma admissions in various populations (Ivey, Simeon and Monteil, 2003; Chen, Xirasagar and Lin, 2006). Under this formulation, four different models were fitted to the data, one for each independent variable, thus avoiding problems of collinearity between variables.

2.4. Model selection

The deviance information criterion (DIC) is widely used for Bayesian model comparison (Spiegelhalter et al., 2014). However, the proposed model is interpreted by the OpenBUGS as a mixture model, and this software is not able to calculate the DIC value

in this situation. Another criterion for model selection is derived from the conditional predictive ordinate (CPO) statistics (Gelfand, Dey and Chang, 1992). For the i-th observation, the CPO_i is given by

$$f\left(\mathscr{D}_i|\mathbf{y}_{[i]}\right) = \int f\left(\mathscr{D}_i|\Theta\right) f\left(\Theta|\mathscr{D}_{[i]}\right) d\Theta,$$

where Θ is the complete vector of parameters, $\mathscr{D}_i$ is each instance of all data $\mathscr{D}$, $\mathscr{D}_{[i]}$ is $\mathscr{D}$ without the current observation i and $f\left(\Theta|\mathscr{D}_i\right)$ is the posterior density of Θ given $\mathscr{D}_{[i]}$, $i = 1,\ldots,n$. Thus, the CPO statistics expresses the posterior probability of observing the value or set of values of $\mathscr{D}_i$ when the model is fitted to all data except $\mathscr{D}_i$. A MCMC approximation of CPO_i (Chen, Shao and Ibrahim, 2000) is given by

$$\widehat{CPO}_i = \left[\frac{1}{B}\sum_{b=1}^{B}\frac{1}{f\left(\mathscr{D}_i|\Theta_b\right)}\right]^{-1}$$

where B is the number of iterations during implementation of the MCMC procedure after the burn-in period and Θ_b is the vector of the samples obtained at the b-th iteration. Thus, approximate CPO statistics can be directly computed with OpenBUGS by defining nodes for $f\left(\mathscr{D}_i|\Theta_b\right)^{-1}$. Assuming approximate normality, inverse values for $\widehat{CPO}_i$ larger than 40 can be considered as possible outliers and higher than 70 as extreme values (Ntzoufras, 2009). The log pseudo marginal likelihood (LPML) is a Bayesian measure of fit or adequacy which is defined based on the CPO statistics (Geisser and Eddy, 1979). For a given model, the LPML value is given by $\widehat{LPML} = \sum_{i=1}^{n}\log\widehat{CPO}_i$.

The larger is the value of LPML, the better is the fit of the model. The corresponding pseudo Bayes' factor (PBF) comparing models m and $m\prime$ is

$$PBF_{mm'} = \exp\left(\widehat{LPML}_m - \widehat{LPML}_{m'}\right).$$

In addition, the discrepancy between the data and an estimation model can be measured by the sum of squared residuals (SSR) given by

$$SSR = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}\left(y_i - \widehat{\mu}_t\right)^2,$$

where $\widehat{\mu}_t$ is obtained by replacing the parameters in (2) by their respective estimates. For fits of different models to a given dataset, a smaller SSR value indicates a better fit to the data.

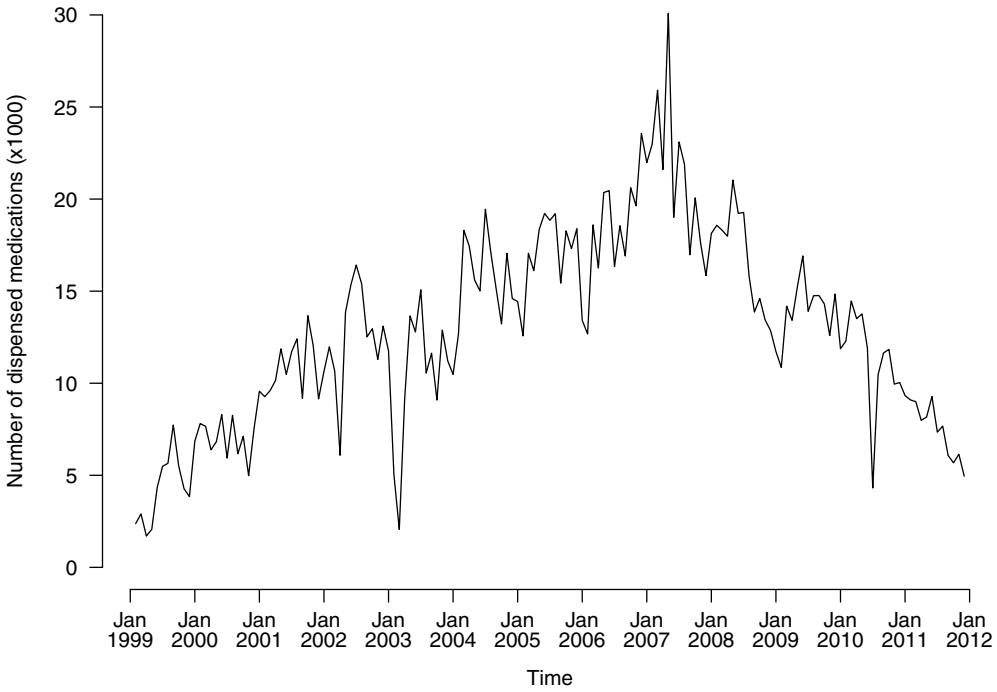

Figure 1: *Monthly number of dispensed medications (salbutamol capsule, 2mg) by the public health services of Ribeirão Preto, Brazil, from February 1999 to December 2011.*

3. Results

The graph in Figure 1 shows a time series of the number of monthly dispensations of the salbutamol from February 1999 to December 2011. This graph gives evidence of the presence of a change-point for the time series. It is observed an increase in the number of dispensations of the medication at the beginning of the considered period and a great reduction in March 2003, probably due to the short period when the drug was missing, followed by a further increase in the number of dispensations, until May, 2007. Thereafter, it is observed that the number of dispensed medications decreases until the end of the period of observation. A great reduction in the number of dispensations was also observed in July 2010. However, in the analysis of these data, only one change-point in the time series will be considered. The reductions in the number of dispensations observed in March 2003 and July 2010 will be treated as months with atypical numbers of dispensed medications, instead of instants in which the trend behaviour of the series has been modified.

In the Bayesian analysis, non-informative prior distributions were considered for all parameters of the model. In this way, it was considered that $c_1 = c_2 = c_3 = c_{4,j} = c_{5,j} = c_6 = c_7 = c_{8,r} = c_{9,r} = 10^6$, $j = 1, \ldots, p$, $r = 1, \ldots, R$, in the prior distributions for α, β_1,

β_2, $\gamma_{1,j}$, $\gamma_{2,j}$, $\psi_{1,r}$ and $\psi_{2,r}$, and $c_{10} = c_{11} = c_{12} = c_{13} = 0.1$ in the prior distributions for σ_1^2 and σ_2^2. In the case of the model with Student-t errors, it is also considered $c_\upsilon = 50$, or say, $\upsilon \sim U(2,50)$. The number of monthly dispensed medications was divided by 1,000 in order to facilitate the convergence of the computational algorithm.

Tables 2 and 3 show the results for the Models 1 to 3 obtained by using the OpenBUGS software. The results in Table 2 consider that the residuals of the models follow a normal distribution, while Table 3 shows results from models with residuals that follow a Student's t-distribution with υ degrees of freedom. In the case of the Models 2 and 3, they were fitted considering one, two or more autoregressive orders, but it was observed that models with order p equal to or greater 2 did not improve the goodness of fit. Thus, we considered $p = 1$ in all the cases.

Table 2 shows that the results of Model 3 have the lowest SSR value and the highest LPML value, suggesting that this model provides the best fit to the data among these three models. The *PBF* value comparing the Models 3 and 2 is 6.05. In all the fitted models, the estimates for β_1 are positive and the estimates for β_2 are negative, showing that the number of dispensed medications is increasing over time until the change-point θ is reached, but decreasing from this value. The 95% credible intervals for $\gamma_{1,1}$ and $\gamma_{2,1}$ do not contain the value zero, evidencing the significance of the autoregressive parameters of order $p = 1$. The results of Model 3 also show that the 95% credible interval for η_2 do not contain the value zero, suggesting the evidence of a yearly seasonal pattern in the series.

Table 2: *Results from the Bayesian change-point statistical models, with residuals following a normal distribution.*

Parameter	Model 1		Model 2		Model 3	
	Bayesian estimate	95% credible interval	Bayesian estimate	95% credible interval	Bayesian estimate	95% credible interval
α	20.80	(20.01 , 21.59)	17.67	(16.18 , 19.13)	18.11	(16.61 , 19.58)
β_1	0.166	(0.149 , 0.182)	0.094	(0.062 , 0.125)	0.104	(0.072 , 0.135)
β_2	−0.268	(−0.297 , −0.240)	−0.178	(0.230 , −0.127)	−0.191	(−0.242 , −0.141)
θ	100.5	(99.0 , 103.0)	102.3	(100.0 , 107.0)	101.9	(100.0 , 107.0)
$\gamma_{1,1}$	–	–	0.444	(0.282 , 0.607)	0.401	(0.237 , 0.564)
$\gamma_{2,1}$	–	–	0.295	(0.099 , 0.500)	0.236	(0.041 , 0.439)
σ_1^2	7.82	(5.88 , 10.36)	6.51	(4.87 , 8.65)	6.24	(4.64 , 8.29)
σ_2^2	4.09	(2.73 , 6.16)	3.57	(2.32 , 5.53)	3.40	(2.22 , 5.22)
η_1	–	–	–	–	0.464	(−0.024 , 0.954)
η_2	–	–	–	–	−0.585	(−1.107 , −0.068)
LPML	−364.8		−354.7		−352.9	
SSR	967.2		805.2		758.4	

Table 3: *Results from the Bayesian change-point statistical models, with residuals following a Student's t-distribution with υ degrees of freedom.*

Parameter	Model 1		Model 2		Model 3	
	Bayesian estimate	95% credible interval	Bayesian estimate	95% credible interval	Bayesian estimate	95% credible interval
α	20.88	(20.11 , 21.64)	17.76	(16.26 , 19.19)	18.19	(16.75 , 19.57)
β_1	0.166	(0.150 , 0.183)	0.096	(0.064 , 0.128)	0.104	(0.073 , 0.135)
β_2	-0.263	(-0.292 , -0.235)	-0.177	(-0.228 , -0.124)	-0.190	(-0.237 , -0.141)
θ	99.9	(97.0 , 102.0)	101.9	(95.0 , 107.0)	102.0	(97.0 , 106.0)
υ (df)	11.1	(3.1 , 40.6)	13.2	(3.4 , 43.8)	9.2	(2.9 , 34.7)
$\gamma_{1,1}$	–	–	0.423	(0.258 , 0.590)	0.383	(0.220 , 0.543)
$\gamma_{2,1}$	–	–	0.303	(0.104 , 0.510)	0.244	(0.063 , 0.437)
σ_1^2	5.40	(3.22 , 8.22)	4.90	(2.94 , 7.23)	4.35	(2.62 , 6.59)
σ_2^2	3.11	(1.70 , 5.16)	2.68	(1.43 , 4.56)	2.16	(1.13 , 3.76)
$\sigma_1^2\upsilon(\upsilon-2)^{-1}$	7.85	(5.27 , 12.29)	6.60	(4.53 , 9.82)	6.69	(4.38 , 10.70)
$\sigma_2^2\upsilon(\upsilon-2)^{-1}$	4.55	(2.65 , 8.00)	3.63	(2.07 , 6.49)	3.35	(1.81 , 6.20)
η_1	–	–	–	–	0.510	(0.062 , 0.951)
η_2	–	–	–	–	-0.631	(-1.091 , -0.161)
LPML	-472.8		-442.8		-465.4	
SSR	968.0		808.8		766.9	

The results in Table 3 indicate that the estimates obtained from the models with residuals following a Student's t-distribution are close to those found when considering a normal distribution (Table 2). The graphs in Figure 2 illustrate the simulated posterior Gibbs samples for the change-point in each of the three assumed models. In addition, plots of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the residuals of the Models 1 to 3 are shown in Appendix B. The ACF and PACF of residuals of the Models 2 and 3 at different lag times were not significantly different from zero. From equation (1), the variances of ε_t before and after the change-point are given by $\sigma_1^2\upsilon(\upsilon-2)^{-1}$ and $\sigma_2^2\upsilon(\upsilon-2)^{-1}$, respectively. Estimators for these quantities are also presented in Table 3, and we can note that they are very similar to those for σ_1^2 and σ_2^2 obtained from the fit of the models based on the normal distribution (Table 2). Alternatively, we also considered models based on the Student's t-distribution with fixed values for υ ranging from 2 to 50. For each possible choice of υ, we obtained the correspondent values for LPML and SSR considering the Models 1 to 3 (results not shown in the tables). We did not find important differences when compared the LPML and SSR values obtained from models with fixed values for υ ranging from 2 to 50. However, we noted a better fit to the data (i.e. higher LPML values and lower SSR values) for values relatively higher for υ, such as $\upsilon = 100$ or $\upsilon = 200$, thus suggesting that models based on the normal distribution can be more adequate for the monthly number of dispensed medications.

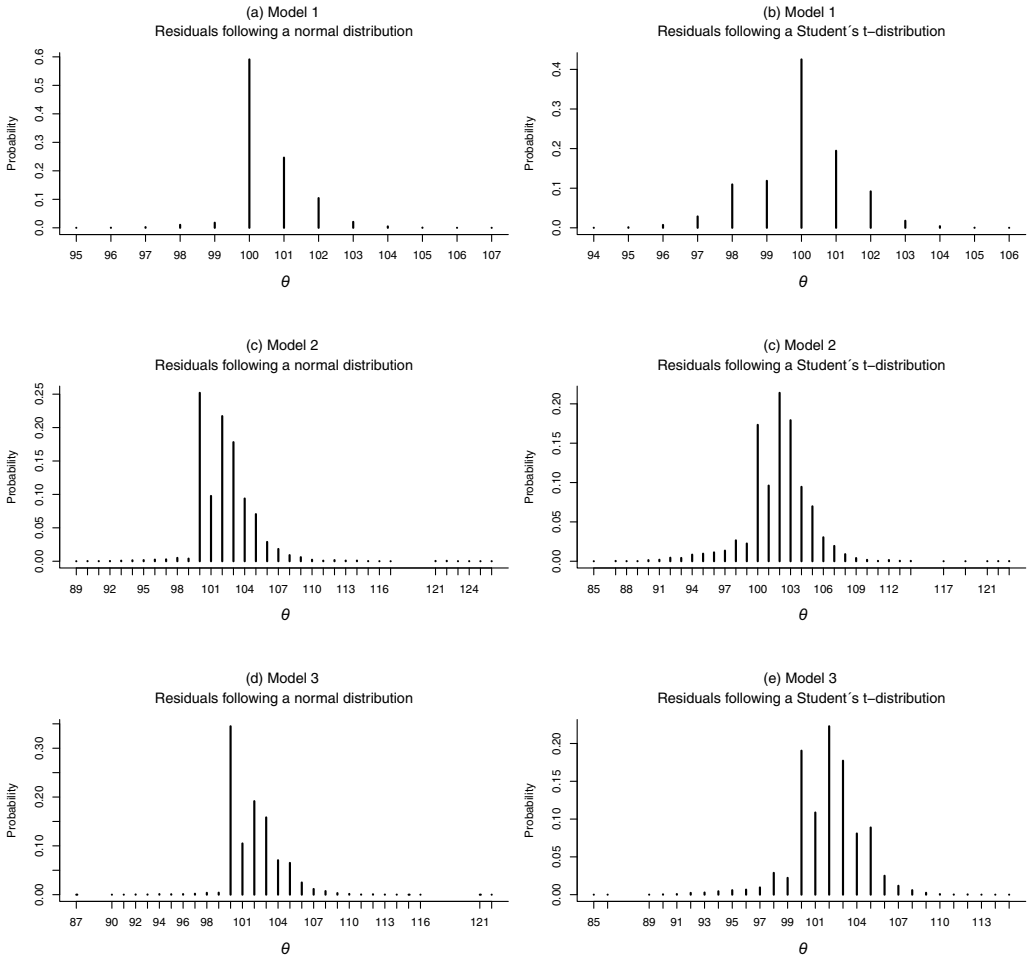

Figure 2: *Plots of the simulated posterior Gibbs samples for the change-point in each of the three assumed models.*

The upper panel of the Figure 3 shows the observed number of dispensed medications and the predicted values obtained from Model 1. Considering the results from the Model 1 with residuals that follow a normal distribution, the predicted values linearly increase up to the change-point θ estimated by $t = 100.5$ (Table 2), corresponding to the month of May, 2007, with a 95% credible interval ranging from April 2007 to August 2007. After this change-point, the predicted values linearly decrease with the coefficient β_2 estimated by -0.268. Considering the Model 1 with residuals that follow a Student's t-distribution, the change-point θ is estimated by $t = 99.9$ (Table 3). However, autocorrelation plots (not shown) for the residuals from Model 1 evidence significant serial correlation between successive values of ε_t, that is, the assumption of independence between the residuals was not attained. Therefore, Model 1 is useful to describe the linear trend of the time series before and after the change-point, but inferences for the

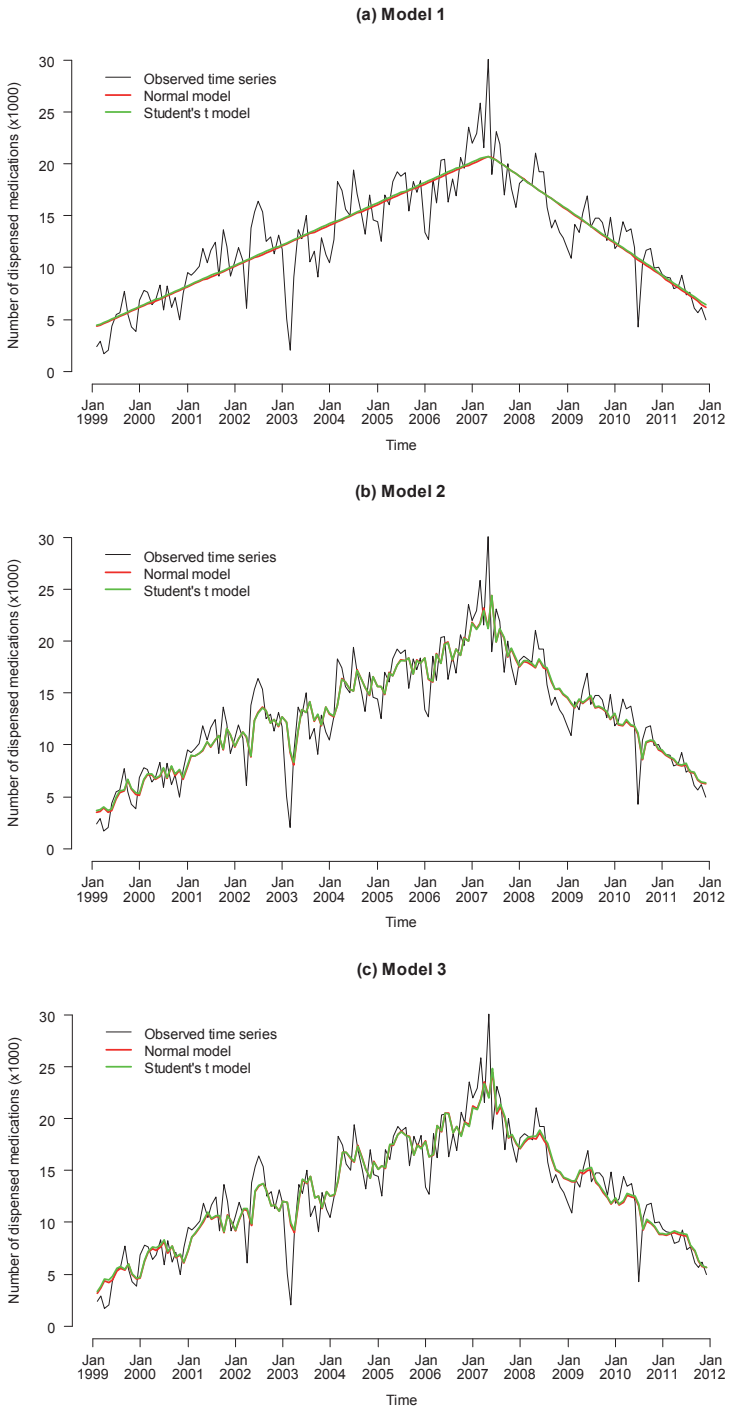

Figure 3: *Comparison between the observed time series and the time series estimated from the Models 1 to 3, with residuals following a normal distribution and a Student's t-distribution with υ degrees of freedom.*

parameters of the model can be harmed from this lack of independence for the residuals. Results from the fit of the Models 2 and 3, considering autoregressive terms of order 1, are also shown in Tables 2 and 3 and visualized in the Figure 3. Considering the fit with residuals that follow a normal distribution, the change-point was now estimated by $t = 102.3$ (Table 2), corresponding to the month of July, 2007, with a 95% credible interval ranging from May, 2007, to November, 2007. Autocorrelation plots (not shown) for the residuals from Models 2 and 3 did not evidence significant serial correlation between successive values of ε_t, indicating a good fit of the model to the data. The central and lower panels of the Figure 3 show the predicted values obtained from Models 2 and 3, respectively. In both models, the estimate for the variance σ_1^2 was greater than the estimate for σ_2^2, suggesting a higher dispersion of the number of dispensed medications before the change-point.

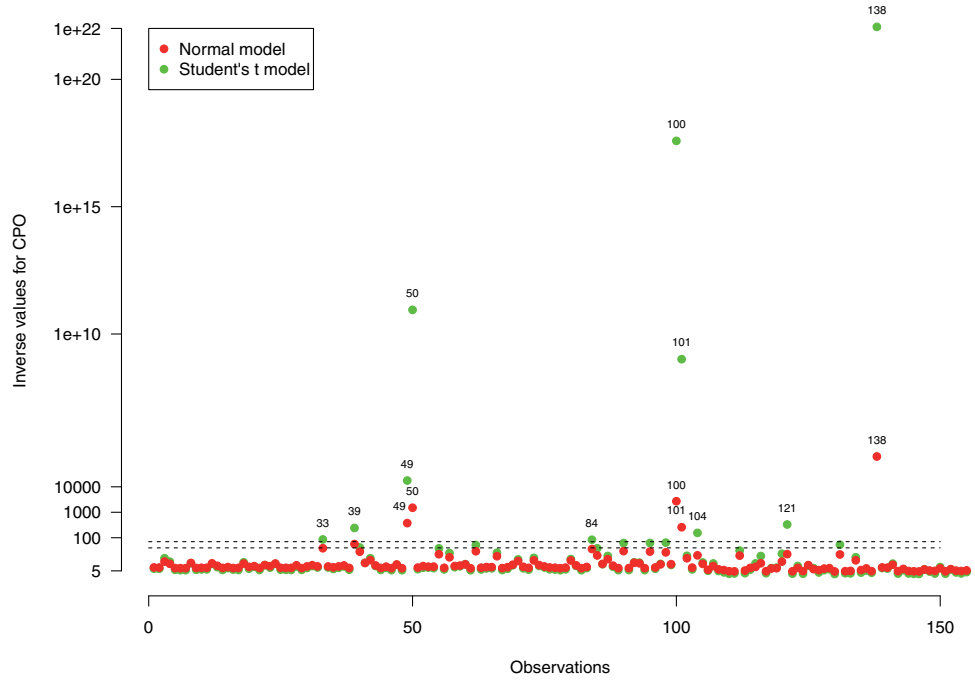

Figure 4: *Comparison between inverse values for CPO obtained from the Model 3 considering residuals following a normal distribution and a Student's t-distribution with υ degrees of freedom. The horizontal dashed lines pass through the values 40 and 70, identifying possible outliers and extreme values, respectively.*

Examination of a plot of inverse values for CPO values can identify possible outliers in the model fitting, thus allowing for comparisons between models. Considering the results from the Model 3, the graph in Figure 4 corresponds to the plot of inverse values for CPO, where the two horizontal dashed lines in the figure pass through the values 40 and 70, identifying possible outliers and extreme values, respectively (see Subsection 2.4). This graph compares the inverse values for CPO obtained from the Model 3 con-

sidering residuals following a normal distribution and a Student's t-distribution with υ degrees of freedom. We can note that the number of the extreme values is greater when considering the model with residuals following a Student's t-distribution, thus reinforcing that the model with errors following a normal distribution is the model that best fits to the data.

We also fitted alternative models that do not take into account the presence of a change-point, but consider the presence of autoregressive effects of high order. However, we observed that these models did not fit well to the data. For example, for models with residuals following a normal distribution and autoregressive effects of orders 4 and 5, we obtained LPML values given by -357.9 and -354.9, respectively, and SSR values given by 891.2 and 866.4, respectively. In addition, for models with residuals following a Student's t-distribution and autoregressive effects of orders 4 and 5, we obtained LPML values given by -436.7 and -435.4, respectively, and SSR values given by 892.6 and 868.1, respectively.

Table 4: *Results from regression models that considers the monthly maximum and minimum absolute temperatures as independent variables.*

Parameter	Maximum absolute temperature (^oC)		Minimum absolute temperature (^oC)	
	Bayesian estimate	95% credible interval	Bayesian estimate	95% credible interval
α	18.41	(16.85 , 19.94)	18.1	(16.58 , 19.58)
β_1	0.110	(0.076 , 0.142)	0.103	(0.071 , 0.135)
β_2	-0.199	(-0.254 , -0.147)	-0.191	(-0.243 , -0.140)
θ	101.4	(100.0 , 106.0)	102.0	(100.0 , 107.0)
$\gamma_{1,1}$	0.384	(0.218 , 0.548)	0.403	(0.238 , 0.566)
$\gamma_{2,1}$	0.202	(0.005 , 0.413)	0.241	(0.042 , 0.447)
σ_1^2	6.13	(4.53 , 8.19)	6.34	(4.71 , 8.47)
σ_2^2	3.51	(2.29 , 5.37)	3.42	(2.22 , 5.29)
η_1	0.164	(-0.497 , 0.814)	0.453	(-0.044 , 0.944)
η_2	-0.413	(-0.985 , 0.164)	-0.433	(-1.42 , 0.551)
ψ_1	-0.195	(-0.470 , 0.084)	-0.025	(-0.196 , 0.145)
ψ_2	-0.132	(-0.436 , 0.174)	-0.030	(-0.207 , 0.147)
LPML		-353.8		-354.9
SSR		740.6		759.5

Tables 4 and 5 show the results from regression models (Model 4) in which monthly maximum and minimum absolute temperatures, average monthly temperature and monthly average precipitation are independent variables. These models assume that the residuals follow a normal distribution. For all these independent variables, we can observe that the 95% credible intervals for the parameters ψ_1 and ψ_2 include the value zero. This implies that we do not have evidence that these climatic variables are as-

Table 5: *Results from regression models that considers the monthly average temperature and precipitation as independent variables.*

Parameter	Monthly average temperature (^oC)		Precipitation (log mm)	
	Bayesian estimate	95% credible interval	Bayesian estimate	95% credible interval
α	18.14	(16.62 , 19.62)	18.1	(16.598 , 19.58)
β_1	0.104	(0.071 , 0.136)	0.103	(0.071 , 0.135)
β_2	−0.193	(−0.245 , −0.142)	−0.192	(−0.245 , −0.140)
θ	101.8	(100.0 , 107.0)	102.0	(100.0 , 107.0)
$\gamma_{1,1}$	0.401	(0.236 , 0.563)	0.403	(0.239 , 0.566)
$\gamma_{2,1}$	0.235	(0.038 , 0.438)	0.234	(0.032 , 0.442)
σ_1^2	6.26	(4.62 , 8.41)	6.26	(4.66 , 8.33)
σ_2^2	3.41	(2.22 , 5.28)	3.50	(2.26 , 5.48)
η_1	0.321	(−0.223 , 0.863)	0.462	(−0.035 , 0.957)
η_2	−0.142	(−1.053 , 0.758)	−0.718	(−1.391 , −0.035)
ψ_1	−0.204	(−0.551 , 0.146)	0.061	(−0.164 , 0.287)
ψ_2	−0.164	(−0.517 , 0.191)	0.054	(−0.185 , 0.296)
LPML	−354.1		−354.8	
SSR	751.1		755.9	

sociated with the monthly number of dispensed medications. These results can be still observed even in similar models that not include the seasonal component S_t. Plots of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the residuals from these models are shown in Appendix B (Figure 7). The ACF and PACF of residuals at different lag times were not significantly different from zero.

4. Discussion

Statistical methods of time series analysis are widely used in public health studies (Zeger, Irizarry and Peng, 2006; Jornet-Sanz et al., 2017). These methods are useful for detecting outbreaks, monitoring the occurrence of a disease at a regional level, analysing epidemiological surveillance data, describing the seasonality of infectious diseases, examining how climate change can affect the disease occurrence over time, and predicting future scenarios of an event of interest. In the present article, we introduced a Bayesian approach that can be used to estimate a change-point model with autoregressive terms. In the context of the monthly number of dispensed asthma medications, this model is useful to provide a better understanding of the corresponding time series, such as seasonal patterns, dependence on previous times and possible association with climatic variables.

Alternatively to the method presented here, a maximum likelihood estimate of the change-point θ can be obtained by using the profile likelihood approach. In this case, the profile likelihood $\ell_p(\theta)$ for θ is defined by maximizing the likelihood function with respect to all the other parameters in the model for a range of values for θ over which the profile likelihood is to be evaluated. Thus, the maximum likelihood estimate for the change-point is

$$\widehat{\theta}_{ML} = \arg_\theta \max \ell_p(\theta) = \arg_\theta \max \sum_{t=1}^{n} \ln f(y_t|\theta, \widehat{\boldsymbol{\xi}}_{ML}),$$

where $\widehat{\boldsymbol{\xi}}_{ML}$ is the vector of maximum likelihood estimates for the other parameters associated with the model. Although it is possible to implement a computer algorithm in order to find the maximum likelihood estimate for θ, presentation of this analysis is out of the scope of the present paper. We opted for the use of Bayesian methods, that make it easy to incorporate prior knowledge about the change-point value. In addition, Bayesian estimation is facilitated using the OpenBUGS software, that only requires the specification of the distribution for the data and the prior distributions for the parameters.

As previously mentioned, the present statistical model was developed using a time series of the dispensation of salbutamol sulfate tablets 2 mg. Currently, this presentation form of salbutamol sulfate is no longer considered the most appropriate because it is associated with a higher number of side effects when compared to other forms, such as the oral spray (Sociedade Brasileira de Pneumologia e Tisiologia, 2012). Salbutamol sulfate is also indicated for the treatment of other diseases such as chronic obstructive pulmonary disease (COPD), preferably via inhalation. This drug can still be used in some other situations, such as inhibition of uncomplicated premature labor in the last gestational trimester, in which oral administration is the preferable choice (Motazedian et al., 2010).

The graphs in Figure 2 shows that the behavior of the time series for the number of monthly dispensations of the salbutamol sulfate tablets of 2 mg from 1999 to 2011 is interpreted in terms of the presence of a change-point. The Brazilian National Drug Policy, introduced in October 30th 1998, established new guidelines for pharmaceutical care in the public health by defining, among other things, a list of essential medications according to the most common health problems reported in the population. Thereafter, and with the decentralization process of drugs distribution for the states and cities (Ordinance GM 176 of March 8th, 1999), it was possible to expand the supply of medications in public health network (Botega and Santos, 2007), which explains the increased number of dispensations of the medication at beginning of year 1999 (Figure 3). Fluctuations in the monthly number of medication dispensed are observed in the Figure 2), with some seasonality. It was hypothesised that local maximum points in the time series are coincident with colder and drier periods, when the airway infections and episodes of bronchospasm occur more frequently, creating a greater demand for the use of the medication in specific periods of the year (Thomazelli et al., 2007; Peterson et al., 2012).

However, the regression models used in this study do not show a significant association between climatic variables and the number of dispensed medication.

Figure 1 shows a great reduction in the dispensation of salbutamol by March 2003 and July 2010, suggesting a period of discontinuity in the supply of the medication. Despite the good results of the decentralization policy of the pharmaceutical care, the provision of essential medications and medicines in some special situations (drugs being part of specialized pharmaceutical care) depended largely on efforts of the Brazilian National Health System managers (Botega and Santos, 2007), which leads to unavailability of the medicament to the population. The reduction in March 2003 was followed by a further increase in the number of dispensations, which as observed until the month of May, 2007. In 2004, as part of the National Policy on Integral Health Care of People with Respiratory Diseases, the cities with primary healthcare services began receiving beclomethasone 250 mcg oral spray, beclomethasone 50 mcg nasal spray and salbutamol 100 mcg oral spray from the Brazilian Ministry of Health for treatment of both asthma and allergic rhinitis, and given that asthma and allergic rhinitis often co-exist in the same individual, the control of one of these diseases favors the control of another, thus contributing to the implementation of better health practices for asthma. At that moment, aminophylline 100mg tablets were being provided by National Health System and now they are no longer supplied, being replaced by salbutamol spray (Botega and Santos, 2007). Figure 2 shows a further reduction in the number of dispensations of salbutamol sulfate from 2007 to the end of 2011, characterizing a change-point that probably reflects the improvement of healthcare provided to patients with asthma when the new medications were introduced.

As a final consideration, the article provides suggestions for future investigations:

(a) Possible extensions of the model in order to accommodate more than one change-point should be considered in future research works.
(b) In the proposed model, we assumed constant variances before and after the change-point. Future works can assume the effect of covariates on these variances, thus improving the fit of the proposed model.
(c) The actual numbers for asthma medication are huge, as discussed in this paper, and therefore the model assumptions are very reasonable. Extensions of the proposed model for low count data are essential for the analysis of a large broad of other epidemiological time series.
(d) By considering the data shown in Table 1, the change-point can be seen in the central part of the time series. Studies with simulated data can be useful to verify the performance of the proposed model in estimating the change-point when the period after change is short (or say, when there are few observations after the change).

Appendix A

The OpenBugs code used to specify the statistical model in its general form and with residuals following a normal distribution is given below. Observations of the independent variable are denoted by `x[t]`. In addition, `cp` denotes the change-point value and `N` is the length of the time series.

```
model
{
for(t in 1:N) {
y[t] ~ dnorm(mu[t], tau[J[t]])
mu[t] <- alpha + beta[J[t]]* (t-cp)
 + gama[J[t]]* (w[t] - mean(w[])) + St[t]
 + phi[J[t]]*(x[t] - mean(x[]))
k[t] <- step(t - cp - 0.5)
J[t] <- 1 + k[t]
punif[t] <- 1/N
St[t] <- eta[1]*sin(2*pi*t/12) + eta[2]*cos(2*pi*t/12)
# Likelihood function
L[t] <- 1/sqrt(2*pi*(pow(sigma[1],1-k[t])
 * pow(sigma[2],k[t])))
 * exp(-(y[t]-mu[t])*(y[t]-mu[t])
 /(2*(pow(sigma[1],1-k[t]) * pow(sigma[2],k[t]))))
# Inverse values for CPO
PO[t] <- 1/L[t]
}
for(i in 2:N) { w[i] <- y[i-1] }
w[1] <- y[1]
pi <- 3.14159265359
# Prior distributions
prec <- 1.0E-6
alpha ~ dnorm(0.0, prec)
cp ~ dcat(punif[])
for(j in 1:2) {
 beta[j] ~ dnorm(0.0, prec)
 eta[j] ~ dnorm(0.0, prec)
 gama[j] ~ dnorm(0.0, prec)
 phi[j] ~ dnorm(0.0, prec)
 tau[j] ~ dgamma(0.1,0.1)
 sigma[j] <- 1/tau[j]
}}
```

Appendix B

Figures 5 and 6 show autocorrelation functions (ACF) and partial autocorrelation functions (PACF) of the residuals of the Models 1, 2 and 3 based on the normal distribution (Figure 5) and Student's t-distribution (Figure 6). Dashed horizontal lines correspond to the significance boundaries for the non-zero terms. By comparing the plots in Figures 5 and 6, we can observe that the ACF and PACF functions from the models with residuals based on normal and Student's t-distributions are quite close one another. The Figures show that there was no significant autocorrelation between residuals at different lag times for the Models 2 and 3.

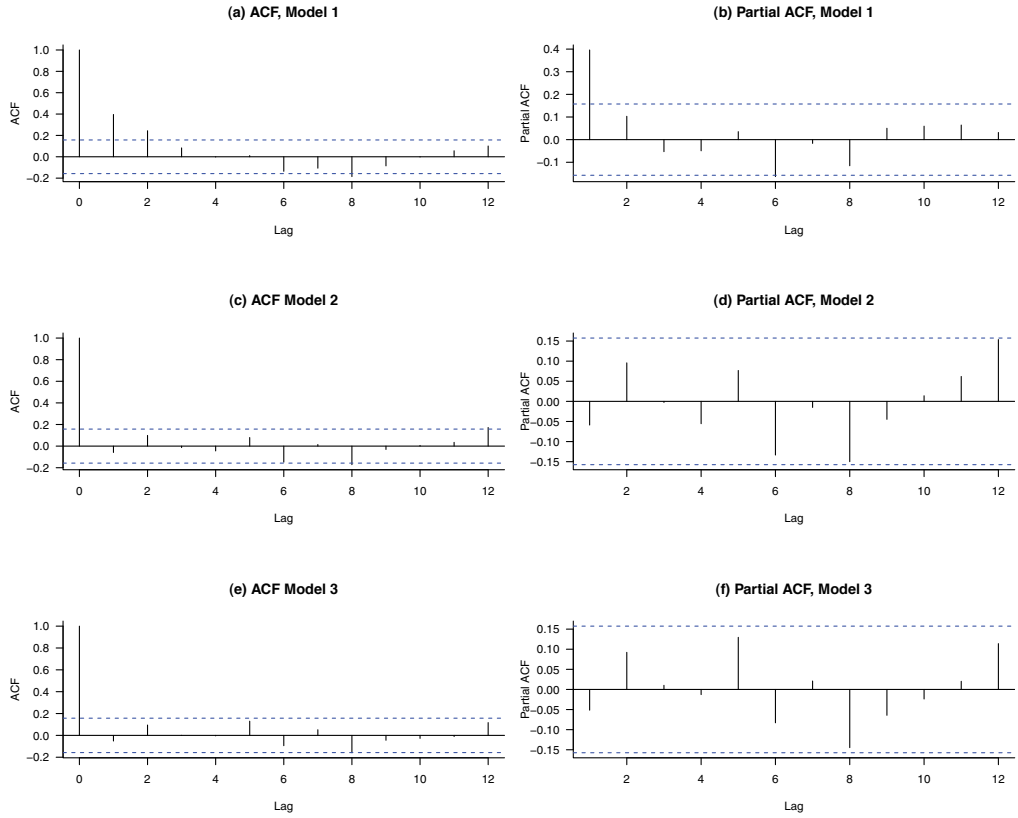

Figure 5: *Autocorrelation function (ACF) and partial ACF (PACF) plots for the residuals considering the Models 1, 2 and 3 based on the normal distribution. In each plot, two horizontal dashed lines denote two standard error limits of sample autocorrelation function.*

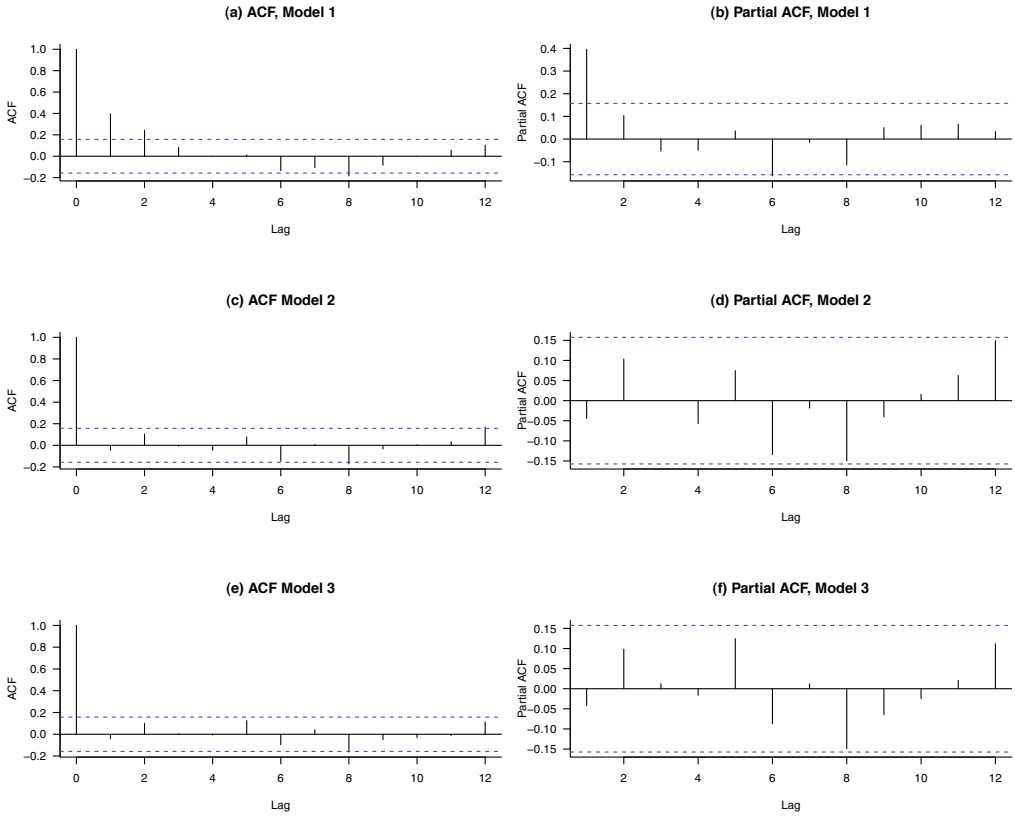

Figure 6: *Autocorrelation function (ACF) and partial ACF (PACF) plots for the residuals considering the Models 1, 2 and 3 based on the Student's t-distribution. In each plot, two horizontal dashed lines denote two standard error limits of sample autocorrelation function.*

Figure 7 shows ACF and PACF of the residuals of the Model 4, based on the normal distribution and including the climatic variables as independent variables. The plots show that there was no significant autocorrelation between residuals at different lag times.

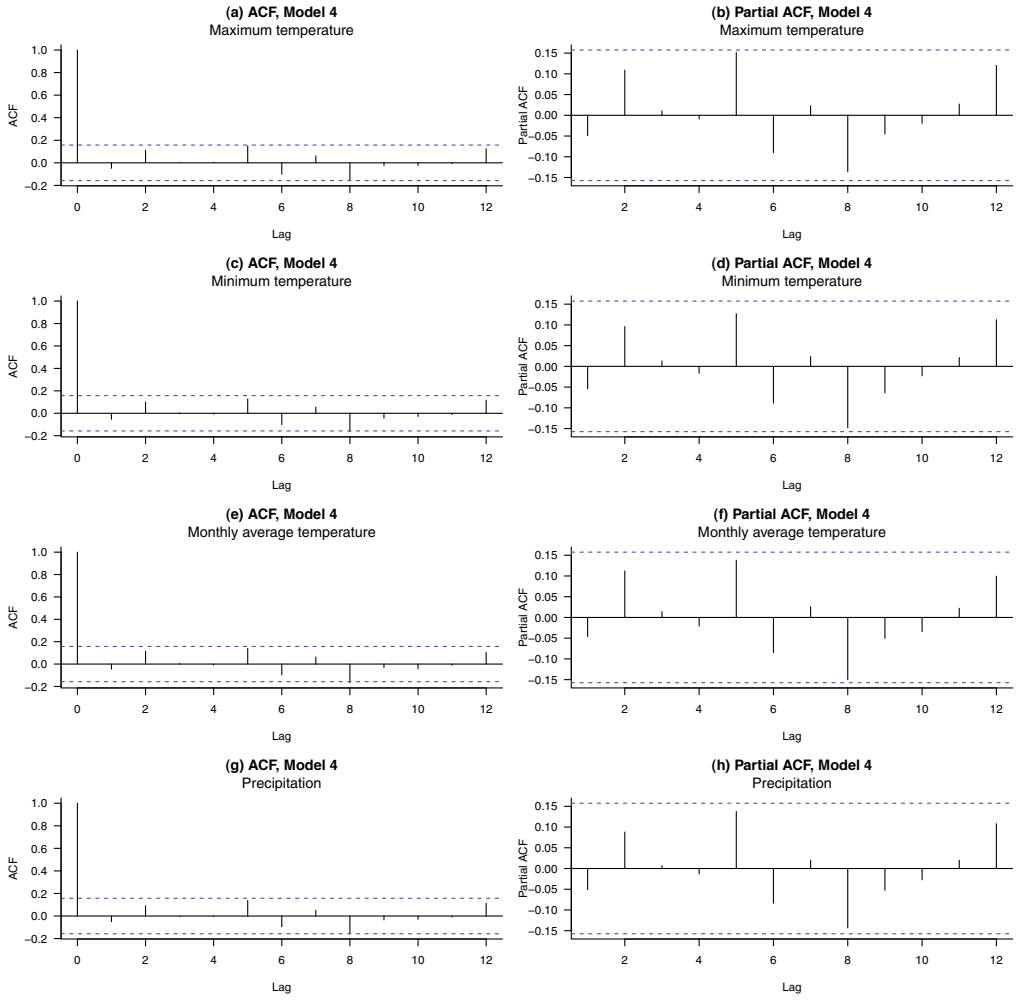

Figure 7: *Autocorrelation function (ACF) and partial ACF (PACF) plots for the residuals considering the Model 4 based on the normal distribution and the climatic variables included as independent variables. In each plot, two horizontal dashed lines denote two standard error limits of sample autocorrelation function.*

References

Achcar, J.A. and Loibel, S. (1998). Constant hazard models with a change-point: a Bayesian analysis using Markov Chain Monte Carlo methods. *Biometrical Journal*, 40, 543–555.

Achcar, J.A., Rodrigues, E.R., Paulino, C.D. and Soares, P. (2010). Non-homogeneous Poisson models with a change-point: an application to ozone peaks in Mexico city. *Environmental and Ecological Statistics*, 17, 521–541.

Achcar, J.A., Rodrigues, E.R. and Tzintzun, G. (2011). Using non-homogeneous Poisson models with multiple change-points to estimate the number of ozone exceedances in Mexico City. *Environmetrics*, 22, 1–12.

Achcar, J.A., Martinez, E.Z., Ruffino-Netto, A., Paulino, C.D. and Soares P. (2008). A statistical model investigating the prevalence of tuberculosis in New York City using counting processes with two change-points. *Epidemiology and Infection*, 136, 1599–1605.

Barry, D. and Hartigan, J.A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88, 309–319.

Bittar, O.J.N.V., Mendes, J.D.V. and Magalhães, A. (2011). *Rede Hospitalar no Estado de São Paulo: Mapear para Regular*. São Paulo: Secretaria de Estado da Saúde de São Paulo.

Botega, A. and Santos, M.R. (2007). *Descentralização das ações de assistência farmacêutica - asma e rinite*. In: Cerci-Neto, A. (org.) Asma em saúde pública. 1st. ed. São Paulo: Manole. p. 23–30.

Casella, G. and George, E.I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46, 167–174.

Carlin, B.P. and Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Chapman and Hall.

Carlin, B.P., Gelfand, A.E. and Smith, A.F.M. (1992). Hierarchical Bayesian analysis of change-point problems. *Journal of the Royal Statistical Society: Series C*, 41, 389–405.

Chen, M.H., Shao, Q.M. and Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Chen, C.H., Xirasagar, S. and Lin, H.C. (2006). Seasonality in adult asthma admissions, air pollutant levels, and climate: a population-based study. *Journal of Asthma*, 43, 287–292.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86, 221–241.

Dey, D.K. and Purkayastha, S. (1997). Bayesian approach to change-point problems. *Communications in Statistics - Theory and Methods*, 26, 2035–2047.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16, 203–213.

Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74, 153–160.

Gelfand, A.E., Dey, D.K. and Chang, H. (1992). *Model determination using predictive distributions with implementation via sampling-based methods*. In Bayesian Statistics 4. (Edited by J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith), 147–167, Oxford University Press.

Ghosh, P. and Vaida, F. (2007). Random change point modelling of HIV immunologic responses. *Statistics in Medicine*, 26, 2074–2087.

Hackl, P. (2012). *Statistical Analysis and Forecasting of Economic Structural Change*. New York: Springer.

Ivey, M.A., Simeon, D.T. and Monteil, M.A. (2003). Climatic variables are associated with seasonal acute asthma admissions to accident and emergency room facilities in Trinidad, West Indies. *Clinical & Experimental Allergy*, 33, 1526–1530.

Jensen, U. and Lautkebohmert, C. (2007). *Change-Point Models*. In: Ruggeri, F., Kenett, R. and Faltin, W. *Encyclopedia of Statistics in Quality and Reliability*. Vol. 1. New York: John Wiley & Sons.

Jornet-Sanz, M., Corberán-Vallet, A., Santonja, F. and Villanueva, R. (2017). A Bayesian stochastic SIRS model with a vaccination strategy for the analysis of respiratory syncytial virus. *SORT - Statistics and Operations Research Transactions*, 1, 159–176.

Lee, T.S. (2010). Change-point problems: bibliography and review. *Journal of Statistical Theory and Practice*, 4, 643–662.

Lavielle, M. and Lebarbier, E. (2001). An application of MCMC methods for the multiple change-points problem. *Signal Process*, 81, 39–53.

Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter D. (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.

Motazedian, S., Ghaffarpasand, F., Mojtahedi, K. and Asadi N. (2010). Terbutaline versus salbutamol for suppression of preterm labor: a randomized clinical trial. *Annals of Saudi Medicine*, 30, 370–375.

Ntzoufras, I. (2009). *Bayesian Modeling using WinBUGS*. Hoboken: Wiley.

Peterson, M.G., Gaeta, T.J., Birkhahn, R.H., Fernández, J.L. and Mancuso, C.A. (2012). History of symptom triggers in patients presenting to the emergency department for asthma. *Journal of Asthma*, 49, 629–636.

Sociedade Brasileira de Pneumologia e Tisiologia. (2012). Diretrizes da Sociedade Brasileira de Pneumologia e Tisiologia para o Manejo da Asma. *Jornal Brasileiro de Pneumologia*, 38, S1–S46.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 485–493.

Thomazelli, L.M., Vieira, S., Leal, A.L., Sousa, T.S., Oliveira, D.B., Golono, M.A., Gillio, A.E., Stwien, K. E., Erdman, D.D. and Durigon, E.L. (2007). Surveillance of eight respiratory viruses in clinical samples of pediatric patients in Southeast Brazil. *Journal of Pediatrics*, 83, 422–428.

Zeger, S.L., Irizarry, R. and Peng, R.D. (2006). On time series analysis of public health and biomedical data. *Annual Review of Public Health*, 27, 57–79.

SORT 42 (1) January-June 2018, 27-44 DOI: 10.2436/20.8080.02.67

Evaluating the complexity of some families of functional data

E. G. Bongiorno[1], A. Goia[2] and P. Vieu[3]

Abstract

In this paper we study the complexity of a functional data set drawn from particular processes by means of a two-step approach. The first step considers a new graphical tool for assessing to which family the data belong: the main aim is to detect whether a sample comes from a monomial or an exponential family. This first tool is based on a nonparametric kNN estimation of small ball probability. Once the family is specified, the second step consists in evaluating the extent of complexity by estimating some specific indexes related to the assigned family. It turns out that the developed methodology is fully free from assumptions on model, distribution as well as dominating measure. Computational issues are carried out by means of simulations and finally the method is applied to analyse some financial real curves dataset.

MSC: 62-09, 62G05, 60G99.

Keywords: Small ball probability, log-Volugram, random processes, complexity class, complexity index, knn estimation, functional data analysis.

1. Introduction

The description and the analysis of a statistical sample $X_1, \ldots, X_n$ often rely on the complexity of the objects being observed. In usual multivariate situations (that is when each X_i is a d-dimensional vector) the degree of complexity is linked with the dimension d of the data which is in general known and statistical procedures are therefore developed to estimate and/or describe some probabilistic characteristic of the underlying random vector X (density function being the most common). For many reasons that we will discuss just below, this general approach cannot be followed in functional data analysis, that is the branch of statistics dealing with observations X_i which are curves, surfaces, images or other objects. Such a topic has attracted a lot of researchers and the interest towards this discipline is certified by monographs (see e.g. Bosq, 2000; Ferraty and

[1] DiSEI, Università del Piemonte Orientale, Via Perrone, 18, 28100, Novara, Italy. enea.bongiorno@uniupo.it

[2] DiSEI, Università del Piemonte Orientale, Via Perrone, 18, 28100, Novara, Italy. aldo.goia@uniupo.it

[3] Institut de Mathématiques de Toulouse, Université Paul Sabatier, France. philippe.vieu@math.univ-toulouse.fr

Received: June 2017
Accepted: February 2018

Vieu, 2006; Horváth and Kokoszka, 2012; Ramsay and Silverman, 2005), collections of recent contributions (see e.g. Aneiros et al., 2017; Bongiorno et al., 2014), special issues (see e.g. Kokoszka et al., 2017; Goia and Vieu, 2016) and recent articles (see among many others Bongiorno and Goia, 2016; Cardot, Cénac and Godichon-Baggioni, 2017; Chen, Delicado and Müller, 2017; Vilar, Raña and Aneiros, 2016). The question of defining the complexity of a functional sample has to be thought in a much more different way. The problem goes back to mathematical analysis in abstract infinite dimensional spaces, and more precisely to the difficulty for choosing some dominating measure (as could be the Lebesgue measure for continuous vectors or the counting measure for discrete ones). This is discussed in details for instance in Bogachev (1998). This has at least two important consequences. Firstly, the notion of density function has to be revisited, and secondly the notion of complexity of the model could not be reduced to a simple dimensionality index (see Bongiorno and Goia, 2017; Delaigle and Hall, 2010; Ferraty, Kudraszow and Vieu, 2012).

An usual way to overpass this difficulty when the sample comes from a variable X valued in some infinite dimensional topological space $\mathscr{F}$ is to consider the Small Ball Probability (SmBP), that is the asymptotic behaviour of $\mathbb{P}(X \in B(\chi,h))$ as h tends to zero. Here $B(\chi,h)$ stands for the ball centered at χ with radius h. Operatively, it is useful to assume that the SmBP satisfies for small h

$$\mathbb{P}(X \in B(\chi,h)) \sim \psi(\chi)\,\phi(h), \tag{1}$$

where, to ensure identifiability of the decomposition, one has to impose some normalization restriction like $\mathbb{E}[\psi(X)] = 1$. This factorization isolates the manner in which the SmBP depends upon χ and h through the *spatial* and *volumetric* terms ψ and ϕ respectively without referring to some dominating measure, and this justifies its utilization in literature (see for instance Gasser, Hall and Presnell, 1998 and Masry, 2005). Although the volumetric term has been studied extensively from a probabilistic point of view and mostly for the Gaussian processes (see the surveys on small tail literature Li and Shao, 2001; Lifshits, 2012 and references therein), from a statistical point it has only been used as a tool for controlling asymptotic behaviour of nonparametric functional estimator (see Ferraty and Vieu, 2006, Chapter 13; Masry, 2005). In fact, functional data analysis literature has focused mostly on the spatial term $\psi(\chi)$ since it naturally leads to define a surrogate density for the process and the methods vary from semi– to non–parametric approaches (see Bongiorno and Goia, 2016, 2017; Ciollaro et al., 2014; Delaigle and Hall, 2010; Delsol and Louchet, 2014; Ferraty et al., 2012) with applications in various statistical problems like defining/estimating functional modes (see Ferraty and Vieu, 2006, Chapter 6; Delaigle and Hall, 2010; Gasser et al., 1998) and classification problems (see Bongiorno and Goia, 2016; Ciollaro et al., 2014; Jacques and Preda, 2014).

To understand how the volumetric term ϕ can be of help in evaluating the complexity, firstly consider the multivariate setting $\mathscr{F} = \mathbb{R}^d$. Here, the complexity parameter is

the dimension d which appears in $\phi(h) = v_d h^d$ (with v_d being the volume of the d–dimensional unit ball), while the function $\psi(\chi)$ represents the d–dimensional density function. In the functional setting, there is an important additional problem coming from the fact that the concentration function $\phi(\cdot)$ may be of many different forms, most of them being not as simple as in the multivariate one: often ϕ can not be expressed in closed form not even asymptotically (see Bongiorno and Goia, 2017; Delaigle and Hall, 2010). On the other hand, there are some remarkable cases whose volumetric term can be explicitly written; in particular, let us look at three specific cases (trajectories drawn from two of them are depicted in Figure 1):

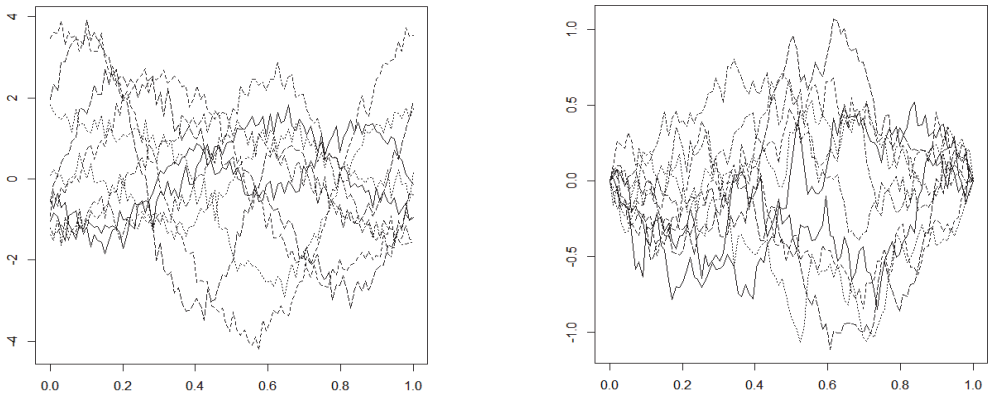

Figure 1: *Ten trajectories drawn from a noised 3–dimensional process and a Brownian Bridge process are depicted on left and right panel respectively.*

Case 1 The functional data have some finite dimensional structure. In this case the concentration function has the monomial form $\phi(h) = c_d h^d$, for some constant term c_d and the complexity of the model is the positive integer parameter d. This happens for instance when the topology on the functional space is constructed by looking only at d directions (of a given orthonormal basis) of the functional elements (see Ferraty and Vieu, 2006, Chapter 13).

Case 2 The functional data have some fractal structure (see Ferraty and Vieu, 2006, Definition 13.1). This is an extension of the first situation in which the concentration function takes the form $\phi(h) = c_\alpha h^\alpha$, for some constant term c_α and the complexity of the model is now the (non integer) positive parameter α.

Case 3 The functional data come from some Gaussian processes. This corresponds for instance to Wiener, Brownian Bridge or diffusion processes in $\mathscr{L}^2_{[0,1]}$ (see Li and Shao, 2001), for which the concentration function has the exponential form $\phi(h) = C_1 h^\gamma \exp\left\{-C_2/h^\beta\right\}$ with $\beta \in (0,\infty)$ and $\gamma \in [0,\infty)$. In this case the complexity of the data is measured by the indexes γ, β which cannot be interpreted as some dimensionality parameters (see Li and Shao, 2001 for deeper

discussion and more examples of exponential type processes). Note that, such an exponential structure is implicitly linked with the existence of some Gaussian dominating measure for the process.

In what follows, the term *monomial* (*exponential* resp.) *family* refers to the set of processes like in Case 1 or 2 (Case 3, resp.). Since these cases are, to the best of our knowledge, the only ones for which the volumetric term can be specified and cover a wide range of situations, we limit our analysis to them.

It is worth noticing that, by analogy with the finite dimensional setting, the function ϕ may reveal some latent features of the process: ϕ can be interpreted as a *roughness/complexity function* and characterizes the family to which the process belongs. Each class of functions ϕ defines a different kind of process (see examples just above), and inside each class the corresponding parameters (d, α, β, γ, ...) will be called the *complexity indexes*.

In light of what explained above, we propose a flexible approach to evaluate the complexity of functional data. The aim of our paper is twofold: firstly one has to detect the kind of process the data belong to (distinguishing between monomial and exponential families), and, once this is done, to estimate the complexity index(es). In the first step, starting from an estimate of ϕ, we introduce a method being *free of dominating measure* and based on a new graphical tool, named *log-Volugram*, that allows us to identify to which family of processes the statistical sample belongs (this is done along Section 2.1). To ensure a high degree of flexibility of the procedure, one has to use estimates being *free from parametric restriction* and models being *distribution-free*: to achieve this goal our procedure is based on kNN nonparametric functional smoothers which combine flexibility, easiness of implementation (because the dependence on a single discrete parameter) and location-adaptive feature. This is why the exploited kNN methodologies in functional data analysis are shortly reviewed at the beginning of Section 2. In the second step, once the class of the process is detected, the complexity index(es) (d, α, β, γ, ...) is(are) estimated and this can be done because of the free-modelling feature of the estimate ϕ. To do this we adopt a strategy commonly used in nonparametric framework: to study some specific submodel one compares a free-model estimate with what would be the true target under the submodel (see Härdle and Mammen, 1993 for earlier works in this direction in the multivariate regression setting). In our setting, the non-parametric estimates of ϕ is compared through a dissimilarity measure with one parametric family among the ones illustrated above and, by minimizing arguments, the complexity index(es) are estimated. This second step is presented in Section 2.2. Practical aspects about the introduced methodology and computational issues are discussed in Section 3.1 whereas the behaviour of the whole procedure is illustrated by means of a wide scope simulation studies in Section 3.2; these show good performances under different experimental conditions. Finally, to show how our two steps procedure can be usefully applied in a real case, we examine its performance in a financial framework to verify the compatibility of the data with standard model assumptions (see Section 4).

2. Methodology

In this section, after reviewing how the volumetric term in factorization (1) can be estimated nonparametrically, we show how to use it in developing new graphical tools that allow us to qualitatively detect the class of process from which the sample is drawn (see Section 2.1). Therefore, Section 2.2 describes how the nonparametric feature of the method allows to get estimates of the index complexity of the sample given the specified family.

The first statistical step consists in estimating both components in the decomposition (1) from a sample. To ensure a wide applicability of the method one has to develop statistical models/procedures being fully nonparametric. In the functional data setting, nonparametric statistics have been popularized in the book Ferraty and Vieu (2006) and are now widely used as long as one is interested in estimating some functional operator (regression, conditional distribution, ...). Among the various nonparametric smoothers, the kNN method is particularly adapted to the functional setting because it provides directly location adaptive estimates without needing highly complicated procedure (see Laloë, 2008; Burba, Ferraty and Vieu, 2009 for introductory works on functional kNN, see Biau, Cérou and Guyader, 2010; Lian, 2011; Kara et al., 2017; Kudraszow and Vieu, 2013 for the most recent advances and see Biau and Devroye, 2015 for a recent general presentation of kNN ideas).

Concerning the estimation of the terms in (1) the kNN estimates has a very simple and appealing form (see Ferraty et al., 2012). In fact, given a sample of n curves $X_1,\ldots,X_n$ drawn from X, a point $\chi \in \mathscr{F}$ and a integer $k < n$, the surrogate density ψ at χ can be estimated by

$$\widehat{\psi}_k(\chi) = \frac{k(n-1)}{\sum_{i=1}^n k_i}, \tag{2}$$

where $k_i = \#\{j \neq i : X_j \in B(X_i, H_{n,k}(\chi))\}$, $H_{n,k}(\chi) = \min\left\{h \in \mathbb{R}^+, \sum_{i=1}^n 1_{B(\chi,h)}(X_i) = k\right\}$ and $1_A(x)$ is the characteristic function of the set A. As a matter of consequence, the single parameter involved in the method is a simple integer one, namely the number k of data contained in each neighbourhood.

At this stage, once the surrogate density is estimated and given the asymptotic factorization (1), one can easily derive nonparametric estimates of the volumetric component ϕ in the following way:

$$\phi_{k,n}(h) = \frac{n^{-1}\sum_{i=1}^n 1_{B(\chi,h)}(X_i)}{\widehat{\psi}_k(\chi)}. \tag{3}$$

Theoretical assessments related to the consistency of estimators (2) and (3) are developed in Ferraty et al., 2012. In order to compute (3) one has to face some practical problems. Firstly, since the asymptotic factorization (1) holds for small h, too large values must be avoided since they may increase the estimation error. At the same time, even too small values of h must be discharged since they force $\phi_{k,n}$ to be null: the ball,

at the numerator on the right-side hand of (3), does not contain sample points. In other words, a suitable range of values $\mathscr{H} = [h_m, h_M]$ for h should be identified; for details see Section 3.1. Secondly, once h is appropriately chosen, one must take into account that the point χ, at which the SmBP is estimated, affects the approximation error of the whole factorization and, hence, the error of both $\widehat{\psi}_k$ and $\phi_{k,n}$. In this view, to circumvent such issue and to avoid an arbitrarily choice of χ, the estimation is averaged over the sample, that is

$$\widehat{\phi}_{k,n}(h) = n^{-1} \sum_{j=1}^{n} \phi_{k,n}^{(j)}(h), \tag{4}$$

where $\phi_{k,n}^{(j)}$ is (3) computed with $\chi = X_j$. In the following, if no ambiguities arise the dependences on k and/or n are dropped.

From such an estimate, one can visualize two graphical tools, that we name *Volugram* and *log-Volugram*. The shape of the latter is of help in discriminating among different family models for $\phi(\cdot)$ and in evaluating the *roughness/complexity indexes*. This is the basis of the descriptive approach to be developed in this paper.

2.1. The (log-)Volugram

The *Volugram* is the plot of $\widehat{\phi}$ computed on the realizations $x_1, \ldots, x_n$ versus h taken in a suitable positive interval sufficiently closed to zero. Because the quantities $\widehat{\phi}(h)$ are fully free from any kind of hypothesis (neither on the model, nor on the distribution of X, nor on any underlying dominating analytic structure), the observation of the shape of the curve $\widehat{\phi}$ can be directly used to have an idea on what is the complexity of the statistical sample. To fix the ideas let us just look at how behaves this Volugram in some simple examples. Figure 2 depicts the Volugrams of the noised 3-dimensional process and of the Brownian Bridge whose trajectories are illustrated in Figure 1. In both cases, estimations are based on samples of size $n = 200$, the number of neighbourhood is fixed to $k = n/2$ and, for the sake of computational practicality, h takes values in $\{H_{n,k}(x_i)\}_{i=1}^{n}$. Moreover, to ensure that the Volugram explores the smallest values of h, the plot is restricted to the 50% smallest values of the latter grid.

As is clear from Figure 2, although the Volugrams behave as one can expect (in both cases $\widehat{\phi}(h)$ decreases to zero as smaller values of h are considered), by looking at the sole Volugram it is not possible to discriminate from which family (exponential or monomial) the sample is drawn from. A practical tool to establish by eye such feature is instead provided by the *log-Volugram* defined as the plot of $\log \widehat{\phi}(h)$ versus $\log h$. Indeed, from a theoretical point of view, the volumetric term of processes in the monomial family satisfy, for small values of h, $\log \phi(h) \sim \alpha \log h$ whereas, in the exponential case, $\log \phi(h) \sim -C_2/h^{\beta}$. In other words, for small values of h, $\log \phi(h)$ is proportional to $\log h$ ($1/h^{\beta}$ respectively) for a process in the monomial (exponential respectively) family and the log-Volugram presents (does not present) a straight line shape. As a matter of

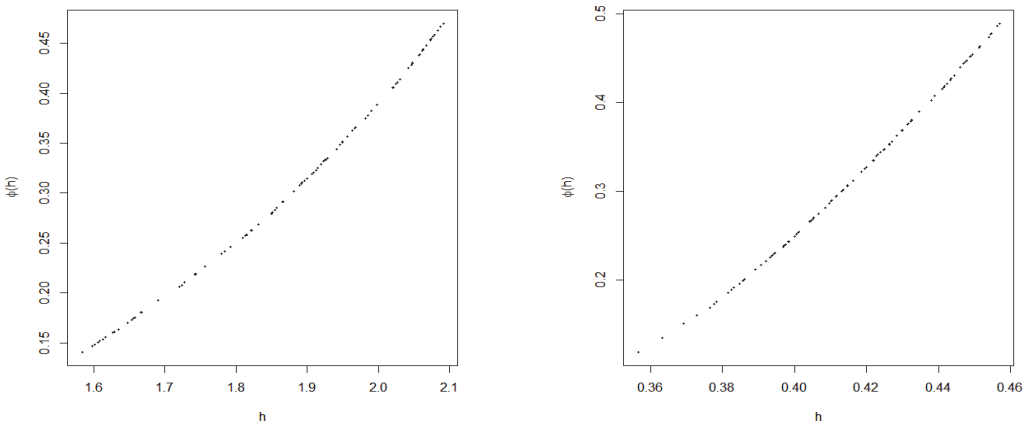

Figure 2: *Volugrams associated to a sample (of size 200) of a noised 3-dimensional process (left) and a Brownian Bridge (right) defined on* $[0,1]$. *In both cases,* $k = \lfloor n/2 \rfloor$ *and h takes values in the* 50% *smallest values of* $\{H_{n,k}(x_i)\}_{i=1}^{n}$.

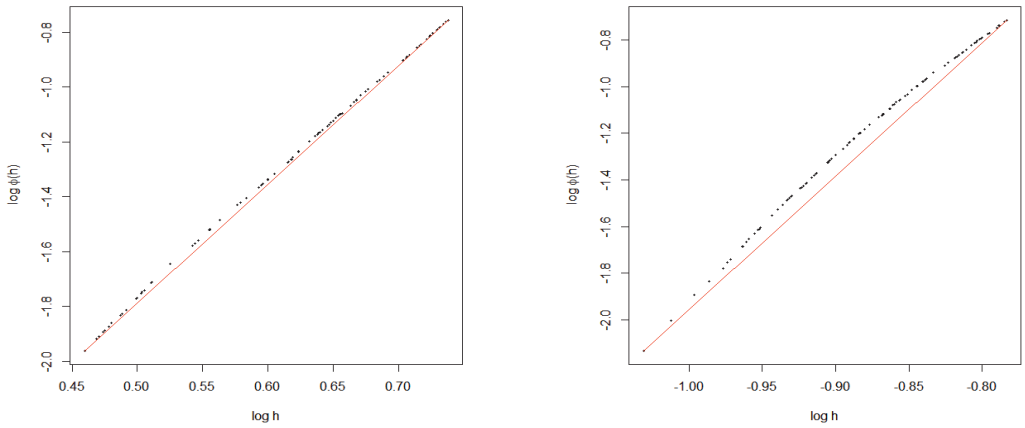

Figure 3: *log-Volugrams associated to a sample (of size 200) of a noised 3-dimensional process (left) and a Brownian Bridge (right) defined on* $[0,1]$. *In both cases,* $k = \lfloor n/2 \rfloor$ *and h takes values in the* 50% *smallest values of* $\{H_{n,k}(x_i)\}_{i=1}^{n}$. *The line passing through the first and the last points (ordered according to the ascending order of h) is drawn as well.*

illustration and using the same data and settings of Figure 2, the correspondent log-Volugrams are depicted in Figure 3. For the sake of comparison, the latter figures are completed by overlapping the line passing through the first and the last points (ordered according to the ascending order of h).

These arguments make clear how the log-Volugram allows, better than the Volugram, to drive the researcher towards the family of processes from which the sample comes. In particular, the more $\{(\log h, \log \widehat{\phi}(h))\}$ are aligned, the greater the compatibility to the monomial model is. On the contrary, deviations from this situation represent an empirical evidence of exponential model. Hence, one can decide that the theoretical

volumetric function $\phi(\cdot)$ is of some specific form depending on a complexity parameter $\theta \in \Theta$ where Θ is a subset of $\mathbb{R}^p$

$$\phi \in \mathscr{C} = \{\phi_\theta, \theta \in \Theta\}.$$

To fix the idea, the left panel of Figure 3 suggests the monomial family $\mathscr{C}_M = \{\phi_\alpha(h) = c_\alpha h^\alpha, \alpha > 0\}$, whilst the right panel leads towards the exponential one $\mathscr{C}_E = \{\phi_{(\gamma,\beta)}(h) = C_1 h^\gamma \exp\{-C_2/h^\beta\}, \beta > 0, \gamma \geq 0\}$.

2.2. Estimating the complexity index

In the second step of the procedure the aim is to gain more insights into the structure of the data by intending to estimate the complexity index θ of the chosen family $\mathscr{C}$ by means of a comparison between the free-model estimate ϕ with one of the parametric family that would be the true target. Precisely, this leads to consider the *centered cosine dissimilarity* between $g(\phi_\theta)$ and $g(\widehat{\phi}_k)$ computed on the observed values and defined by

$$\Delta(\widehat{\phi}_k, \phi_\theta) = 1 - \frac{\langle \widetilde{g}(\phi_\theta), \widetilde{g}(\widehat{\phi}_k)\rangle^2}{\|\widetilde{g}(\phi_\theta)\|^2 \|\widetilde{g}(\widehat{\phi}_k)\|^2}, \qquad k = 1, 2, \ldots, (n-1),\ \theta \in \Theta, \tag{5}$$

where $\langle f_1, f_2\rangle = \int_{\mathscr{H}} f_1 f_2$ with $\mathscr{H}$ being a suitable interval included in $(0, \infty)$, $\|f\|^2 = \langle f, f\rangle$ and $\widetilde{g}(\phi) = g(\phi) - \int_{\mathscr{H}} g(\phi)$ with $g(\cdot)$ a suitable continuous real valued function defined on $(0, +\infty)$. Note that centered cosine dissimilarity is invariant for affine transformations. Practical aspects in computing (5), including how $g(\cdot)$ and $\mathscr{H}$ are chosen, are treated in details in Section 3.1. The idea is to estimate the complexity index that minimizes $\Delta(\widehat{\phi}_k, \phi_\theta)$ over suitable grids $\mathscr{T}$ for θ, and $\mathscr{K}$ for k. Let us now show how such dissimilarity behaves in the simple examples that are following through the paper. Figure 4 depicts the heat-map of Δ (top panels) and the curves $\{\Delta(\widehat{\phi}_k, \phi_\theta) : k \in \mathscr{K}\}$ (bottom panels). These heuristically show that Δ reaches a minimum which appears rather stable with respect to the choice of k.

That spontaneously leads to estimate the complexity index by minimizing (5) for a fixed k, that is

$$\widehat{\theta} = \arg\min_{\theta \in \mathscr{T}} \Delta(\widehat{\phi}_k, \phi_\theta).$$

At this stage, it is worth noticing that if the shape of log-Volugram produces doubts in the choice of the family, it is always convenient to firstly classify the sample as drawn from the exponential family and estimate β. If a misspecification of the model occurred, then the estimation of the complexity index tends to assume the minimum values in the grid $\mathscr{T}$, see Figure 5; i.e. the exponential part of the volumetric term can be considered negligible. This can be used as a feedback procedure to avoid this kind of misspecification error.

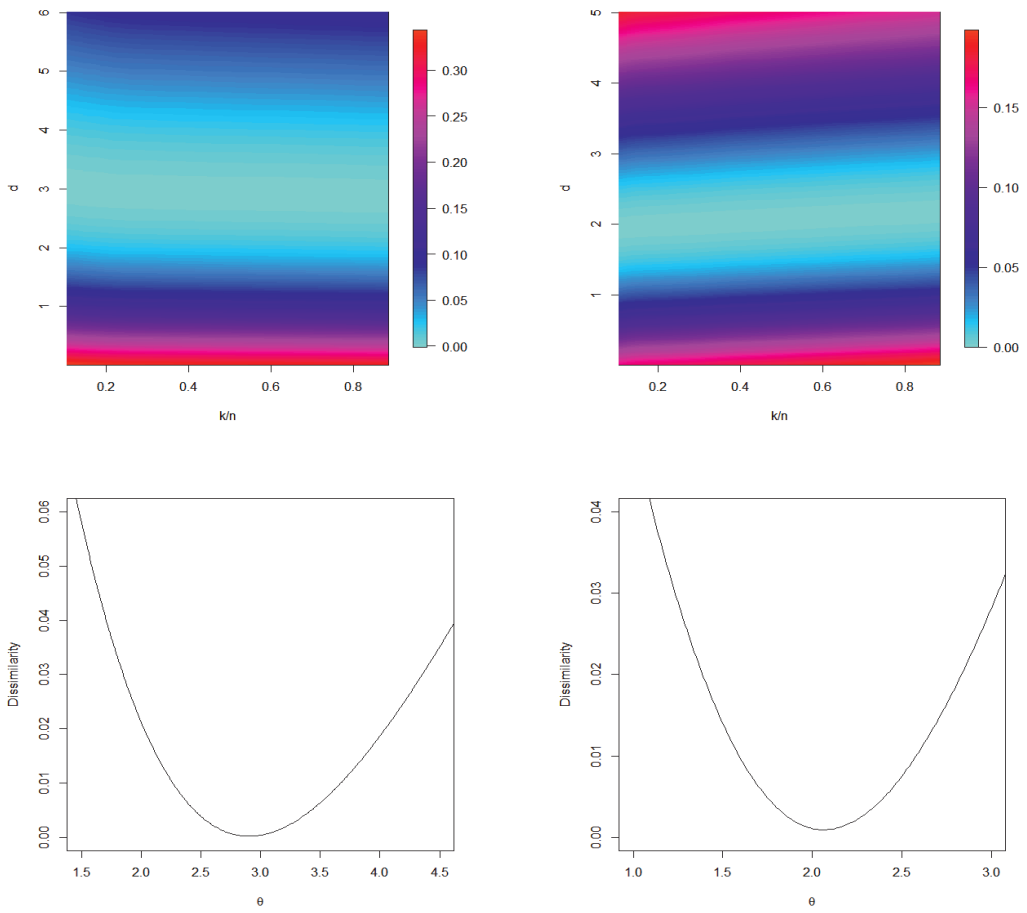

Figure 4: *Top panels: the heat maps of* $\Delta(\widehat{\phi}_k, \phi_\theta)$ *as a function of* k/n *and* θ *associated to a sample (of size 200) of a noised 3-dimensional process (left) and a Brownian Bridge (right) defined on* $[0,1]$*. Bottom panels: graphs of* $\Delta(\widehat{\phi}_k, \phi_\theta)$ *with* $k = \lfloor n/2 \rfloor$*, as a function of* θ*, associated to the same samples.*

In conclusion, the method detects the good class as explained above and, within the selected family, it seems also capable to find a good estimation of the complexity index. Simulations described in what follows confirm these abilities.

3. Algorithm in action

In this section we firstly describe the algorithm in Section 3.1 and, soon after in Section 3.2, we show its performance over a set of selected simulations under different experimental conditions.

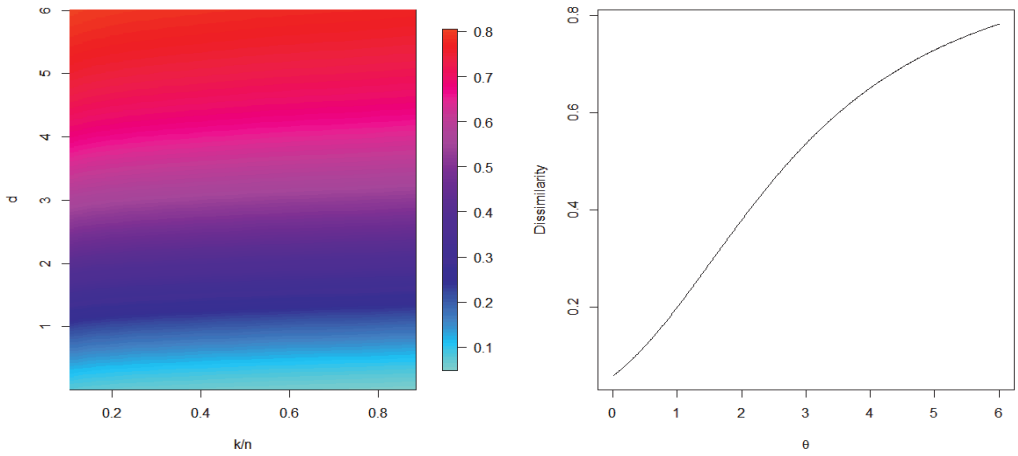

Figure 5: *The heat map of $\Delta(\widehat{\phi}_k, \phi_\theta)$ (left panel) and the plot of $\Delta(\widehat{\phi}_{k_0}, \phi_\theta)$ against θ (right panel) when the noised 3-dimensional process (in left panel of Figure 3) is confused with a process from the exponential family.*

3.1. Procedure description

Here we detail the algorithm features; some of them depend on the family identified at the first step of the methodology as described in Section 2.1.

Although the algorithm could be implemented for potentially any $\mathscr{F}$ whose topology is induced by a semimetric ρ, here, for simplicity, $\mathscr{F}$ is $\mathscr{L}^2_{[0,1]}$: the separable Hilbert space of square integrable function on $[0,1]$ with usual inner product, norm and induced metric. Thus the realizations $x_1, \ldots, x_n$ of a sample $X_1, \ldots, X_n$, drawn from the $\mathscr{F}$-valued random element X, are considered.

In computing the dissimilarity measure $\Delta(\widehat{\phi}_k, \phi_\theta)$, we have to specify $g(\cdot)$, $\mathscr{H}$, $\mathscr{T}$ and $\mathscr{K}$.

For what concerns the transformation $g(\cdot)$, if the monomial class $\mathscr{C}_M$ is suggested by the log-Volugram, g is the identity function, whereas, for the exponential class $\mathscr{C}_E$, it is the logarithm transformation. In both cases, the transformed empirical volumetric term $g(\widehat{\phi})$ is then compared with a term in the simple form ch^θ for small values of h. In fact, if $\phi \in \mathscr{C}_M$, then $\phi(h) = c_\alpha h^\alpha$ with $\alpha \in (0, \infty)$. If $\phi \in \mathscr{C}_E$,

$$\log \phi(h) = \log C_1 + \gamma \log h - C_2 h^{-\beta} \sim -C_2 h^{-\beta} \tag{6}$$

and then, in the exponential case, the leading complexity parameter is β. Indeed, at the best of our knowledge, for the most of processes related to Brownian motion with known SmBP asymptotic, it holds $\log \phi(h) \sim -C_2 h^{-2}$ (see, for instance, Nikitin and Pusev, 2013). In particular, $C_2 = 1/8$ when X is Wiener, Brownian Bridge (BB), Geometric

Brownian Motion (GBM), Ornstein-Uhlenbeck. Anyway, note that (6) is more accurate if $\gamma = 0$, and this happens, for instance, in the case of Brownian Bridge that consequently becomes a benchmark process. In practice, beside the BB, we have specialized our method to deal with those processes suspected to be Wiener or GBM since these can be led back to a BB by means of suitable transformations. In more details, if $X(t) = W(t)$ is Wiener on $t \in [0,1]$, then

$$W(t) - tW(1), \tag{7}$$

is a BB on $[0,1]$, whereas if $X(t)$ is the GBM identified by the stochastic differential equation

$$\begin{cases} dX(t) = \mu X(t)dt + \sigma X(t)dW(t), & t \in [0,1], \\ X(0), \sigma > 0, \end{cases} \tag{8}$$

whose solution is $X(t) = X(0)\exp\left\{\left(\mu - \sigma^2/2\right)t + \sigma W(t)\right\}$, $t \in [0,1]$, then

$$[\log(X(t)/X(0)) - (\mu - \sigma^2/2)t]/\sigma, \qquad t \in [0,1] \tag{9}$$

is a Wiener process for which transformation (7) can be applied, leading to a BB on $[0,1]$. The estimation of γ remains an open problem for processes different from the BB, the Wiener process and the GBM.

For what concerns $\mathscr{H} = [h_m, h_M]$, h_m is chosen in order to guarantee that there exists at least an observed curve x_i for which $B(x_i, h_m)$ includes some $x_j \neq x_i$; whereas, the range $\mathscr{H}$ should become closer to zero as the sample size increases.

Finally, $\mathscr{T}$ is a equally spaced mesh over an interval that varies with the experimental setting; our suggestion is to start with a wide range of values with a relatively rough step, then to restrict the region of search by using a finer grid. To reveal possible dependencies on k, in the simulation study, we use $\mathscr{K} = \{\lfloor \delta n \rfloor : \delta = 1/4, 1/3, 1/2\}$ with $\lfloor \delta n \rfloor$ being the smaller integer greater than δn. Such a choice is coherent with many rules introduced in literature (see, for instance, Devroye, Györfi and Lugosi, 1996; Duda, Hart and Stork, 2012; Györfi et al., 2006).

3.2. Numerical Experiments

In this section we present the results of numerical experiments aimed to evaluate the ability of the method in estimating the complexity parameter by varying the underlying process and the sample size.

We generate 1000 Monte Carlo samples each one constituted by n independent random curves $X_1, \ldots, X_n$ drawn from a process X with $n = 50, 100, 200, 500$. From each sample the complexity index is estimated and its distribution analysed. In particular, we consider noised finite dimensional processes and infinite dimensional ones.

About the noised finite dimensional processes, curves are generated according to

$$X(t) = \sum_{j=1}^{d} a_j \xi_j(t) + \mathscr{E}(t), \qquad t \in [0,1]$$

where $\{\xi_j\}_{j=1}^{d}$ are the first d elements of the Fourier basis

$$\xi_j(t) = \begin{cases} \sqrt{2}\sin(2\pi m t - \pi), & j = 2m-1 \\ \sqrt{2}\cos(2\pi m t - \pi), & j = 2m \end{cases} \qquad m \in \mathbb{N},$$

$\{a_j\}_{j=1}^{d}$ are i.i.d. as $\mathscr{N}(0,1)$ and $\mathscr{E}(t)$ is a Gaussian white noise with $\sigma = 0.02$ representing a measurement error. Here, $d = 3$ and $d = 6$ are considered.

For what concerns the infinite dimensional processes, we consider the Wiener process and the Geometric Brownian Motion (GBM). Each curve in both families are discretized over an equispaced grid on $[0,1]$ consisting of 100 points: the resulting data-sets are $n \times 100$ matrices with entries $x_{i,j}, i = 1,\dots,n, j = 1,\dots,100$. GBM trajectories are simulated from the stochastic differential equation (8), with $X(0) = 1$, $\mu = 0$, $\sigma = 1$, using the Euler-Maruyama approximation scheme (Kloeden and Platen, 1992, Section 9.1). Coherently with what stated in the previous section, Wiener curves are transformed by means of (7) whereas GBM trajectories by (9) and, successively, (7). To operationalize (9), maximum likelihood estimates of parameters are computed for each discretized curve in the sample: for each $i = 1,\dots,n$, μ and σ are estimated by $\hat{\mu} = 100^{-1}\sum_{j=1}^{100} x_{i,j}$ and $\hat{\sigma}^2 = 100^{-1}\sum_{j=1}^{100}(x_{i,j} - \hat{\mu})^2$ respectively. Integrals in (5) are approximated on such grid by using a rectangular numerical rule. In what follows, $\mathscr{T}$ is an equispaced grid with step 0.01 and $k = \lfloor \delta n \rfloor$ with $\delta = 1/4, 1/3, 1/2$.

Table 1 collects the results from the Monte Carlo experiments from which we can appreciate the good performances of complexity index estimator. In particular, in all the cases no relevant bias arises, variability of the estimator is moderate, especially, in relative terms with respect to the true parameter. As expected, variability decreases with n whereas, in the finite dimensional case, it slightly increases with the complexity: the larger d is, the larger the variability in relative terms with respect to the true parameter is. These comments hold true for all the chosen k, therefore, for practical purposes, an heuristic choice like $k = \lfloor n/2 \rfloor$ is reasonable.

The distributions of estimated values $\widehat{d}$ and $\widehat{\beta}$ over the 1000 simulations when $n = 500$ and $k = 250$ are plotted in Figure 6: dashed vertical lines are superimposed to kernel density estimates in correspondence of extreme quantiles of order 0.025 and 0.975, in order to delimit a Monte Carlo empirical 95% confidence interval. All distributions appear rather symmetric and bell-shaped: anyway, the Shapiro-Wilk test tends to reject the normality assumption in all the cases at the level 5%.

Table 1: *Synthetic indicators of the estimated complexity indexes obtained from 1000 MC replication under different experimental conditions.*

Process	$\delta \rightarrow$	1/2		1/3		1/4	
family	$n \downarrow$	Mean	St.dev	Mean	St.dev	Mean	St.dev
Finite dimensional	50	2.966	0.305	2.994	0.312	3.023	0.316
(with $d = 3$)	100	2.985	0.167	3.007	0.169	3.023	0.170
	200	3.014	0.100	3.034	0.100	3.046	0.100
	500	3.072	0.072	3.090	0.071	3.101	0.070
Finite dimensional	50	5.938	1.559	5.962	1.596	5.986	1.629
(with $d = 6$)	100	5.891	0.911	5.926	0.933	5.947	0.947
	200	5.894	0.497	5.923	0.509	5.940	0.521
	500	5.985	0.249	6.008	0.255	6.023	0.260
GBM ($\beta = 2$)	50	1.995	0.190	1.927	0.187	1.897	0.188
	100	2.011	0.117	1.949	0.113	1.916	0.112
	200	2.015	0.072	1.956	0.070	1.924	0.069
	500	2.005	0.039	1.952	0.038	1.921	0.038
Wiener ($\beta = 2$)	50	1.987	0.190	1.920	0.187	1.888	0.188
	100	2.005	0.117	1.942	0.113	1.909	0.112
	200	2.009	0.071	1.950	0.069	1.918	0.069
	500	1.999	0.039	1.946	0.038	1.915	0.037

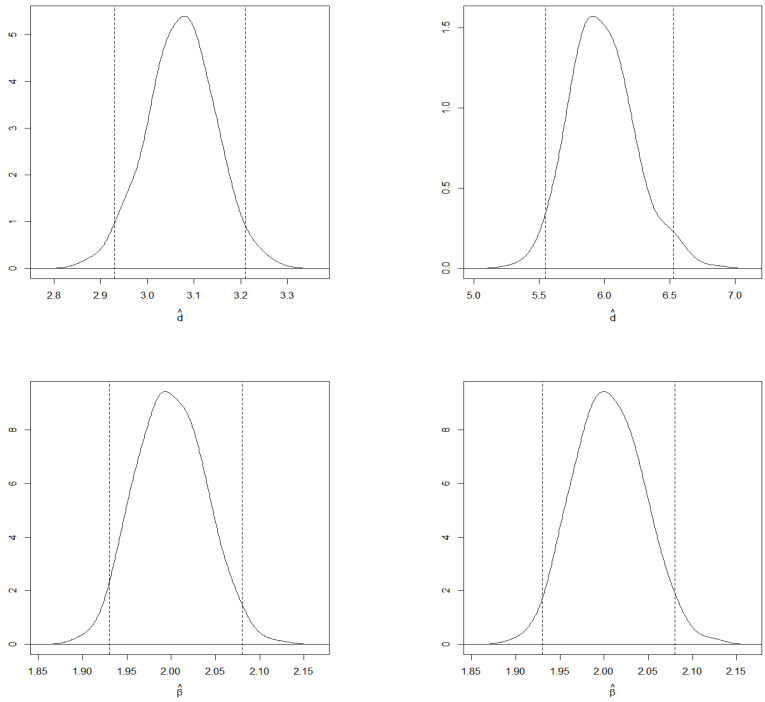

Figure 6: *Kernel density estimates of $\widehat{d}$ and $\widehat{\beta}$ for the finite dimensional processes ($d = 3$ and $d = 6$, left and right top panels respectively) and for the GBM and Wiener processes ($\beta = 2$, left and right bottom panels respectively) when $n = 500$. Dashed vertical lines correspond to the 95% Monte Carlo confidence interval limits.*

4. Application to financial data

A common problem in finance is the modelling of stock prices time series, for example in implementing parametric option pricing models via Monte Carlo simulations. Thanks to its nice properties, the GBM has gained a central place in theoretical and applied financial literature, becoming a prototype for a generation of models; see, for example, Fusai and Roncoroni (2007) and Campbell, Lo and MacKinlay (1997).

In this section we illustrate how the proposed methodology can provide a tool for practitioners in detecting the family of processes to which the observed time series belongs, and for a rough evaluation of the complexity of such data. To do this, we analyze in details the case of the S&P500 during the period 14th October 2016, 15th January 2017 with 1 minute frequency for a total of 63 market days and 390 observations per day (we deleted shorter days). Data are collected by using the link https://www.google.com/finance/getprices?i=60&p=200d&f=d,o,h,l,c,v&df=cpct&q=.INX. The corresponding trajectory is depicted in the left panel of Figure 7. To qualitatively assess that the observed trajectory is compatible with a GBM process, we apply our method on a sample derived from above dataset: given the high frequency of measurements, each market day is divided into three non-overlapping parts having the same size to which correspond three trajectories. Consequently, the sample is formed by $n = 189$ each one discretized over an equally spaced grid of 130 points.

In order to implement the two steps of our method, the sample must be transformed as explained in detail in Section 3.2. In particular, given the assumption that the underlying process is a GBM, since drift and volatility of a stock process vary with time, it is reasonable to model each curve x_i with specific parameters μ_i and σ_i. They are estimated by using the maximum likelihood approach illustrated in the previous section,

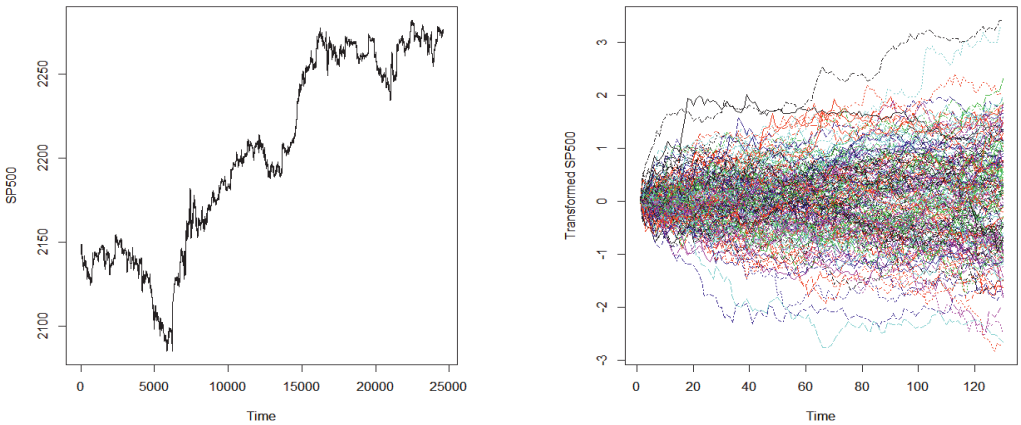

Figure 7: *Left panel - Trajectory of S&P500 value from 14th October 2016 to 15th January 2017 with 1 minute frequency. Right panel - The functional sample: each functional observation is one third of a market day trajectory after transformation* (9).

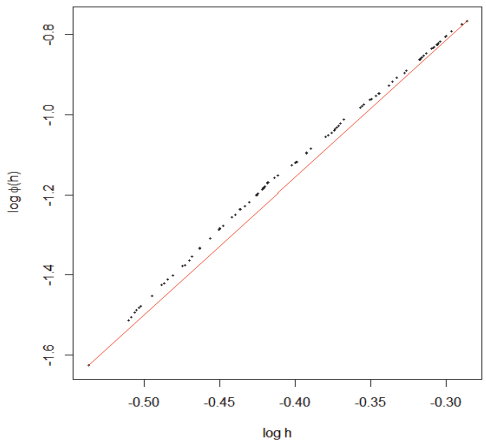

Figure 8: *The log-Volugram for the transformed S&P500 sample.*

starting from discretized points of each curve. The sample of curves which arise from these manipulations is plotted in the right panel of Figure 7.

In the same spirit of Section 2.1, we plot the log-Volugram with $k = \lfloor n/2 \rfloor$, see Figure 8. Its shape drives our analysis towards the exponential family. The heat map of dissimilarity Δ and the dissimilarity computed at $k = \lfloor n/2 \rfloor$ are drawn in Figure 9. The minimization of Δ leads to $\widehat{\beta} = 1.94$.

This first analysis supports the assumption that S&P500 could be modelled as a GBM with varying parameters at least for a short time period.

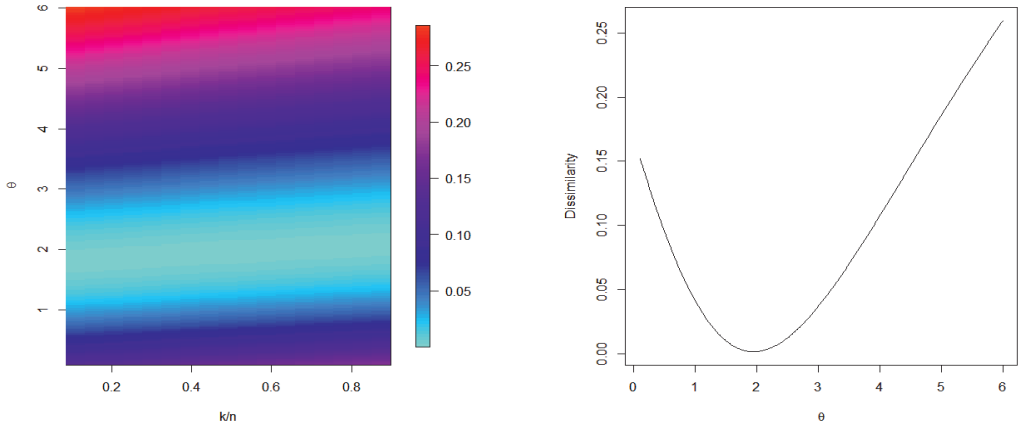

Figure 9: *The heat map of Δ (left panel) and Δ at $k = \lfloor n/2 \rfloor$ (right panel) for the transformed S&P500 sample.*

In order to evaluate the stability of results with respect to the way in which we built the sample of functional data, we repeated the analysis using different cutting criteria: besides dividing each market day in three parts, we tried also with two parts consisting of 195 points, five parts of 78 points and 6 parts of 65 (all the intervals are not overlapped). Resulting samples have sizes $n = 126, 315, 378$ whereas the results obtained with $k = \lfloor n/2 \rfloor$ are $\widehat{\beta} = 1.94, 1.96, 1.98$ respectively. They confirm the compatibility of data with a GBM (with time varying parameters) assumption that, hence, can be used as a good approximating model for performing option pricing.

5. Comments

This paper has provided flexible tools for analysing the complexity of a functional statistical sample. In order to ensure its high degree of applicability the procedure is free from any structural assumption from several points of view: from an analytic point of view (it is free from any dominating measure assumption in the underlying infinite dimensional space), from a probabilistic point of view (it is free from any distribution assumption on the underlying stochastic process), from a statistical point of view (it is free from any parametric assumption on the model), and from a computational point of view (the method depends on a single discrete parameter). This has been possible by using kNN ideas that combine good theoretical properties and ease of implementation. In a first step, the method provides some graphical tools (the so-called Volugram or log-Volugram) which are used to detect the class of complexity of the data, while in a second step it provides an automatic estimate of the index of complexity inside of the detected class. The methodology provides excellent results in evaluating the complexity family and index on simulated and real datasets.

Acknowledgements

We wish to thank the Associate Editor and an anonymous referee for their valuable comments. This work was partially developed during a visiting research period at the Institut de Mathématiques de Toulouse, Université Paul Sabatier. E. Bongiorno and A. Goia thank for the hospitality. E. Bongiorno and A. Goia 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 CRoNoS - COST Action IC1408 is acknowledged.

References

Aneiros, G., Bongiorno, E.G., Cao, R. and Vieu, P. (2017). *Functional Statistics and Related Fields*. Springer.

Biau, G., Cérou, F. and Guyader, A. (2010). Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Transactions on Information Theory*, 56, 2034–2040.

Biau, G. and Devroye, L. (2015). *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, Cham.

Bogachev, V.I. (1998). *Gaussian Measures*. Vol. 62 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI.

Bongiorno, E.G. and Goia, A. (2016). Classification methods for Hilbert data based on surrogate density. *Computational Statistics & Data Analysis*, 99, 204–222.

Bongiorno, E.G. and Goia, A. (2017). Some insights about the small ball probability factorization for Hilbert random elements. *Statistica Sinica*, Forthcoming.

Bongiorno, E.G., Goia, A., Salinelli, E. and Vieu, P. (Eds.) (2014). *Contributions in Infinite-Dimensional Statistics and Related Topics*. Società Editrice Esculapio.

Bosq, D. (2000). *Linear Processes in Function Spaces*. Vol. 149 of Lecture Notes in Statistics. Springer-Verlag, New York.

Burba, F., Ferraty, F. and Vieu, P. (2009). k-nearest neighbour method in functional nonparametric regression. *Journal of Nonparametric Statistics*, 21, 453–469.

Campbell, J.Y., Lo, A.W.-C. and MacKinlay, A.C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.

Cardot, H., Cénac, P. and Godichon-Baggioni, A. (2017). Online estimation of the geometric median in Hilbert spaces: Nonasymptotic confidence balls. *Annals of Statistics*, 45, 591–614.

Chen, K., Delicado, P. and Müller, H.-G. (2017). Modelling function-valued stochastic processes, with applications to fertility dynamics. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79, 177–196.

Ciollaro, M., Genovese, C., Lei, J. and Wasserman, L. (2014). *The Functional Mean-Shift Algorithm for Mode Hunting and Clustering in Infinite Dimensions*. Preprint.

Delaigle, A. and Hall, P. (2010). Defining probability density for a distribution of random functions. *Annals of Statistics*, 38, 1171–1193.

Delsol, L. and Louchet, C. (2014). Segmentation of hyperspectral images from functional kernel density estimation. In: *Contributions in Infinite-Dimensional Statistics and Related Topics*. Esculapio, Bologna, pp. 101–106.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Vol. 31 of Applications of Mathematics (New York). Springer-Verlag, New York.

Duda, R.O., Hart, P.E. and Stork, D.G. (2012). *Pattern Classification*. John Wiley & Sons.

Ferraty, F., Kudraszow, N. and Vieu, P. (2012). Nonparametric estimation of a surrogate density function in infinite-dimensional spaces. *Journal of Nonparametric Statistics*, 24, 447–464.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer Series in Statistics. Springer, New York.

Fusai, G. and Roncoroni, A. (2007). *Implementing Models in Quantitative Finance: Methods and Cases*. Springer Science & Business Media.

Gasser, T., Hall, P. and Presnell, B. (1998). Nonparametric estimation of the mode of a dis-tribution of random curves. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 60, 681–691.

Goia, A. and Vieu, P. (2016). An introduction to recent advances in high/infinite dimensional statistics [Editorial]. *Journal of Multivariate Analysis*, 146, 1–6.

Györfi, L., Kohler, M., Krzyzak, A.and Walk, H. (2006). *A Distribution-Free Theory of Non-Parametric Regression*. Springer Science & Business Media.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, 21, 1926–1947.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer, New York.

Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8, 231–255.

Kara, L.-Z., Laksaci, A., Rachdi, M. and Vieu, P. (2017). Data-driven kNN estimation in nonparametric functional data analysis. *Journal of Multivariate Analysis*, 153, 176–188.

Kloeden, P.E. and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Vol. 23 of Applications of Mathematics (New York). Springer-Verlag, Berlin.

Kokoszka, P., Oja, H., Park, B. and Sangalli, L. (2017). Special issue on functional data analysis. *Econometrics and Statistics*, 1, 99–100.

Kudraszow, N.L. and Vieu, P. (2013). Uniform consistency of kNN regressors for functional variables. *Statistics & Probability Letters*, 83, 1863–1870.

Laloë, T. (2008). A k-nearest neighbor approach for functional regression. *Statistics & Probability Letters*, 78, 1189–1193.

Li, W.V., Shao and Q.-M. (2001). Gaussian processes: inequalities, small ball probabilities and applications. In: *Stochastic Processes: Theory and Methods*. Vol. 19 of Handbook of Statistics North-Holland, Amsterdam, pp. 533–597.

Lian, H. (2011). Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electronic Journal of Statistics*, 5, 31–40.

Lifshits, M.A. (2012). *Lectures on Gaussian Processes*. Springer Briefs in Mathematics. Springer, Heidelberg.

Masry, E. (2005). Nonparametric regression estimation for dependent functional data: asymptotic normality. *Stochastic Processes and their Applications*, 115, 155–177.

Nikitin, Y.Y. and Pusev, R.S. (2013). Exact small deviation asymptotics for some Brownian functionals. *Theory of Probability and Its Applications*, 57, 60–81.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, 2nd Edition. Springer Series in Statistics. Springer, New York.

Vilar, J.M., Raña, P. and Aneiros, G. (2016). Using robust FPCA to identify outliers in functional time series, with applications to the electricity market. *SORT*, 40, 321–348.

SORT 42 (1) January-June 2018, 45-58 DOI: 10.2436/20.8080.02.68

Preliminary test and Stein-type shrinkage LASSO-based estimators

M. Norouzirad and M. Arashi*

Abstract

Suppose the regression vector-parameter is subjected to lie in a subspace hypothesis in a linear regression model. In situations where the use of least absolute and shrinkage selection operator (LASSO) is desired, we propose a restricted LASSO estimator. To improve its performance, LASSO-type shrinkage estimators are also developed and their asymptotic performance is studied. For numerical analysis, we used relative efficiency and mean prediction error to compare the estimators which resulted in the shrinkage estimators to have better performance compared to the LASSO.

MSC: 62F15, 62H05.

Keywords: Double shrinking, LASSO, preliminary test LASSO, restricted LASSO, Stein-type shrinkage LASSO.

1. Introduction

Consider the linear regression model with form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$ is a vector of responses, $\boldsymbol{X}$ is an $n \times p$ non-stochastic design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is an unknown vector of parameters, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ is the vector of random errors, with $E(\boldsymbol{\epsilon}_n) = \boldsymbol{0}$ and $E(\boldsymbol{\epsilon}_n \boldsymbol{\epsilon}_n^\top) = \sigma^2 \boldsymbol{I}_n (\sigma^2 < \infty)$, $\boldsymbol{I}_n$ the identity matrix of order n.

In general, the main goal of the linear regression model (1) is the estimation of parameters and prediction of response for a given design matrix. The estimation problem is usually solved through the ordinary least squares (OLS) method. Provided $\boldsymbol{C}_n = \boldsymbol{X}^\top \boldsymbol{X}$ is well-conditioned, we use the OLS estimator given by $\tilde{\boldsymbol{\beta}}_n = \boldsymbol{C}_n^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$. The corresponding estimator of σ^2 is $s_e^2 = (\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}_n)^\top (\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}_n)/m$, $m = n - p$.

Corresponding author: m_arashi_stat@yahoo.com

* Department of Statistics, Faculty of Mathematical Sciences. Shahrood University of Technology, Shahrood, Iran

Received: February 2017

Accepted: February 2018

Assume the following regularity conditions:

A1: $\max_{1\leq i\leq n} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{x}_i \to 0$ as $n\to\infty$ where $\boldsymbol{x}_i^{\mathrm{T}}$ is the ith row of design matrix $\boldsymbol{X}$.

A2: $\lim_{n\to\infty} n^{-1}\boldsymbol{C}_n = \boldsymbol{C}$, where $\boldsymbol{C}$ is finite and positive-definite matrix.
Then, asymptotically $\tilde{\boldsymbol{\beta}}_n \sim \mathscr{N}_p(\boldsymbol{\beta}, \sigma^2\boldsymbol{C}^{-1})$, which is independent of $(ms_e^2)/\sigma^2 \sim \chi^2_m$ (asymptotically).

Now, suppose that we are provided with some prior information about the whole or subset of covariates. This prior information can be utilized to improve the overall estimation of the regression coefficients using shrinkage estimation (Ahmed and Raheem, 2012).

There are many notable studies incorporating prior information, in the form of restrictions, to improve estimation in the sense that the restricted and shrinkage estimators have lesser risk and prediction error values.

Saleh (2006) gives extensive overviews on preliminary test and shrinkage estimators using the OLS, ridge and maximum likelihood (ML) estimators as starting points. Fallahpour et al. (2012) developed shrinkage estimators by using the weighted semiparametric OLS estimator. Hossain and Ahmed (2014) start by maximum partial likelihood estimator and propose shrinkage and positive shrinkage estimators, while Roozbeh (2015, 2016) develops shrinkage estimators in a ridge regression. Other related studies include Hossain et al. (2015), Hossain and Howlader (2016), Hossain et al. (2016), Yuzbasi and Ahmed (2016) and Yuzbasi et al. (2017), to mention a few.

However, in this study, we have different concerns. As a prelude, Tibshirani (1996) proposed a new method for variable selection that produces an accurate, stable, and parsimonious model, called least absolute shrinkage and selection operator (LASSO) that is obtained by

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j| \right\}, \qquad \lambda_n \geq 0, \tag{2}$$

where λ_n is the tuning parameter, controlling the level of sparsity in $\hat{\boldsymbol{\beta}}^{\mathrm{L}}$.

Now, the questions are as follows:

1. How can we build the theory if we start with the LASSO instead of using the OLS/ML estimator?
2. What will the form of shrinkage estimators be under restriction, when LASSO is used as the starting point?
3. Is it possible to derive asymptotic properties of the preliminary test and shrinkage estimators based on the LASSO?

In this paper, we cover the above issues. Hence, we organize the paper as follows: In Section 2, the restricted LASSO estimator is defined for inference under restriction and the concept of double shrinking is introduced (covering questions 1 and 2 above). Section 3 contains the asymptotic distributions of the proposed estimators (covering question 3 above). An extensive numerical study is carried out in Section 4 and we conclude our study in Section 5.

2. Restricted LASSO and double shrinking

The LASSO estimator has been denoted as $\hat{\boldsymbol{\beta}}_n^{\mathrm{L}}$ and termed as unrestricted LASSO estimator (ULE). Now, suppose that some non-sample information (a priori restriction on the parameters) about the covariates is available. A set of q linear restrictions on the vector $\boldsymbol{\beta}$ can be written as $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$. Or, we can suppose that our model is subjected to lie in the linear subspace restriction,

$$\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}, \tag{3}$$

where $\boldsymbol{H}$ is a $q \times p$ $(q \leq p)$ matrix of known elements, and $\boldsymbol{h}$ is a q vector of known components. The rank of $\boldsymbol{H}$ is q, which implies that the restrictions are linearly independent. This restriction may be (i) a fact known from theoretical or experimental considerations, (ii) a hypothesis that may have to be tested or (iii) an artificially imposed condition to reduce or eliminate redundancy in the description of model (Sengupta and Jammalamadaka, 2003).

Our proposal is to consider the following estimator as the restricted LASSO estimator (RLE),

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{RL}} = \hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \boldsymbol{C}_n^{-1}\boldsymbol{H}^{\mathsf{T}}(\boldsymbol{H}\boldsymbol{C}_n^{-1}\boldsymbol{H}^{\mathsf{T}})^{-1}(\boldsymbol{H}\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \boldsymbol{h}). \tag{4}$$

The above closed form RLE cannot be achieved via routine optimization techniques. Indeed, we proposed it by the analogy of OLS estimator of $\boldsymbol{\beta}$ subject to the restriction $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$.

When (3) is satisfied, $\hat{\boldsymbol{\beta}}_n^{\mathrm{RL}}$ has smaller asymptotic risk than $\hat{\boldsymbol{\beta}}_n^{\mathrm{L}}$. However, for $\boldsymbol{H}\boldsymbol{\beta} \neq \boldsymbol{h}$, $\hat{\boldsymbol{\beta}}_n^{\mathrm{RL}}$ may be biased and inconsistent in many cases. Now, how can we decide on ULE or RLE, since we do not know whether the restriction holds? To solve this, it is plausible to follow Fisher's recipe and define the preliminary test LASSO estimator (PTLE) by taking $\hat{\boldsymbol{\beta}}_n^{\mathrm{L}}$ or $\hat{\boldsymbol{\beta}}_n^{\mathrm{RL}}$ according to the acceptance or rejection of the null hypothesis, $\mathscr{H}_o$: $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$.

This estimator will have the form

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{PTL}} = \hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - (\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \hat{\boldsymbol{\beta}}_n^{\mathrm{RL}})I(\mathscr{L}_n \leq \mathscr{L}_{n,\alpha}), \tag{5}$$

where $\mathscr{L}_{n,\alpha}$ is the upper α-level critical value of the exact distribution of the test statistic $\mathscr{L}_n$ under $\mathscr{H}_o$. We will propose a relevant test statistic later in Section 3.

The PTLE is highly dependent on the level of significance α and has discrete nature which is simplified to one of the extremes $\hat{\boldsymbol{\beta}}_n^{\mathrm{L}}$ or $\hat{\boldsymbol{\beta}}_n^{\mathrm{RL}}$ according to the output of the test. In this respect, making use of a continuous and α-free estimator may make more sense. Now, we propose a double shrinking idea which reflects a relevant estimator. It is well-known that the LASSO estimator shrinks coefficients toward the origin. However, when the restriction $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$ is subjected to the model, it is of major importance that the estimator be shrunken toward the restricted one as well. Hence, there must be shrinking toward two directions or double shrinking concept, say. Consequently, we combine the idea of James and Stein (1961) shrinkage and LASSO to propose the following Stein-type shrinkage LASSO estimator (SSLE)

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{SSL}} = \hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - k_n(\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \hat{\boldsymbol{\beta}}_n^{\mathrm{RL}})\mathscr{L}_n^{-1}, \qquad k_n = \frac{m(q-2)}{m+2}, \tag{6}$$

where k_n is the shrinkage constant.

The estimator $\hat{\boldsymbol{\beta}}_n^{\mathrm{SSL}}$ may go past the estimator $\hat{\boldsymbol{\beta}}_n^{\mathrm{RL}}$. So, we define the positive-rule Stein-type shrinkage LASSO estimator (PRSSLE) given by

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n^{\mathrm{PRSSL}} &= \hat{\boldsymbol{\beta}}_n^{\mathrm{RL}} + (1 - k_n\mathscr{L}_n^{-1})I(\mathscr{L}_n > k_n)(\hat{\boldsymbol{\beta}}_n^{L} - \hat{\boldsymbol{\beta}}_n^{\mathrm{RL}}),\\ &= \hat{\boldsymbol{\beta}}_n^{\mathrm{SSL}} - (1 - k_n\mathscr{L}_n^{-1})I(\mathscr{L}_n \leq k_n)(\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \hat{\boldsymbol{\beta}}_n^{\mathrm{RL}}).\end{aligned} \tag{7}$$

We note that, as the test based on $\mathscr{L}_n$ is consistent against fixed $\boldsymbol{\beta}$ such that $\boldsymbol{H}\boldsymbol{\beta} \neq \boldsymbol{h}$, the PTLE, SSLE and PRSSLE are asymptotically equivalent to the ULE for fixed alternative. Hence, we will investigate the asymptotic risks under local alternatives and compare the performance of the estimators.

3. Some asymptotic results

For the purpose of this section, we consider the class of local alternatives, $\mathscr{K}_{(n)}$ defined by

$$\mathscr{K}_{(n)} : \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h} + n^{-\frac{1}{2}}\boldsymbol{\xi}, \quad \boldsymbol{\xi} = (\xi_1, \ldots, \xi_q)^{\intercal} \in \mathbb{R}^q.$$

Let $\hat{\boldsymbol{\beta}}_n^*$ be any estimator of $\boldsymbol{\beta}$. We define the asymptotic cumulative distribution function (c.d.f.) of $\hat{\boldsymbol{\beta}}_n^*$, under $\mathscr{K}_{(n)}$, as

$$G_p(x) = \lim_{n\to\infty} P_{\mathscr{K}_{(n)}}\left\{\sqrt{n}s_e^{-1}(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}) \leq \boldsymbol{x}\right\}.$$

If the asymptotic c.d.f. exists, then the asymptotic distributional bias (ADB) and quadratic bias (ADQB) are given by

$$\boldsymbol{b}(\hat{\boldsymbol{\beta}}_n^*) = \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta})\right] = \int \boldsymbol{x} dG_p(\boldsymbol{x}),$$
$$B(\hat{\boldsymbol{\beta}}_n^*) = \sigma^{-2}[\boldsymbol{b}(\hat{\boldsymbol{\beta}}_n^*)]^{\top}\boldsymbol{C}[\boldsymbol{b}(\hat{\boldsymbol{\beta}}_n^*)],$$

respectively, where $\sigma^2\boldsymbol{C}^{-1}$ is the mean squared error (MSE)-matrix of $\tilde{\boldsymbol{\beta}}_n$ as $n \to \infty$. Defining

$$\boldsymbol{M}(\hat{\boldsymbol{\beta}}_n^*) = \int \boldsymbol{x}\boldsymbol{x}^{\top} dG_p(\boldsymbol{x}) = \lim_{n\to\infty} E\left[n(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta})^{\top}\right],$$

as the asymptotic distributional MSE (ADMSE), we have the weighted risk of $\hat{\boldsymbol{\beta}}_n^*$ given by

$$\mathrm{R}(\hat{\boldsymbol{\beta}}_n^*) = \mathrm{tr}[\boldsymbol{M}(\hat{\boldsymbol{\beta}}_n^*)] = \lim_{n\to\infty} E[n(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta})^{\top}(\hat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta})]$$

as the asymptotic distributional quadratic risk (ADQR).

Suppose the LASSO is weakly consistent, i.e., $\lambda_n = o(n^{1/2})$. Up to this point, we implemented a test statistic based on the OLS estimator, however, constructing a test based on the LASSO estimator will give the same asymptotic behaviour under weak consistency. A test statistic based on the ULE will have form

$$\mathscr{L}_n = \frac{(\boldsymbol{H}\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \boldsymbol{h})^{\top}(\boldsymbol{H}\boldsymbol{C}_n^{-1}\boldsymbol{H}^{\top})^{-1}(\boldsymbol{H}\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \boldsymbol{h})}{s_L^2}, \tag{8}$$

where

$$s_L^2 = \frac{1}{m}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_n^{\mathrm{L}})^{\top}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_n^{\mathrm{L}}) \tag{9}$$

Using Theorem 2 of Knight and Fu (2000), Theorem 7.8.2.3 of Saleh (2006), and $\sqrt{n}$-consistency, we have the following important result.

Theorem 1 *Under the assumptions of Theorem 2 and $\lambda_n = o(n^{1/2})$, we have*

(i) $W_n^{(1)} = \sqrt{n}(\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \boldsymbol{\beta}) \overset{\mathscr{D}}{=} W = \sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$.

(ii) $W_n^{(2)} = \sqrt{n}(\hat{\boldsymbol{\beta}}_n^{\mathrm{RL}} - \boldsymbol{\beta}) \xrightarrow{\mathscr{D}} \mathscr{N}_p(-\boldsymbol{\delta}, \sigma^2\boldsymbol{A})$ *where* $\boldsymbol{\delta} = \boldsymbol{C}^{-1}\boldsymbol{H}^{\top}(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}^{\top})^{-1}\boldsymbol{\xi}$ *and* $\boldsymbol{A} = \boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{H}^{\top}(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}^{\top})^{-1}\boldsymbol{H}\boldsymbol{C}^{-1}$.

(iii) $W_n^{(3)} = \sqrt{n}(\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \hat{\boldsymbol{\beta}}_n^{\mathrm{RL}}) \xrightarrow{\mathscr{D}} \mathscr{N}_p(\delta, \sigma^2(C^{-1} - A))$.

(iv) $W_n^{(4)} = \boldsymbol{H}\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \boldsymbol{h} \xrightarrow{\mathscr{D}} \mathscr{N}_q(\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{h}, \sigma^2(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}^{\top}))$.

(v) $\begin{bmatrix} W_n^{(1)} \\ W_n^{(3)} \end{bmatrix} \xrightarrow{\mathscr{D}} \mathscr{N}_{2p}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\delta} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{C}^{-1} & \boldsymbol{C}^{-1} - \boldsymbol{A} \\ \boldsymbol{C}^{-1} - \boldsymbol{A} & \boldsymbol{C}^{-1} - \boldsymbol{A} \end{bmatrix}\right)$.

(vi) $\begin{bmatrix} W_n^{(2)} \\ W_n^{(3)} \end{bmatrix} \xrightarrow{\mathscr{D}} \mathscr{N}_{2p}\left(\begin{bmatrix} \boldsymbol{\delta} \\ -\boldsymbol{\delta} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C}^{-1} - \boldsymbol{A} \end{bmatrix} \right).$

(vii) $\begin{bmatrix} W_n^{(1)} \\ W_n^{(4)} \end{bmatrix} \xrightarrow{\mathscr{D}} \mathscr{N}_{p+q}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{H\beta} - \boldsymbol{h} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{C}^{-1} & \boldsymbol{C}^{-1}\boldsymbol{H}^{\mathsf{T}} \\ \boldsymbol{HC}^{-1} & \boldsymbol{HC}^{-1}\boldsymbol{H}^{\mathsf{T}} \end{bmatrix} \right).$

(viii) $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{\mathrm{SSL}} - \boldsymbol{\beta}) \xrightarrow{\mathscr{D}} \boldsymbol{W} - k\left\{ \dfrac{\boldsymbol{C}^{-1}\boldsymbol{H}^{\mathsf{T}}(\boldsymbol{HC}^{-1}\boldsymbol{H}^{\mathsf{T}})^{-1}(\boldsymbol{HW}+\boldsymbol{\xi})}{\sigma^{-2}(\boldsymbol{HW}+\boldsymbol{\xi})^{T}(\boldsymbol{HC}^{-1}\boldsymbol{H}^{\mathsf{T}})^{-1}(\boldsymbol{HW}+\boldsymbol{\xi})} \right\}.$

(ix) $\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}}_n^{\mathrm{PRSSL}} - \boldsymbol{\beta}) \xrightarrow{\mathscr{D}} \quad & \boldsymbol{W} - k\left\{ \dfrac{\boldsymbol{C}^{-1}\boldsymbol{H}^{\mathsf{T}}(\boldsymbol{HC}^{-1}\boldsymbol{H}^{\mathsf{T}})^{-1}(\boldsymbol{HW}+\boldsymbol{\xi})}{\sigma^{-2}(\boldsymbol{HW}+\boldsymbol{\xi})^{\mathsf{T}}(\boldsymbol{HC}^{-1}\boldsymbol{H}^{\mathsf{T}})^{-1}(\boldsymbol{HW}+\boldsymbol{\xi})} \right\} \\ & + \boldsymbol{C}^{-1}\boldsymbol{H}^{\mathsf{T}}(\boldsymbol{HC}^{-1}\boldsymbol{H}^{\mathsf{T}})^{-1}(\boldsymbol{HW}+\boldsymbol{\xi}) \\ & \times \left\{ 1 - \frac{k}{\sigma^{-2}(\boldsymbol{HW}+\boldsymbol{\xi})^{\mathsf{T}}(\boldsymbol{HC}^{-1}\boldsymbol{H}^{\mathsf{T}})^{-1}(\boldsymbol{HW}+\boldsymbol{\xi})} \right\} \\ & \times \boldsymbol{I}(\mathscr{L} < k). \end{aligned}$

where $\boldsymbol{W} \xrightarrow{\mathscr{D}} \mathscr{N}_p(\boldsymbol{0}, \sigma^2\boldsymbol{C}^{-1})$.

Based on the part (a) of Theorem 1, the distribution of the test statistics is obtained by Theorem 2.

Theorem 2 *Under the foregoing regularity conditions and local alternatives* $\mathscr{K}_{(n)}$*, if the LASSO satisfies the weakly consistent condition, i.e.,* $\lambda_n = o(n^{1/2})$*, the test statistics* $\mathscr{L}_n$ *defined in Eq. 8 converges in distribution to* $\mathscr{L}$*, which has the non central chi-square distribution with q degrees of freedom, non centrally parameter* $\Delta^2 = \sigma^{-2}\boldsymbol{\xi}^{\mathsf{T}}(\boldsymbol{HC}^{-1}\boldsymbol{H}^{\mathsf{T}})^{-1}\boldsymbol{\xi} = \sigma^{-2}\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{C}\boldsymbol{\delta}$ *where* $\boldsymbol{\delta} = \boldsymbol{C}^{-1}\boldsymbol{H}^{\mathsf{T}}(\boldsymbol{HC}^{-1}\boldsymbol{H}^{\mathsf{T}})^{-1}\boldsymbol{\xi}$*, and*

$$\mathscr{L} = \frac{(\boldsymbol{HW}+\boldsymbol{\xi})^{\mathsf{T}}(\boldsymbol{HC}^{-1}\boldsymbol{H}^{\mathsf{T}})^{-1}(\boldsymbol{HW}+\boldsymbol{\xi})}{\sigma^2}.$$

Proof. Rewrite the numerator of test statistics in Eq. (8) as

$$\begin{aligned} & \left(\boldsymbol{H}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \boldsymbol{\beta})\right) + \sqrt{n}(\boldsymbol{H\beta} - \boldsymbol{h})\right)^{\mathsf{T}} \left(\boldsymbol{H}(n\boldsymbol{C}_n^{-1})\boldsymbol{H}^{\mathsf{T}}\right)^{-1} \\ & \times \left(\boldsymbol{H}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \boldsymbol{\beta})\right) + \sqrt{n}(\boldsymbol{H\beta} - \boldsymbol{h})\right) \end{aligned} \tag{10}$$

Using part (i) of Theorem 1, $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \boldsymbol{\beta}\right)$ has the same asymptotic distribution as $\boldsymbol{W}$. Hence, under $\mathscr{K}_{(n)}$ and the regularity condition **A2**, Eq. (10) has the same distribution as

$$(\boldsymbol{HW}+\boldsymbol{\xi})^{\mathsf{T}}\left(\boldsymbol{HC}^{-1}\boldsymbol{H}^{\mathsf{T}}\right)^{-1}(\boldsymbol{HW}+\boldsymbol{\xi}) \tag{11}$$

On the other hand, by (i) of Theorem 1, it is obvious that $s_L^2 \to \sigma^2$. Using this fact together with Eq. (11), the result follows by Slutsky's theorem. ∎

The results of Theorems 1 and 2 can be used to derive ADB, ADQB, and ADQR.

To verify the consistency of the estimators, we have the following theorem and subsequent remarks.

Theorem 3 *Under the foregoing regularity conditions and local alternatives $\mathscr{K}_{(n)}$, we have the following as $n \to \infty$,*

(i) $\hat{\boldsymbol{\beta}}_n^{\mathrm{RL}} \xrightarrow{\mathscr{D}} \mathrm{argmin}(Z) - \boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal}(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal})^{-1}(\boldsymbol{H}\mathrm{argmin}(Z) - \boldsymbol{h})$.

(ii) $\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \hat{\boldsymbol{\beta}}_n^{\mathrm{RL}} \xrightarrow{\mathscr{D}} \boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal}(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal})^{-1}(\boldsymbol{H}\mathrm{argmin}(Z) - \boldsymbol{h})$.

(iii) $\hat{\boldsymbol{\beta}}_n^{\mathrm{PTL}} \xrightarrow{\mathscr{D}} \mathrm{argmin}(Z) - \boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal}(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal})^{-1}(\boldsymbol{H}\mathrm{argmin}(Z) - \boldsymbol{h})I(\mathscr{L} < \mathscr{L}_\alpha)$.

(iv) $\hat{\boldsymbol{\beta}}_n^{\mathrm{SSL}} \xrightarrow{\mathscr{D}} \mathrm{argmin}(Z) - k\boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal}(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal})^{-1}(\boldsymbol{H}\mathrm{argmin}(Z) - \boldsymbol{h})\mathscr{L}^{-1}$.

(v) $\hat{\boldsymbol{\beta}}_n^{\mathrm{PRSSL}} \xrightarrow{\mathscr{D}} \mathrm{argmin}(Z) - (k\mathscr{L}^{-1} + (1 - k\mathscr{L}^{-1})I(\mathscr{L} \le k))\boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal}(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal})^{-1}$
$\times(\boldsymbol{H}\mathrm{argmin}(Z) - \boldsymbol{h})$.

where $\mathscr{L}_\alpha$ is the upper critical value of chi-squared distribution with q d.f., $k = q - 2$, and $Z(\boldsymbol{\phi}) = (\phi - \boldsymbol{\beta})^{\intercal}\boldsymbol{C}(\phi - \boldsymbol{\beta}) + \lambda_0 \sum_{j=1}^{p} |\phi_j|$.

Proof. According to Theorem 2 of Knight and Fu (2000), if $\boldsymbol{C}$ is a nonsingular matrix and $\lambda_n/n \to \lambda_0 \ge 0$, then $\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} \xrightarrow{\mathscr{D}} \mathrm{argmin}(Z)$. To prove (i), by Slutsky's theorem, Eq. (4), and regularity condition (A2), we have

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \boldsymbol{C}_n^{-1}\boldsymbol{H}^{\intercal}(\boldsymbol{H}\boldsymbol{C}_n^{-1}\boldsymbol{H}^{\intercal})^{-1}(\boldsymbol{H}\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \boldsymbol{h}) \xrightarrow{\mathscr{D}} \mathrm{argmin}(Z) - \boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal}(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal})^{-1} \times(\boldsymbol{H}\mathrm{argmin}(Z) - \boldsymbol{h}).$$

(ii) By Eq. (4), we have $\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \hat{\boldsymbol{\beta}}_n^{\mathrm{RL}} = \boldsymbol{C}_n\boldsymbol{H}^{\intercal}(\boldsymbol{H}\boldsymbol{C}_n^{-1}\boldsymbol{H}^{\intercal})^{-1}(\boldsymbol{H}\hat{\boldsymbol{\beta}}_n^{\mathrm{L}} - \boldsymbol{h})$, which converges to $\boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal}(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal})^{-1}(\boldsymbol{H}\mathrm{argmin}(Z) - \boldsymbol{h})$. the result follows by Slutsky's theorem and regularity condition (A2). (iv) From Theorem 2, $I(\mathscr{L}_n \le \mathscr{L}_{n,\alpha}) \xrightarrow{\mathscr{D}} I(\mathscr{L} \le \mathscr{L}_\alpha)$. Making use of Eq. (5), (iii), and Slutsky's theorem, we have

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{PTL}} \xrightarrow{\mathscr{D}} \mathrm{argmin}(Z) - \boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal}(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}^{\intercal})^{-1}(\boldsymbol{H}\mathrm{argmin}(Z) - \boldsymbol{h})I(\mathscr{L} < \mathscr{L}_\alpha)$$

To prove (iv) and (v), since $k_n \to k = q - 2$, the result is obvious using Eq. (6), (iii), and Slutsky's theorem. ∎

Similar results as in Theorem 3 can be obtained using Theorem 2 of Knight and Fu (2000).

Remark 1 *Under the assumptions of Theorem 3 and $\lambda_n = o(n)$, we have the following results,*

(i) $\hat{\boldsymbol{\beta}}_n^{\mathrm{RL}} \xrightarrow{\mathscr{D}} \boldsymbol{\beta} - \boldsymbol{\delta}; \quad \boldsymbol{\delta} = \boldsymbol{C}^{-1}\boldsymbol{H}^{\mathsf{T}}(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}^{\mathsf{T}})^{-1}(\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{h})$.

(ii) $\hat{\boldsymbol{\beta}}_n^{\mathrm{PTL}} \xrightarrow{\mathscr{D}} \boldsymbol{\beta} - \boldsymbol{\delta} I(\mathscr{L} < \mathscr{L}_\alpha)$.

(iii) $\hat{\boldsymbol{\beta}}_n^{\mathrm{SSL}} \xrightarrow{\mathscr{D}} \boldsymbol{\beta} - \boldsymbol{\delta}\mathscr{L}^{-1}$.

(iv) $\hat{\boldsymbol{\beta}}_n^{\mathrm{PRSSL}} \xrightarrow{\mathscr{D}} \boldsymbol{\beta} - \left\{k\mathscr{L}^{-1} + (1 - k\mathscr{L}^{-1})I(\mathscr{L} < k)\right\}\boldsymbol{\delta}$.

Remark 2 *Under $\mathscr{H}_0$, all estimators are consistent for $\boldsymbol{\beta}$.*

4. Numerical analysis

In this section, we evaluate performance of the proposed estimators using a simulation study along with a real example.

4.1. Simulation

In this section, we conduct a Monte Carlo simulation to analyse relative efficiencies with respect to different levels of sparsity. In particular, we use $\mathrm{RE}(\hat{\boldsymbol{\beta}}^*; \hat{\boldsymbol{\beta}}^{\mathrm{L}}) = \mathrm{R}(\hat{\boldsymbol{\beta}}^{\mathrm{L}})/\mathrm{R}(\hat{\boldsymbol{\beta}}^*)$, where $\hat{\boldsymbol{\beta}}^*$ is one of the proposed estimators in this paper.

We generate a matrix $\boldsymbol{X}$ from a multivariate normal distribution with mean vector $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The off-diagonal elements of the covariance matrix are considered to be equal to r with $r = 0, 0.2, 0.9$. We consider $n = 100$ and various p ranging 10, 15, and 20.

One of the most applicable $\boldsymbol{H}$ and $\boldsymbol{h}$ is to select variables. Sometimes, an expert claims that some variables do not affect regression model. If we suppose $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathsf{T}}, \boldsymbol{\beta}_2^{\mathsf{T}})^{\mathsf{T}}$, then $\boldsymbol{\beta}_2 = \mathbf{0}$ is equivalent to the variables that may be ignored for predicting model.

Let us consider $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^{\mathsf{T}}, \boldsymbol{\beta}_2^{\mathsf{T}}\right)^{\mathsf{T}} = \left(\mathbf{1}_{p-q}^{\mathsf{T}}, \mathbf{0}_q^{\mathsf{T}}\right)^{\mathsf{T}}$, where $\mathbf{1}_{p-q}$ and $\mathbf{0}_q$ stand for the vectors of 1 and 0 with dimensions $p-q$ and q, respectively. In order to investigate the behaviour of the proposed estimators, we define $\Delta^* = \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|$, where $\boldsymbol{\beta}_0 = \left(\mathbf{1}_{p-q}^{\mathsf{T}}, \mathbf{0}_q^{\mathsf{T}}\right)^{\mathsf{T}}$ and $\|\cdot\|$ is the Euclidean norm. If $\Delta^* = 0$, then $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ while $\boldsymbol{\beta} = (\mathbf{1}_{p-q}^{\mathsf{T}}, \boldsymbol{\Delta}^{\mathsf{T}})^{\mathsf{T}}$ when $\Delta^* > 0$, where $\boldsymbol{\Delta} = (\Delta, \ldots, \Delta)^{\mathsf{T}}$ is the q-dimensional vector of Δ values. When we increase the number of Δ^*, it indicates the degree of violation of the null hypothesis.

In our simulation study, without loss of generality, we assume $\boldsymbol{\beta}$ is a p-vector in which the first s components of $\boldsymbol{\beta}$ are $\mathbf{1}$ and other $(p-s)$ components are zero. The responses were simulated from the following model:

$$y_i = \sum_{i=1}^{p} x_i\beta_i + e_i, \qquad e_i \sim \mathscr{N}(0,1)$$

Each realization was repeated 1000 times to obtain risk of the estimated regression parameters. Thus, risks are calculated for the ULE, RLE, PTLE, SSLE and PRSSLE. The results are tabulated in Tables 1-3.

The findings of Tables 1-3 may be summarized as:

a) When the null hypothesis is true ($\Delta^2 = 0$), RLE behaves better than other estimator. As we depart from the null hypothesis, the performance of this estimator decreases.

b) For large Δ^2, the performance of estimators decreases; even, when the correlation is low, the unrestricted LASSO performs better.

c) Neither PTLE nor Stein-type shrinkage LASSO estimator dominates each other.

d) The positive rule Stein-type shrinkage LASSO uniformly dominates Stein-type LASSO estimator.

e) It is well - known that shrinkage and positive-rule shrinkage estimators are always better than unrestricted estimator. Here, the results confirm that also.

Table 1: *Relative efficiencies (standard errors) of the estimators for fixed* Δ^2, $r = 0$, $s = 6$ *different values of* p.

	ULE		RLE		PTLE		SSLE		PRSSLE	
p					$\Delta^2 = 0$					
10	1	(0.003)	1.63	(0.011)	1.52	(0.011)	1.23	(0.008)	1.35	(0.009)
15	1	(0.002)	2.32	(0.013)	2.19	(0.013)	1.75	(0.007)	2.05	(0.010)
20	1	(0.001)	3.51	(0.007)	2.98	(0.007)	2.44	(0.004)	2.99	(0.006)
p					$\Delta^2 = 0.1$					
10	1	(0.001)	1.57	(0.003)	1.46	(0.003)	1.21	(0.003)	1.32	(0.003)
15	1	(0.001)	2.18	(0.007)	2.02	(0.007)	1.66	(0.005)	1.95	(0.006)
20	1	(0.001)	3.48	(0.015)	2.92	(0.015)	2.40	(0.008)	2.98	(0.011)
p					$\Delta^2 = 0.5$					
10	1	(0.001)	0.85	(0.000)	0.86	(0.001)	1.05	(0.001)	1.07	(0.001)
15	1	(0.002)	1.57	(0.002)	1.42	(0.003)	1.37	(0.003)	1.54	(0.003)
20	1	(0.001)	2.86	(0.004)	2.34	(0.004)	2.19	(0.004)	2.58	(0.004)
p					$\Delta^2 = 1$					
10	1	(0.001)	0.36	(0.000)	0.90	(0.001)	1.00	(0.001)	1.00	(0.001)
15	1	(0.000)	0.82	(0.000)	0.86	(0.000)	1.12	(0.000)	1.14	(0.000)
20	1	(0.003)	1.81	(0.002)	1.43	(0.004)	1.81	(0.005)	1.90	(0.005)
p					$\Delta^2 = 5$					
10	1	(0.001)	0.02	(0.007)	1.00	(0.007)	0.94	(0.004)	0.94	(0.006)
15	1	(0.001)	0.06	(0.000)	1.00	(0.000)	0.98	(0.000)	0.98	(0.000)
20	1	(0.001)	0.13	(0.000)	1.00	(0.001)	1.00	(0.001)	1.00	(0.001)

Table 2: *Relative efficiencies (standard errors) of the estimators for fixed* Δ^2, $r = 0.2$, $s = 6$ *different values of* p.

	ULE		RLE		PTLE		SSLE		PRSSLE	
p					$\Delta^2 = 0$					
10	1	(0.003)	1.95	(0.014)	1.71	(0.014)	1.34	(0.008)	1.44	(0.011)
15	1	(0.002)	2.70	(0.015)	2.37	(0.016)	1.90	(0.012)	2.28	(0.014)
20	1	(0.000)	4.77	(0.010)	3.55	(0.009)	2.98	(0.006)	3.63	(0.008)
p					$\Delta^2 = 0.1$					
10	1	(0.001)	1.93	(0.004)	1.61	(0.004)	1.33	(0.002)	1.40	(0.003)
15	1	(0.001)	2.69	(0.009)	2.30	(0.010)	1.92	(0.006)	2.21	(0.008)
20	1	(0.001)	4.72	(0.010)	3.39	(0.009)	2.96	(0.006)	3.62	(0.008)
p					$\Delta^2 = 0.5$					
10	1	(0.001)	0.97	(0.000)	0.91	(0.001)	1.17	(0.002)	1.17	(0.002)
15	1	(0.002)	1.87	(0.003)	1.30	(0.004)	1.70	(0.004)	1.74	(0.004)
20	1	(0.001)	3.74	(0.005)	1.93	(0.006)	2.75	(0.005)	2.91	(0.005)
p					$\Delta^2 = 1$					
10	1	(0.001)	0.37	(0.000)	0.99	(0.001)	1.08	(0.001)	1.08	(0.001)
15	1	(0.000)	0.85	(0.000)	0.97	(0.000)	1.37	(0.000)	1.37	(0.000)
20	1	(0.003)	1.85	(0.000)	1.07	(0.003)	2.10	(0.007)	2.10	(0.007)
p					$\Delta^2 = 5$					
10	1	(0.001)	0.01	(0.001)	1.00	(0.001)	0.99	(0.001)	0.99	(0.001)
15	1	(0.001)	0.04	(0.000)	1.00	(0.000)	1.02	(0.001)	1.02	(0.001)
20	1	(0.001)	0.85	(0.000)	1.00	(0.001)	1.13	(0.001)	1.13	(0.001)

The linear regression model is fitted to this dataset in order to predict the response variable. The LASSO of Tibshirani (1996) (the UL in our study), restricted LASSO (RL), preliminary test LASSO (PTL), Stein-type shrinkage LASSO (SSL), and positive rule Stein-type shrinkage (PRSSL) estimators are used to estimate the unknown regression coefficients.

Since one of the biggest problems in estimation is to determine $\boldsymbol{H}$ and $\boldsymbol{h}$, we suppose that $\boldsymbol{H} = \boldsymbol{I}_7$. This choice is just for simplicity and also to avoid errors obtained by incorrect selection of parameters.

In order to show the impact of correctness or incorrectness of hypothesis, we consider the following two cases:

Case I. Let $\boldsymbol{h} = (0,0,10,0.2,0.7,0.06,0)^\top$. The null hypothesis changes into $\mathscr{H}_o : \boldsymbol{\beta} = \boldsymbol{h}$ and thus, the variables POPULATION, INCOME, and AREA are insignificant.

Case II. Let $\boldsymbol{h} = (0,0,0,0,0,0,0)^\top$. The null hypothesis changes into $\mathscr{H}_o : \boldsymbol{\beta} = \mathbf{0}$ and thus, all variables are insignificant.

Table 3: *Relative efficiencies (standard errors) of the estimators for fixed* Δ^2, $r = 0.9$, $s = 6$ *different values of* p.

	ULE		RLE		PTLE		SSLE		PRSSLE	
p					$\Delta^2 = 0$					
10	1	(0.002)	6.76	(0.051)	3.34	(0.052)	1.86	(0.019)	1.86	(0.019)
15	1	(0.001)	8.46	(0.006)	5.76	(0.061)	4.28	(0.036)	4.68	(0.042)
20	1	(0.000)	14.74	(0.064)	14.48	(0.064)	6.13	(0.065)	11.31	(0.057)
p					$\Delta^2 = 0.1$					
10	1	(0.001)	6.35	(0.062)	2.95	(0.017)	1.81	(0.004)	1.82	(0.004)
15	1	(0.001)	8.31	(0.004)	5.77	(0.041)	4.28	(0.024)	4.68	(0.031)
20	1	(0.001)	14.11	(0.085)	12.56	(0.085)	5.96	(0.052)	10.68	(0.082)
p					$\Delta^2 = 0.5$					
10	1	(0.001)	3.28	(0.003)	1.39	(0.005)	1.69	(0.005)	1.69	(0.005)
15	1	(0.002)	5.41	(0.017)	2.71	(0.020)	3.85	(0.028)	3.92	(0.028)
20	1	(0.001)	10.40	(0.020)	7.10	(0.021)	6.77	(0.032)	8.68	(0.031)
p					$\Delta^2 = 1$					
10	1	(0.002)	1.18	(0.000)	0.96	(0.001)	1.55	(0.005)	1.55	(0.005)
15	1	(0.000)	2.50	(0.000)	1.26	(0.001)	3.14	(0.003)	3.14	(0.003)
20	1	(0.002)	5.28	(0.012)	2.57	(0.019)	5.33	(0.053)	6.58	(0.053)
p					$\Delta^2 = 5$					
10	1	(0.005)	0.02	(0.000)	1.00	(0.005)	0.87	(0.004)	0.87	(0.004)
15	1	(0.002)	0.06	(0.000)	1.00	(0.002)	1.58	(0.004)	1.58	(0.004)
20	1	(0.001)	0.19	(0.000)	1.00	(0.001)	2.68	(0.006)	2.68	(0.006)

Table 4: *Description of the variables of state.x77.*

Variables	Description	Role
LifeExp	Average years of life expectancy at birth	Response
Population	in thousands	Predictor
Income	dollars per capita	Independent
Illiteracy	Percentage of those unable to read and write	Independent
Murder	number of murders and non-negligent manslaughters per 100000 people	Independent
HS Grad	percentage of adults who were high-school graduates	Independent
Frost	mean number of days per year with low temperatures below freezing	Independent
Area	in square miles	Independent

4.2. Real data

In this section, we study the performance of proposed LASSO-based shrinkage estimators using state.x77 dataset (available by default in R software). Descriptions of the variables in this dataset are given in Table 4.

Table 5: *5-fold cross validation relative average prediction errors for state data.*

	RLE	PTLE			SSLE	PRSSLE
		0.01	0.05	0.10		
Case I	22.2615	1.0009	1.0004	1.0004	1.0200	1.0208
Case II	1.0017	1.0000	1.0000	1.0000	1.0000	1.0008

The performance of the estimators are evaluated using average five-fold cross validation error. By choosing 1000 as a large enough number for repeating process in a bootstrap simulation scheme, Table 5 shows the relative average prediction errors in the two cases.

Based on Table 5, RLE is the best estimator because the hypothesis $\boldsymbol{H\beta} = \boldsymbol{h}$ is nearly true, but PRSSLE has lower prediction error than other estimators in case I. This estimator is followed by SSLE. Indeed, by departing from the null hypothesis, these estimators will behave similar to the LASSO in case II. If the level of significance α for constructing PTLE increases, then the prediction error decreases.

5. Conclusion

In this paper, we proposed improved LASSO-based estimators by imposing a subspace restriction to the linear regression model. Particularly, we introduced preliminary-test LASSO, Stein-type shrinkage LASSO, and positive-rule shrinkage LASSO estimators. Asymptotic performance of the proposed estimators studied in case $n > p$. The proposed methodology for improving the LASSO can also be applied to the high-dimensional case $p > n$. Indeed the test statistic for $\mathscr{H}_o : \boldsymbol{H\beta} = \boldsymbol{h}$ plays a determining role.

In addition to the given theorems for the asymptotic behaviour of the proposed estimators, using a simulation study, we compared the performance of estimators numerically for various configurations of p, correlation coefficient between the predictors (r), and the error in variance (σ^2). For different non-centrality parameter Δ, degree of model misspecification, the number of non-zero βs varied, and then the performance of estimators evaluated. We found that the positive-rule shrinkage LASSO estimator has the best performance among all. When we deviated from the null model, neither PTLE nor SSLE dominated one another and the PTLE performed better as α became large. Relative efficiency of the proposed estimators increased when there were more near-zero parameters in the model. As an application, a real dataset was analysed, where a five-fold cross-validation averages and standard deviations of the prediction errors were evaluated for the LASSO and its other four variants. The new estimators dominated the LASSO in average prediction error sense.

Acknowledgments

We would like to thank the Editor, Prof. Maria Durban, and anonymous referees for their valuable suggestions and comments which substantially improved the presentation and led to adding many details in the paper. Second author M. Arashi's work is based on the research supported in part by the National Research Foundation of South Africa (Grant NO. 109214).

References

Ahmed, S.E. and Raheem, S.M.E. (2012). Shrinkage and absolute penalty estimation in linear regression models, *Wires: Computational Statistics*, 4, 541–553.

Fallahpour, S., Ahmed, S.E. and Doksum, K.A. (2012). L1 penalty and shrinkage estimation in partially linear models with random coefficient autoregressive errors, *Applied Stochastic Models in Business and Industry*, 28, 236–250.

Hossain, S. and Ahmed, S.E. (2014). Penalized and Shrinkage Estimation in the Cox Proportional Hazards Model. *Communications in Statistics-Theory and Methods*, 43, 1026–1040.

Hossain, S., Ahmed, S.E. and Doksum, K.A. (2015). Shrinkage, pretest, and penalty estimators in generalized linear models. *Statistical Methodology*, 24, 52–68.

Hossain, S. and Ahmed, S.E. and Yi, Y. (2016). Shrinkage and pretest estimators for longitudinal data analysis under partially linear models. *Journal of Nonparametric Statistics*, DOI:10.1080/10485252.2016.1190358.

Hossain, S. and Howlader, H. (2016). Shrinkage estimation in lognormal regression model for censored data. *Journal of Applied Statistics*, DOI: 10.1080/02664763.2016.1168365.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. Berkeley, Calif.: University of California Press, 361–379.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, 28, 1356–1378.

Roozbeh, M. (2015). Shrinkage ridge estimators in semiparametric regression models. *Journal of Multivariate Analysis*, 136, 56–74.

Roozbeh, M. (2016). Robust ridge estimator in restricted semiparametric regression models. *Journal of Multivariate Analysis*, 147, 127–144.

Saleh, A.K.M.E. (2006). *Theory of preliminary test and stein-type estimation with applications*, John Wiley & Sons, New York.

Sengupta, D. and Jammalamadaka, S.R. (2003). *Linear models: An integrated approach*, World Scientific Publishing Company.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B.*, 58, 267–288.

Yuzbasi, B. and Ahmed, S.E. (2016). Shrinkage and penalized estimation in semi-parametric models with multicollinear data. *Journal of Statistical Computation and Simulation*, 86, 3543–3561.

Yuzbasi, B., Ahmed, S.E. and Gungor, M. (2017). Improved Penalty Strategies in Linear Regression Models. *REVSTAT-Statistical Journal*, Accepted.

SORT 42 (1) January-June 2018, 59-72 DOI: 10.2436/20.8080.02.69

Heteroscedasticity irrelevance when testing means difference

Pablo Flores M.[1] and Jordi Ocaña[2]

Abstract

Heteroscedasticity produces a lack of type I error control in Student's t test for difference between means. Pretesting for it (e.g., by means of Levene's test) should be avoided as this also induces type I error. These pretests are inadequate for their objective: not rejecting the null hypotheses is not a proof of homoscedasticity; and rejecting it may simply suggest an irrelevant heteroscedasticity. We propose a method to establish irrelevance limits for the ratio of variances. In conjunction with a test for dispersion equivalence, this appears to be a more affordable pretesting strategy.

MSC: 62F03.

Keywords: Homoscedasticity, equivalence test, indifference zone, pretest, Student's t test

1. Introduction

Student's t test for determining possible inequalities between two population means is subject to normality and homoscedasticity assumptions. (Readers not familiar with basic statistical techniques, such as Student's or Welch's test, may refer to sources like the SPSS tutorial at https://libguides.library.kent.edu/SPSS/IndependentTTest.) Presumably, these assumptions are or are not confirmed by means of other (pre)tests on the same data. The pretests (or the order in which they are applied) may vary. When the null hypothesis of a normality test (Shapiro-Wilk, Kolmogorov-Smirnov, etc.) is rejected, the traditional procedure is to assume that the sample does not come from a normal distribution. In such cases, a non-parametric approach is adopted, for example the Wilcoxon's test to compare the location parameters of two independent samples – possibly under the additional yet false assumption of its supposedly higher robustness to dispersion differences. Otherwise, when the null hypothesis of normality is not rejected, this assumption

[1] Grupo de Investigación en Ciencia de Datos CIDED, Escuela Superior Politécnica de Chimborazo, Facultad de Ciencias, Panamericana Sur km 1 1/2, EC060155 Riobamba, Ecuador; email: p_flores@espoch.edu.ec

[2] Departament de Genètica, Microbiologia i Estadística, Secció d'Estadística, Universitat de Barcelona, Facultat de Biologia, Diagonal 643, 08028 Barcelona, Spain; email: jocana@ub.edu

Received: December 2017
Accepted: May 2018

is taken as true; and pretesting then proceeds to the next step by means of a test with perfect homoscedasticity as the null hypothesis (F, Levene, Bartlett, Cochran, etc.). If its null hypothesis is not rejected, then homoscedasticity is taken as true. This leads to the use of Student's t test as an adequate procedure for comparing means. Otherwise, heteroscedasticity is assumed and a procedure like Welch's test (Welch, 1947) is adopted.

Although it is not unusual to find such pretesting recommendations, several studies (Hsu, 1938; Overall, Atlas and Gibson, 1995; Scheffé, 1970) show that these strategies alter the overall type I error probability (TIEP) especially when sample sizes are unequal. Zimmerman (2004) performed a simulation study using different sample sizes, levels of heteroscedasticity and levels of significance to estimate the overall TIEP. The results showed that when Student's test is performed without any homoscedasticity pretesting, and when Levene's pretest is used to decide between Student's or Welch's test, the overall TIEP is severely inflated. On the other hand, the TIEP for Welch's test remains close to the significance level for all heteroscedasticity levels. In strategies that alter the TIEP, the largest variance associated with the largest sample size deflates the TIEP while it is inflated when the largest variance is associated with the smallest sample size. The severity of this distortion increases with the heteroscedasticity level. In addition, overall TIEP distortion increases as the significance level of the pretest decreases, the overall TIEP ceases to be affected at high levels of significance in the preliminary test, e.g., at the non-usual value $\alpha = 0.20$.

Rasch, Kubinger and Moder (2011), state that pretesting to validate the assumptions in the comparison of means test leads to alterations in the type I and type II error probabilities. These authors show that using a pretest for normality (Kolmogoroff-Smirnov) and a pretest for equality of variances (Levene) causes an increase in the overall TIEP. In contrast, when Welch's test is used directly (without a pretest), these overall TIEP distortions largely disappear. They conclude that pretesting does not pay off. Instead, applying Welch's test directly without pretesting is best, and it should be recommended in textbooks as well as implemented in statistical software as the standard option for comparing means. In addition, the authors advise that Wilcoxon's and Student's t-test should never be used.

The next section introduces some concepts and notation in equivalence testing. Section three describes the algorithm that we have used to determine these irrelevance limits. In the fourth section, a simulation study comparing the previously cited pretesting strategies is presented. In the fifth section, two illustrative examples are presented. Finally, in the last section the main conclusions are discussed.

2. Equivalence testing concepts and some additional notation

The above results contribute to other evidence indicating that pretesting in order to fulfill validity conditions (not only in the problem of means comparison) is not a reliable strategy. However, one may ask if this inadequacy is (fully or partially) due to the fact that these pretests are intrinsically inappropriate for their goal: Note that their null hypothesis states complete fulfilment of the normality or homoscedasticity assumptions. As is well known, not rejecting the null hypothesis is not a proof of its correctness, while rejecting it may simply indicate an irrelevant departure from perfect normality or homoscedasticity. In other words, asserting that there is a non-significant difference between variances should not be confused with there being homogeneity. In the words of Altman and Bland (1995) "*Absence of evidence is not evidence of absence*".

Figure 1 schematically shows these ideas in the specific case of the homoscedasticity assumption, which will constitute the focus of the present paper.

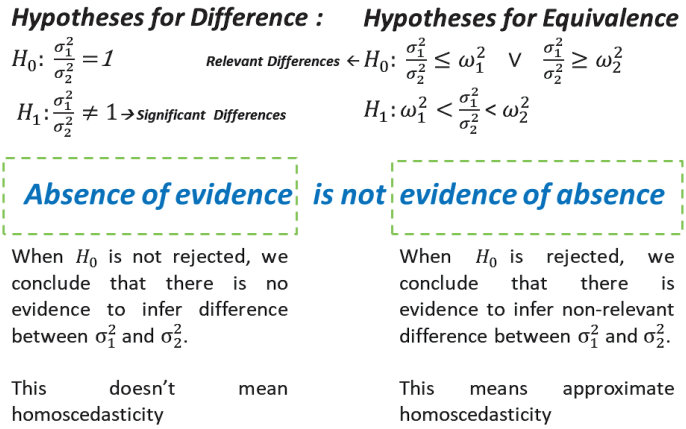

Figure 1: *Traditional and equivalence approach.*

Wellek (2010) (p. 164), proposes an approach that is based on equivalence testing. In this class of tests, the alternative hypothesis states equivalence, i.e., perfect fit (to normal) or equality (of variances) **except for irrelevant deviations** while the null hypothesis states relevant ones. In this approach, the relevant differences between variances are stated in the null hypothesis; thus the assumption of near homoscedasticity is reinforced if the null is rejected.

In brief, Wellek's test may be described as follows: For the hypotheses

$$\begin{aligned} H_0 &: \frac{\sigma_1^2}{\sigma_2^2} \leq \omega_1^2 \wedge \frac{\sigma_1^2}{\sigma_2^2} \geq \omega_2^2 && \text{No equivalence (relevant difference of variances)} \\ H_1 &: \omega_1^2 < \frac{\sigma_1^2}{\sigma_2^2} < \omega_2^2 && \text{Equivalence (non-relevant difference)} \end{aligned} \tag{1}$$

with $\omega_1^2 < 1 < \omega_2^2$, a uniformly more powerful invariant test is one whose critical region is given by:

$$\{\widetilde{C}^{(1)}_{\alpha,n_1-1,n_2-1}(\omega_1^2,\omega_2^2) < Q < \widetilde{C}^{(2)}_{\alpha,n_1-1,n_2-1}(\omega_1^2,\omega_2^2)\},$$

where Q stands for the test statistic:

$$Q = \frac{S_X^2}{S_Y^2} = \frac{(n_2-1)}{(n_1-1)} \frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2}{\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2} \tag{2}$$

and the critical constants $\widetilde{C}^{(1)}_{\alpha,n_1-1,n_2-1}(\omega_1^2,\omega_2^2)$, $\widetilde{C}^{(2)}_{\alpha,n_1-1,n_2-1}(\omega_1^2,\omega_2^2)$, are determined by the equations:

$$F_{n_1-1,n_2-1}(\widetilde{C}_2/\omega_1^2) - F_{n_1-1,n_2-1}(\widetilde{C}_1/\omega_1^2) = \alpha = F_{n_1-1,n_2-1}(\widetilde{C}_2/\omega_2^2) - F_{n_1-1,n_2-1}(\widetilde{C}_1/\omega_2^2).$$

$F_{n_1-1,n_2-1}()$ corresponds to the cumulative distribution function of a centred F distribution, with n_1-1 degrees of freedom in the numerator and n_2-1 in the denominator.

One of the most important aspects of equivalence testing is to establish the equivalence limits. Wellek does not propose a technical criterion to determine them, instead he provides some hints based on what he calls "a common statistical sense", which may not be enough in many applications. For this reason, here we develop a procedure that allows us to calculate these limits in the specific problem of determining (enough) homoscedasticity when the end objective is to perform a comparison of means and assuming that normality is fulfilled. As its input, the procedure requires objective information on the experimental design and (admittedly, less objective) information on the tolerable possible distortion in the TIEP (perhaps with the help of "common statistical sense").

3. Irrelevance limits for the ratio of dispersions of two Gaussian distributions

As mentioned above, using an equivalence dispersion test for two Gaussian distributions as an homoscedasticity pretest overcomes the logical difficulty of approaches like the F test when it is used for the same purpose. However, the equivalence approach has a notable ambiguity: The values of the equivalence or irrelevance limits (ω_1^2, ω_2^2) that define the hypotheses to be tested must be specified. Criteria such as common statistical sense or the researchers prior knowledge on their subject of interest may be subjective and insufficient.

If the equivalence test refers to a parameter involved in a validity requirement for another test – for example, the ratio of variances for Student's t test – then one possibility is to define an irrelevance limit $\delta > 0$ for the difference between the true TIEP and the significance level α. Obviously $\alpha \pm \delta$ must be inside the $(0,1)$ interval. This irrelevance (or permissiveness or indifference) parameter δ is the maximum distance above and

below α that is acceptable as an irrelevant affectation of the TIEP. This approach may seem to be imprecise and prone to arbitrariness but it follows the line of thought (fairly correct in our opinion) that these validity conditions are just idealizations. Possibly, perfect normality and perfect homoscedasticity are never present in nature. As Box (1979) states about normality, a normal distribution does not exist in the real world, but models known to be false often derive in useful approximate results; what is really important is not whether the populations "are normal" but knowing if the approximate model is good enough to be useful. In our approach the approximation will or will not be considered good based on how close the true TIEP is to the nominal significance level. In this same sense, Cochran (1942) suggested that a distance of 20% of the true TIEP from the nominal significance level is an acceptable approximation. This authoritative criterion, known sometimes as "Cochran's Criterion", could be used as the default in algorithms implementing the method proposed here.

In Student's t test, the true TIEP is a continuous function of the population ratio of variances $\omega^2 = \sigma_1^2/\sigma_2^2$ and its value equals the nominal significance level α at $\omega^2 = 1$. From this point of complete homoscedasticity (and depending also on the sample sizes), this TIEP function may be of an increasing or decreasing nature. As a consequence, ω_1^2 and ω_2^2 define an interval around 1 and they correspond to the ratio ω^2 values where the TIEP equals $\alpha - \delta$ or $\alpha + \delta$.

Given a nominal significance level α, a degree of permissiveness δ and sample sizes n_1, n_2, the procedure for obtaining the pair (ω_1^2, ω_2^2) is based on a simulation iterative process. More precisely, starting from a ratio in the neighbourhood of $\omega^2 = 1$, the true TIEP of Student's t test is obtained by simulation, as the proportion of null hypothesis rejections. This process is iterated by progressively decrementing or incrementing this ratio until crossing the threshold $\alpha \pm \delta$ and until the TIEP reaches these limits with a given precision. The following additional safeguard is included: Provided that the resulting TIEP in each simulation iteration is just an estimation of the true TIEP, the algorithms implementing the method may require that a confidence interval for the true TIEP must be fully included inside $\alpha \pm \delta$.

The simulation process is fast because, to repeatedly generate Student's t statistic values, it is not necessary to simulate pairs of independent Gaussian full data samples of sizes n_1 and n_2, respectively, and then compute the t statistic from them. Instead, provided that we are simulating under a Student's t test scenario of true null hypothesis, the difference of the sample means (the numerator of the t statistic) can be directly generated from a Gaussian distribution with zero mean and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. In addition, the sum of squares necessary for computing the pooled variance estimate, $\sum_{i=1}^{n_1}\left(X_{1i} - \overline{X}_1\right)^2 + \sum_{i=1}^{n_2}\left(X_{2i} - \overline{X}_2\right)^2$ can be directly generated as the sum of two independent values (and also independently from the difference between sample means) that are generated from a distribution $\sigma_i^2 \chi_{n_i-1}^2, i = 1,2$, where χ_ν^2 stands for a chi-square distribution with ν degrees of freedom. A further simplification comes from the fact that the only relevant parameter is the ratio of variances and not the variances themselves; thus, one of the variances to be simulated can be fixed at one. What is more,

because complete symmetry exists between the equivalence limits in balanced cases, it is sufficient to obtain only one of them, e.g., the second one, ω_2^2, and then compute $\omega_1^2 = 1/\omega_2^2$. Finally, a variance reduction technique based on the method of"control variates" is applied to avoid the need for very large numbers of simulation replicates to deliver acceptable precision. This technique is also applied in the simulations described in the next chapter, and it is explained in the Appendix.

Table 1: *Indifference zone* (ω_1^2, ω_2^2) *with* $\delta = 0.2\alpha$.

	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$n = (5,5)$	(0.130 - 7.691)	(0.225 - 4.428)	(0.397 - 2.519)
$n = (3,7)$	(0.709 - 1.412)	(0.779 - 1.289)	(0.819 - 1.163)
$n = (7,3)$	(0.711 - 1.410)	(0.776 - 1.325)	(0.832 - 1.166)
$n = (10,10)$	(0.002 - 501.0)	(0.097 - 10.325)	(0.282 - 3.542)
$n = (6,14)$	(0.727 - 1.408)	(0.783 - 1.292)	(0.846 - 1.157)
$n = (14,6)$	(0.716 - 1.362)	(0.787 - 1.264)	(0.859 - 1.148)
$n = (5,10)$	(0.679 - 1.387)	(0.741 - 1.286)	(0.819 - 1.196)
$n = (10,5)$	(0.716 - 1.452)	(0.786 - 1.331)	(0.862 - 1.256)

For illustrative purposes, Table 1 displays the irrelevance limits for some sample sizes (balanced and unbalanced) and significance level scenarios. These values were obtained from 100000 simulation replicates. The results show that, first, there is more heteroscedasticity permissiveness (wider irrelevance intervals) in the balanced scenarios than in the unbalanced ones and, second, that larger sample sizes correspond to wider irrelevance intervals in the balanced cases.

4. Results on pretesting homoscedasticity

4.1. Overall TIEP affectation when the F pretest is used to verify the homoscedasticity assumption

Many tests have been developed for the hypotheses

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$
$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1, \tag{3}$$

to eventually prove heteroscedasticity – and not homoscedasticity. As has been previously stated, some studies use Levene's test as their pretesting option. Provided that the test for heteroscedasticity irrelevance considered in this paper is based on the ratio

Q of sample variances and the Fisher-Snedecor **F** distribution, for the sake of comparison we consider here the traditional F test that is based on the Q statistic to prove heteroscedasticity and we then use it as a reference for comparison with the equivalence approach. However, very similar results were delivered by complementary simulations using Levene's test and other tests for heteroscedasticity (not presented here). At this point, it would be fair to advise against the widespread use of the F test given its lack of robustness in front of departures from normality (see, for example, point 4.3 in Rasch and Guiard, 2004). These drawbacks do not invalidate the results in the present paper because we assume and simulate under perfect normality of data conditions. However, these considerations may be of obvious practical interest.

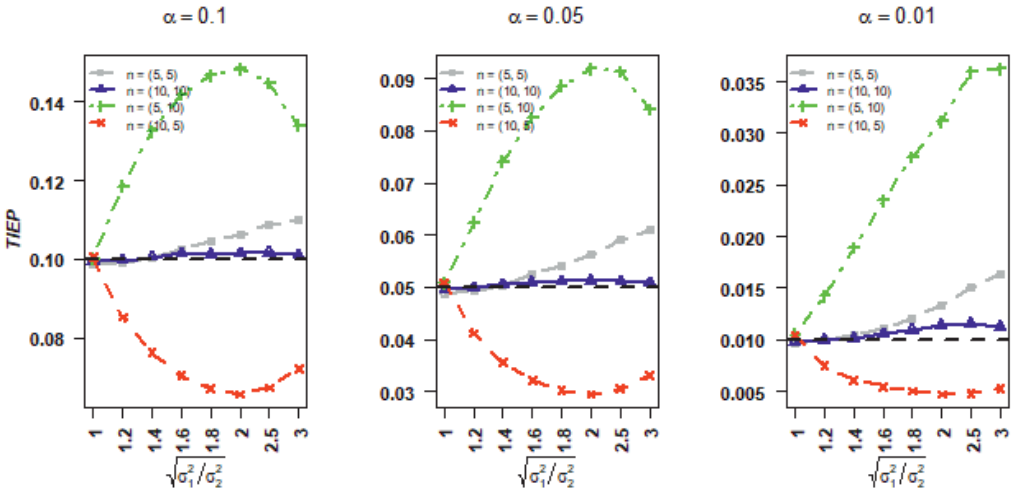

Figure 2: *Overall TIEP estimation when Student's t or Welch's test are conditioned to the result of the F test: If the null hypothesis of variances equality is not rejected, then Student's t test is applied; otherwise Welch's test is applied. The scale of the TIEP axis differs in accordance with the different significance levels under consideration. The relative distance from the nominal significance level is of importance here.*

Figure 2 illustrates similar results to those obtained in the references cited in this paper. They were obtained from 100000 simulation replicates and correspond to scenarios defined by crossing significance levels of $\alpha = 0.1$, 0.05 and 0.01; several heteroscedasticity degrees given by the ratio $\omega = \sqrt{\sigma_1^2/\sigma_2^2}$; and sample sizes that are balanced ($n = (5,5)$ and $n = (10,10)$), and unbalanced ($n = (5,10)$ and $n = (10,5)$), always under equality of population means. Independently of the significance level for comparison of means, all F pretests were performed at a fixed 0.05 significance level.

These results agree with those obtained in the previous studies: There is inflation or deflation in the overall TIEP when the decision to use Student's t or Welch's test is conditioned to the result of a pretest (here, the F test) to (supposedly) verify homoscedasticity. This affectation is clearly less concerning in the case of balanced sample sizes as well as with growing sample sizes. However, when there are few observations and/or unbal-

ancing, the affectation increases considerably as the level of heteroscedasticity grows; so once again we verify that performing this type of pretest is a bad strategy.

4.2. Overall TIEP affectation when the equivalence dispersion pretest is used to verify the homoscedasticity assumption

Figure 3 shows comparable simulation results when the Wellek's equivalence pretest is used, and once the zone of indifference (ω_1^2, ω_2^2) has been determined for each significance level (of the comparison of means test) and sample sizes scenario. The δ values correspond to those suggested by Cochran's criterion, with a tolerance limit for the TIEP equal to 20% of the significance level. We observe much greater control of the TIEP (not perfect, but in any case within the irrelevance limits) with values much closer to the significance level than when pretesting was entrusted to the F test. Independently of the comparison of means significance level, all of the equivalence pretests were performed at a fixed 0.05 significance level.

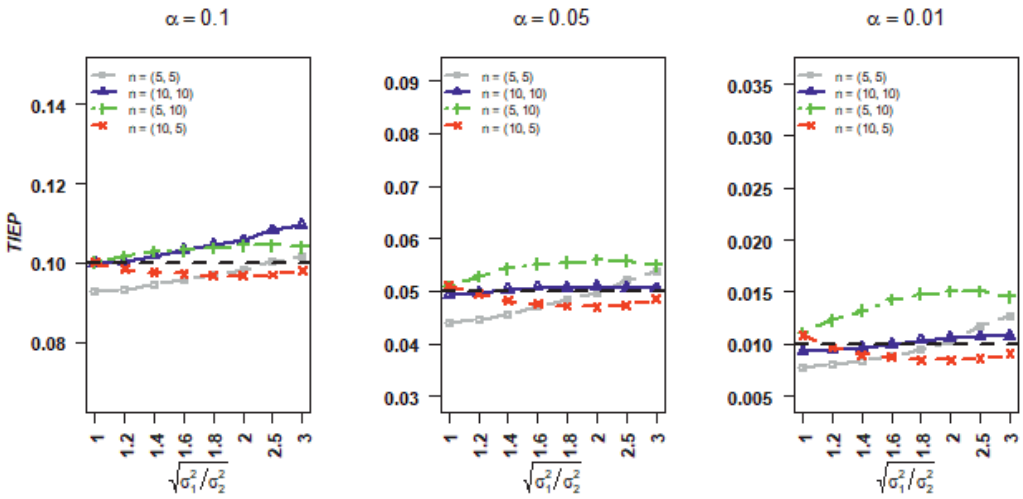

Figure 3: *Overall TIEP estimation when Student's t test or Welch's test are conditioned to the result of the equivalence Wellek's test: If the null hypothesis of relevant ratio of variances is rejected, then Student's t test is applied; otherwise Welch's test is applied.*

4.3. Pretesting vs non-pretesting strategies

Figure 4 shows that, for all sample sizes under consideration, performing Student's t test directly without prior verification of the homoscedasticity assumption greatly inflates or deflates the TIEP as the heteroscedasticity increases. The inflation/deflation of TIEP depends on sample size and especially on balancing/unbalancing; so, for unbalanced

cases, this TIEP's affectation is much greater. For unbalanced cases, the TIEP is below the significance level when the largest sample corresponds to greater variance; whereas, when the smallest sample corresponds to greater variance, the estimated TIEP is above the significance level.

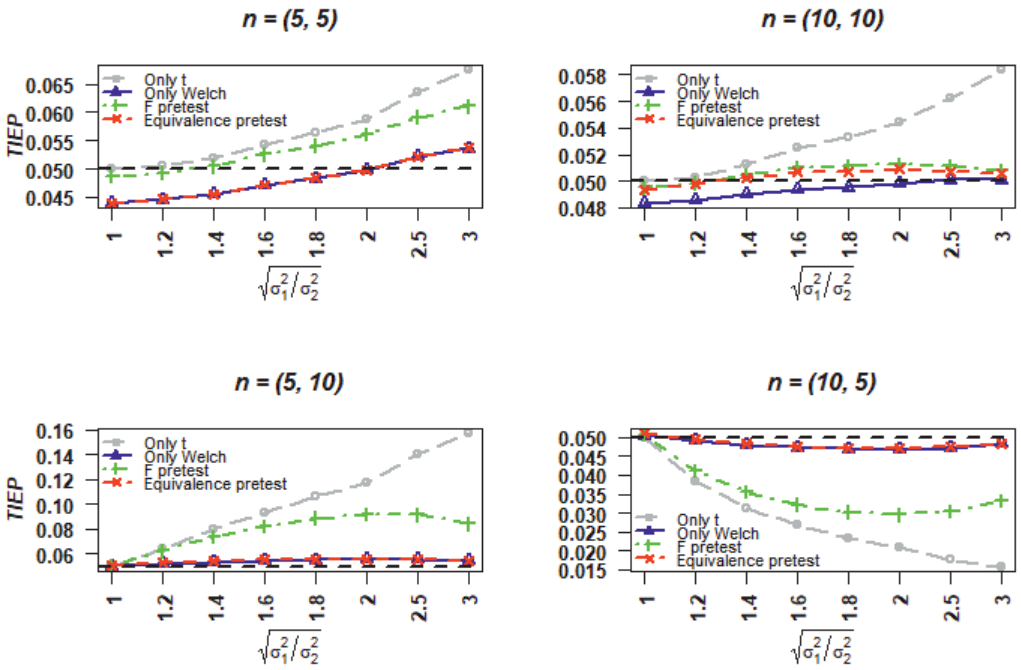

Figure 4: *TIEP affectation using* $\alpha = 0.05$ *and* $\delta = 0.01$ *jointly comparing the pretesting and non-pretesting strategies. Note again that the TIEP scales differ.*

When a traditional pretest such as the F test is used to verify homoscedasticity before deciding on Student's t test or Welch's test should be used, the TIEP also inflates/deflates in the same way as the previous case, although with less intensity, and it becomes less concerning for increasing balanced sample sizes.

Very similar behaviour occurs when Welch's test is used directly, without pretesting, and when pretesting is based on the Wellek's equivalence test. Both strategies are quite stable, with true TIEP values close to the nominal significance level. For low and unbalanced sample sizes, the equivalence test has low power; the null hypothesis stating a disturbing level of heteroscedasticity is rarely rejected; due to there being not enough evidence to prove a non-disturbing level of heteroscedasticity, the cautionary approach of using Welch's test is taken (which seems a more reliable strategy than assuming homoscedasticity based on being unable to prove heteroscedasticity by means of Student's test).

The strategy of using exclusively Welch's test and the strategy based on Wellek's pretest only slightly differ for large and preferably balanced sample sizes. Then, as

the heteroscedasticity irrelevance test reaches enough power, more often there is some evidence to think on a non-disturbing heteroscedasticity and to use Student's t test instead of Welch's test. Although both strategies are tenable (while the other two should be advised against), it is difficult to say which strategy is best. Equivalence pretesting translates into a less conservative strategy, but both have true TIEP values that are very close to the nominal significance level, i.e., always within the $\alpha \pm \delta$ limits.

5. Illustrative examples

To illustrate these methods, we will use two datasets available at the website of the University of Sheffield. The data files and the *R* scripts with the functions implementing the methods described above are available on request to the authors. In all these examples, tests were performed at a nominal significance level of 0.05 and irrelevance in the TIEP distortion was fixed in a 20% level and, therefore, the Cochran's criterion was applied.

The first dataset is available at: https://www.sheffield.ac.uk/mash/statistics2/data.

These data are part of a study trying to relate margarine (or more precisely, its active ingredient, stanol ester) as part of a low fat, low cholesterol diet, with the reduction on cholesterol levels. Here we compare this response on 18 subjects, which are assigned in a balanced way to two margarine types, A and B.

From the algorithm described in Section 3, and provided that both sample sizes (A and B) are $n_1 = n_2 = 9$, all values of the true ratio of variances ranging from 0.1076 to 9.2928 are acceptable to keep the true TIEP of Student's t test inside the limits 0.05 ± 0.01. We feed these equivalence limits (0.1076 and 9.2928) into Wellek's algorithm to determine the critical region of the equivalence test (Section 2). The resulting critical region is $0.3420 < Q < 2.9236$. Provided that the sample variances are 1.7090 and 0.6820, and thus the resulting test statistic is $Q = 2.5059$, then the null hypothesis stating the existence of a relevant heteroscedasticity is rejected. Therefore, applying Student's t test may be considered acceptable. Its resulting p-value is 0.2771 and, therefore, it is impossible to reject the null hypothesis of means equality. Under the "always Welch - no pretests" strategy the resulting p-value is very similar, 0.2801, obviously with the same conclusion.

The second dataset considered here is available at: https://www.sheffield.ac.uk/polopoly_fs/1.570199!/file/stcp-Rdataset-Diet.csv.

These data correspond to a study relating loss in body weight with three diets. We will consider only two groups: diets 1 and 3, and the loss in body weight after 6 weeks of treatment will be used as the observed variable. The respective sample sizes are higher than in the previous example and they are unbalanced: $n_1 = 24$ and $n_3 = 27$. Given these sample sizes and the previously fixed tolerance in the TIEP, 0.05 ± 0.01, the resulting equivalence or heteroscedasticity irrelevance limits are 0.0008 and 3.5068. These equivalence limits conduct to the critical region of the equivalence test defined

by $0.0015 < Q < 1.7900$. For the sample variances 5.0183 and 5.7387, for diet 1 and diet 3, respectively, and then for the ratio $Q = 0.8744$, the hypothesis of a relevant heteroscedasticity is rejected. Consequently, applying Student's *t* test may be considered acceptable. It provides a p-value of 0.0066, which conducts to the rejection of the null hypothesis of equality of means in favor of the two-sided alternative of difference. Again, Welch's test would come to the same conclusion, with a 0.0065 p-value.

Additional examples are available in the *R* scripts mentioned at the beginning of this section.

6. Conclusions and discussion

This paper reinforces the arguments against traditional pretests, such as the *F* test (or Levene's, Bartlett's, Cochran's, etc.) for testing the homoscedasticity assumption prior to Student's *t* test for comparison of means. It seems to support the categorical statement of Rasch et al. (2011) that directly advises against using Student's *t* test and instead promotes making routine use of Welch's test. Our results only qualify this conclusion slightly. Since there is only a small difference between directly applying Welch's test without any previous homoscedasticity verification and pretesting by means of an equivalence/irrelevance dispersion test, and because also both strategies seem to be reliable, it is difficult to recommend any one of them over the others. In any case, the decision should be made on the basis of balancing what is preferable: on the one hand, we have an always small difference in type I error control, which is slightly less conservative in equivalence pretesting; and, on the other, we have the opposite situation when applying only Welch's test without pretesting – which in any case is a simpler procedure.

When choosing between an equivalence pretesting approach or a more robust test against the failure to fulfil validity conditions, all doubts will disappear in situations lacking this second option. For example, it is our opinion that generalizing to more than 2 groups in Welch's test (Welch, 1951) leads to poor control of the TIEP. This could spark interest in continuing this study by expanding it to more general situations. An obvious first step would be to study the suitability of Wellek's test for heteroscedasticity irrelevance for more than two groups (Wellek, 2010, p. 227) as a pretest for the one-way ANOVA.

Acknowledgements

This research is partially supported by Grant MTM2015-64465-C2-1-R (MINECO/FEDER) from the Ministerio de Economía y Competitividad (Spain) and by grant 2014 SGR 464, Generalitat de Catalunya.

The authors are also very grateful to all three referees of this paper, for their very valuable and constructive comments.

A. Appendix: A variance reduction technique when the simulation output is a proportion

In the simulations described in this paper, the parameter to be estimated was a probability. Vegas and Ocaña (1992) and Ocaña and Vegas (1995) developed a simulation variance reduction technique based on the "control variates" method, specifically devoted to this situation. To implement control variates, the simulation output of each simulation replicate, say Y, (here it is an "indicator" variable: 1 if in the end the null hypothesis of equality of means has been rejected, 0 otherwise) should be paired with a correlated "control variate", say C, with known expectation, $E(C)$. In the present study, C was the outcome of Student's t test under the same simulated data but adapted to come from a perfect homoscedasticity scenario, with known $E(C) = \alpha$. In fact, the generation process was the inverse. First, a scenario of perfect homoscedasticity was simulated to obtain C; then, these (homoscedastic) simulated values were subsequently transformed to represent each desired degree of heteroscedasticity in order to obtain Y.

Assume that, after performing m simulation replicates, the simulation output (absolute frequencies) and the associated probabilities (here with $p_{.1} = \alpha$) can be summarized as shown in the following table:

	$C=0$	$C=1$	
$Y=0$	m_{00}	m_{01}	$m_{0.}$
$Y=1$	m_{10}	m_{11}	$m_{1.}$
	$m_{.0}$	$m_{.1}$	m

	$C=0$	$C=1$	
$Y=0$	p_{00}	p_{01}	$p_{0.}$
$Y=1$	p_{10}	p_{11}	$p_{1.}$
	$p_{.0}$	$p_{.1}$	1

Ocaña and Vegas (1995) showed that

$$\tilde{p}_{1.} = p_{.0}\tilde{p}_{10} + p_{.1}\tilde{p}_{11} = p_{.0}\frac{m_{10}}{m_{00}+m_{10}} + p_{.1}\frac{m_{11}}{m_{01}+m_{11}}$$

is an unbiased estimator of $p_{1.}$, which is more efficient than the raw relative frequency, $m_{1.}/m$. Its variance can be estimated by means of:

$$\tilde{\sigma}^2_{\tilde{p}_{1.}} = \frac{\tilde{p}_{00}\tilde{p}_{10}}{np_{.0}+p_{.0}-2} + \frac{\tilde{p}_{01}\tilde{p}_{11}}{np_{.1}+p_{.1}-2}.$$

References

Altman, D.G. and J.M. Bland (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, 311, 485.

Box, G.E. (1979). Robustness in the strategy of scientific model building. *Robustness in Statistics*, 1, 201–236.

Cochran, W.G. (1942). The χ^2 correction for continuity. *Iowa State College Journal of Science*, 16, 421–436.

Hsu, P. (1938). Contribution to the theory of "student's" t-test as applied to the problem of two samples. *Statistical Research Memoirs*.

Ocaña, J. and E. Vegas (1995). Variance reduction for bernoulli response variables in simulation. *Computational Statistics and Data Analysis*, 19, 631–640.

Overall, J.E., R.S. Atlas and J.M. Gibson (1995). Tests that are robust against variance heterogeneity in $k \times 2$ designs with unequal cell frequencies. *Psychological Reports*, 76, 1011–1017.

Rasch, D. and V. Guiard (2004). The robustness of parametric statistical methods. *Psychology Science*, 46, 175–208.

Rasch, D., K.D. Kubinger and K. Moder (2011). The two-sample t test: pre-testing its assumptions does not pay off. *Statistical Papers*, 52, 219–231.

Scheffé, H. (1970). Practical solutions of the behrens-fisher problem. *Journal of the American Statistical Association*, 65, 1501–1508.

Vegas, E. and J. Ocaña (1992). Variance reduction for bernouilli response variables. In *Computational Statistics*, pp. 103–107. Springer.

Welch, B. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, 38, 330–336.

Welch, B.L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika*, 34, 28–35.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. USA: CRC Press.

Zimmerman, D.W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57, 173–81.

SORT 42 (1) January-June 2018, 73-98 DOI: 10.2436/20.8080.02.70

Empirical analysis of daily cash flow time-series and its implications for forecasting

Francisco Salas-Molina[1], Juan A. Rodríguez-Aguilar[2], Joan Serrà[3]
Montserrat Guillen[4] and Francisco J. Martin[5]

Abstract

Usual assumptions on the statistical properties of daily net cash flows include normality, absence of correlation and stationarity. We provide a comprehensive study based on a real-world cash flow data set showing that: (i) the usual assumption of normality, absence of correlation and stationarity hardly appear; (ii) non-linearity is often relevant for forecasting; and (iii) typical data transformations have little impact on linearity and normality. This evidence may lead to consider a more data-driven approach such as time-series forecasting in an attempt to provide cash managers with expert systems in cash management.

MSC: 62J02, 62J05, 62P20.

Keywords: Statistics, forecasting, cash flow, non-linearity, time-series.

1. Introduction

Cash management is concerned with the efficient use of a company's cash and short-term investments such as marketable securities. The focus is placed on maintaining the amount of available cash as low as possible, while still keeping the company operating efficiently. In addition, companies may place idle cash in short-term investments (Ross, Westerfield and Jordan, 2002). Then, the cash management problem can be viewed as a trade-off between holding and transaction costs. If a company tries to keep balances too low, holding cost will be reduced, but undesirable situations of shortage will force to sell available marketable securities, hence increasing transaction costs. In contrast, if the balance is too high, low trading costs will be produced due to unexpected cash flow, but the company will carry high holding costs because no interest is earned on cash. There–

[1] Universitat de València, Av. Tarongers, s/n, 46022 València, Spain. francisco.salas-molina@uv.es

[2] IIIA-CSIC, Campus UAB, 08913 Cerdanyola, Spain. jar@iiia.csic.es

[3] Telefonica Research, Pl. Ernest Lluch, 5, 08019 Barcelona, Spain. joan.serra@telefonica.com

[4] Universitat de Barcelona, Diagonal, 690, 08034 Barcelona, Spain. mguillen@ub.edu

[5] BigML, Inc, 2851 NW 9th Suite, Conifer Plaza, Corvallis, OR 97330, US. martin@bigml.com
Received: March 2018
Accepted: June 2018

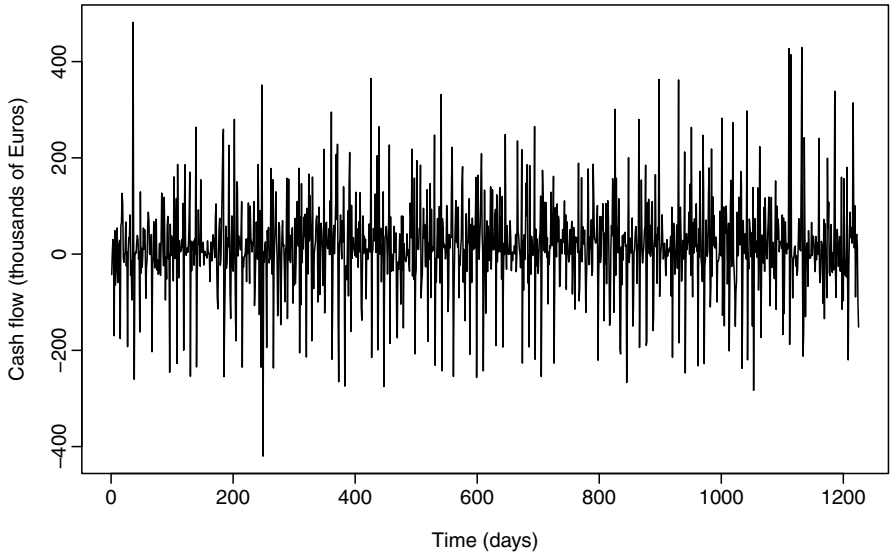

Figure 1: *Example of a cash flow time-series.*

fore, there is a target cash balance which each company must optimize according to the particular characteristics of its cash flows. An example of a raw cash flow time-series is shown in Figure 1, where observations do not apparently follow any seasonal pattern and whose evolution over time seems to be quite stable in terms of mean and variance, similarly to a white noise signal.

Testing the validity of time-series assumptions is an ongoing issue in finance (Marathe and Ryan, 2005; Ewing and Thompson, 2007; Cavaliere and Xu, 2014; Horváth, Kokoszka and Rice, 2014; Arratia, Cabana and Cabana, 2016; Torabi, Montazeri and Grané, 2016). Since Baumol (1952), a number of cash management models have been proposed to control cash balances. These models are based either on the specific statistical properties of cash balances or on cash flow forecasts. A comprehensive review of models, from the first proposals to the most recent contributions, can be found in Gregory (1976), Srinivasan and Kim (1986), and da Costa Moraes, Nagano and Sobreiro (2015). Most of them are based on assuming a given probability distribution for cash flows such as: (i) a random walk in the form of independent Bernouilli trials as in Miller and Orr (1966); (ii) a Wiener process as in Constantinides and Richard (1978), Premachandra (2004), and Baccarin (2009); (iii) a double exponential distribution as in Penttinen (1991). From these and other works, we observe that common assumptions on the statistical properties of cash flow time-series include:

- Normality: cash flows follow a Gaussian distribution with observations symmetrically centered around the mean, and with finite variance.
- Absence of correlation: the occurrence of past cash flows does not affect the probability of occurrence of the next ones.

- Stationarity: the probability distribution of cash flows does not change over time and, consequently, its statistical properties such as the mean and variance remain stable.
- Linearity: cash flows are proportional either to another (external) explanatory variable or to a combination of (external) explanatory variables.

Surprisingly, little and/or contradictory empirical evidence on these assumptions has been provided besides individual cases through time. Early on, negative normality tests were reported in Homonoff and Mullins (1975) for the times series samples of a manufacturing company. Contrastingly, later on, Emery (1981) reported normally distributed cash flow, after data transformation, for two out of three companies, and a small serial dependence for all of them. Pindado and Vico (1996) provided negative normality and independence results on 36 companies, but considering daily cash flow for only a single month. Previous works also reported day-of-week and day-of-month effects on cash flows, in line with the works of Stone and Wood (1977), Miller and Stone (1985), and Stone and Miller (1987). Recently, Gormley and Meade (2007) described the time-series from a multinational company with a non-normal distribution and serial dependence.

We consider that the evidence derived from these works is inconclusive due to: (i) the disagreement between the conclusions of some of the works; (ii) the limited number of companies analysed; and (iii) the short time range of the observations. Moreover, none of the previous works considered the presence of non-linear patterns for forecasting purposes. In this work, we provide an analysis of the statistical properties of 54 real cash flow data sets from small and medium companies in Spain as a representative sample of the most common type of companies in Europe. Indeed, small and medium companies contribute to 99.8% of all enterprises, 57.4% of value added, and 66.8% of employment across the EU28 (Muller et al., 2015). To the best of our knowledge, this is the most comprehensive empirical study on daily cash flow so far. We base this statement on both the length and number of data sets, which amounts to 58005 observations in total, with a minimum, average and maximum time range of 170, 737, 1508 working days, respectively. In addition, we consider a wider range of statistical properties. A further contribution of the present work is to make all the aforementioned data publicly available online[1]. Finally, from a forecasting perspective, we also aim to identify the family of forecasters that best accommodate to cash flow time-series data sets. To this end, we propose a new and simple cross-validated test for non-linearity that provides further knowledge to cash managers in their search for better forecasting models.

Our results show the unlikely occurrence of normality, absence of correlation and stationarity in the data sets under study. These results are consistent with the cited reports of Homonoff and Mullins (1975), based on only one time-series, and Pindado and Vico (1996), based on a very short time range, raising doubts about the claim of indepen-

1. http://www.iiia.csic.es/ jar/54datasets3.csv

dence. We also report that normality could not be achieved through removing outliers, contrary to what was reported by Emery (1981), based on only three time-series. Our analysis also confirms the influence of seasonality as suggested in Miller and Stone (1985) and Stone and Miller (1987). Thus, we consider that our results provide stronger evidence against normality, uncorrelatedness and stationarity than previous works. Note that we do not claim that these results can be extrapolated to all kind of companies. On the contrary, we provide further evidence against standard assumptions in cash management. This evidence may lead to consider a more data-driven approach such as time-series forecasting in order to provide cash managers with expert systems in cash management (Nedović and Devedžić, 2002).

In an attempt to achieve Gaussian and stationary time-seres, practitioners typically use the Box-Cox transformation (Box and Cox, 1964), and time-series differencing (Makridakis, Wheelwright and Hyndman, 2008). Furthermore, some kind of outlier treatment is also a recommended practice. Then, we also study the impact of outlier treatment by replacing them with linear interpolations between two consecutive observations. However, in our study, we find little benefit when these methods are applied to our data sets. As a result, we point out the underlying question about data transformation in relation to the properties of a time-series. Is it always possible to achieve a Gaussian and linear time-series through data transformations? We rely here both on common statistical tests and on our novel non-linearity test to answer this question and we find that: (i) outlier treatment and Box-Cox transformation are not always enough to achieve normality; (ii) outlier treatment produces mixed results in terms of noise reduction and information loss; (iii) outlier treatment and Box-Cox transformations do not produce linearity. These results suggest that non-linear models conform a justifiable alternative for cash flow time-series forecasting, beyond the current conjectures of the literature.

The remaining of the paper is organized as follows. In Section 2, we provide a statistical summary of the contributed 54 real cash flow data sets including normality, correlation and stationarity. In Section 3, we propose a new cross-validated test for non-linearity based on the comparison of a linear model and a non-linear model. Later, we present in Section 4 detailed results on the impact of data transformations on linearity. Finally, we provide some concluding remarks in Section 5.

2. Data summary

The data set contains daily cash flows from 54 different companies from the manufacturing and the service sector in Spain with annual revenue up to €10 million each. No company from the primary sectors is included in the sample. We select only small and medium companies since it is the most common size of companies in both Spain and Europe (Muller et al., 2015). This data set covers a date range of about eight years and is available online. An instance in the data set contains the following fields or columns:

Table 1: *Data sets statistical summary. Mean, standard deviation, minimum, maximum in thousands of €.*

Id	Length	Null %	Mean	Std	Kurtosis	Skewness	Min	Max
1	856	35.7	0.01	3.38	594.81	22.37	−9.07	90.27
2	684	29.8	0.26	5.80	58.98	3.69	−56.51	62.66
3	856	8.5	0.36	35.35	163.62	6.28	−303.20	671.04
4	1201	34.9	−0.12	14.32	78.14	−6.30	−223.38	72.76
5	849	19.4	0.00	1.67	56.10	−0.48	−18.26	16.42
6	799	20.7	0.01	6.63	33.21	−2.42	−68.97	56.27
7	772	38.5	0.07	5.36	86.75	6.74	−24.41	82.91
8	695	21.7	0.05	3.15	14.27	−2.57	−24.21	11.31
9	852	18.8	0.73	56.54	18.92	−0.78	−411.41	473.36
10	744	13.2	0.12	6.95	70.63	0.60	−81.13	78.72
11	639	62.6	−0.05	8.56	391.86	−17.65	−191.53	30.74
12	503	2.6	0.48	35.30	449.38	20.70	−47.27	771.38
13	697	24.7	0.52	24.24	18.81	2.06	−99.39	227.45
14	604	4.6	0.10	13.23	8.51	1.05	−63.23	92.71
15	605	4.1	0.68	11.67	4.43	0.33	−54.75	55.61
16	596	6.4	0.01	1.46	107.82	6.68	−8.48	22.61
17	1102	25.1	0.58	13.31	215.97	11.96	−118.01	250.13
18	552	3.1	0.16	2.16	70.23	5.10	−16.14	26.36
19	503	2.4	−0.31	2.58	6.43	0.50	−15.06	15.28
20	848	27.8	0.02	1.07	96.19	3.86	−12.07	16.04
21	829	18.7	−0.06	5.99	33.36	−1.62	−70.00	53.17
22	494	1.6	−0.46	27.28	22.64	−1.96	−244.29	138.87
23	604	9.1	1.63	20.85	79.99	5.41	−124.19	269.27
24	1097	8.4	0.96	20.36	95.45	6.48	−73.33	317.85
25	587	10.9	0.49	13.94	119.60	6.93	−116.01	201.13
26	751	11.6	−0.02	1.77	15.73	0.15	−10.73	15.56
27	332	8.1	0.29	1.64	10.60	2.14	−4.36	11.84
28	855	5.1	0.00	4.64	13.83	1.77	−18.10	39.01
29	609	13.6	0.04	6.07	108.66	−6.35	−90.04	55.89
30	554	8.1	0.03	1.47	68.26	5.47	−4.81	19.82
31	372	29.6	0.37	8.05	31.46	−2.41	−80.44	34.95
32	1103	24.8	0.28	4.03	11.07	0.54	−25.76	24.50
33	854	31.0	−0.19	6.81	115.63	−1.74	−94.33	95.59
34	1508	11.5	−0.06	10.13	19.89	−2.32	−96.82	49.65
35	501	7.4	0.20	5.40	11.41	−0.58	−31.42	29.19
36	359	11.4	0.42	1.85	12.24	2.44	−7.87	11.84
37	361	3.0	−0.69	17.82	139.06	−1.38	−228.88	218.42
38	170	9.4	−1.20	7.10	43.34	−5.73	−61.93	19.66
39	1104	29.0	0.02	0.95	7.95	−0.07	−5.67	6.57
40	198	0.0	0.78	12.38	0.58	1.02	−25.63	36.91
41	341	17.6	−0.25	8.34	15.80	1.22	−44.29	64.34
42	566	11.0	0.01	1.82	308.62	−15.80	−37.02	7.48
43	750	3.2	0.34	13.10	7.66	−0.04	−65.84	73.40
44	287	4.2	0.52	11.46	81.19	−0.05	−118.74	120.34
45	1465	49.8	0.04	9.12	43.51	−2.89	−107.20	75.47
46	565	44.8	0.54	5.58	75.41	2.91	−51.16	73.83
47	503	4.4	1.98	46.81	46.03	1.37	−338.39	478.26
48	605	13.1	0.21	22.71	34.31	−1.68	−207.04	203.09
49	993	50.5	−0.08	1.36	27.18	−2.18	−10.78	12.73
50	605	45.0	−0.01	27.37	43.79	−2.01	−262.52	221.96
51	1225	0.2	15.09	96.96	2.77	0.12	−419.88	481.66
52	1225	0.4	8.94	49.39	36.23	2.81	−325.46	700.66
53	1223	39.7	0.47	9.13	203.12	−10.25	−196.88	38.48
54	1225	52.3	0.46	77.91	151.93	4.28	−1021.36	1532.10

- Date: standardized YYYY-MM-DD dates from 2009-01-01 to 2016-28-08.
- Company: company identifier from 1 to 54.
- NetCF: daily net cash flow in thousands of €.
- DayMonth: categorical variable with the day of the month from 1 to 31.
- DayWeek: categorical variable with the day of the week from 1 (Monday) to 7 (Sunday).

Table 1 shows the statistical summary of daily net cash flow on non-holidays, grouped by company. Small and medium companies are likely to experiment daily null cash flows, meaning that no monetary movement is observed at a particular working day even under regular activity. As a result, the occurrence of null cash flows is an important characteristic of small and medium companies due to the size of companies. Indeed, almost 30% of the companies in our data set present more than 25% of null cash flow observations even at working days. This fact implies that a null cash flow prediction will be right at least 25% of the times for this group of data sets. Therefore, two good baseline forecasting models for comparative purposes would be an *always-predict-null* or an *always-predict-mean* forecaster (Makridakis et al., 2008).

In addition, the average net cash flow shows that a high percentage of companies present either positive or negative mean with the exception of companies 5 and 28. High positive kurtosis indicates a peaked data distribution in comparison to the normal distribution that has zero kurtosis. The skewness is a measure of the symmetry of the data distribution. Negative skewness indicates that the left tail is longer, and positive skewness indicates that the right tail is longer.

2.1. Normality

First, we study if our cash flows follow a Gaussian distribution. In fact, the observed kurtosis and skewness can be used as a first normality test of the data distribution for each company. Table 1 shows that no company presents zero kurtosis and skewness. Only company 40, with kurtosis 0.58 and skewness 1.02, could be considered close to normality. The proportion of null cash flows is also a strong evidence against normality. Since this situation is likely to be common for SMEs and, due to the high proportion of this type of companies in Europe, we believe that cash managers should test normality before applying cash management models based on this assumption.

Two additional tests can be used to either verify or reject the hypothesis of normality: the Shapiro-Wilk test for normality (Royston, 1982) and the Lilliefors (Kolmogorov-Smirnov) test for normality (Lilliefors, 1967). The results from these two tests applied to the original time-series (summarized in Table 2) allow us to reject the hypothesis of normally distributed cash flows for all the companies in our data set (no exception). However, the presence of correlation and possible changes in the mean of data sets may limit the reliability of these tests. We overcome this problem by performing an additional normality test. More precisely, we check the normality of the residuals of fitting

an ARMA model to each of the time-series as suggested by Ducharme and Lafaye de Micheaux (2004). To obtain ARMA models, we follow the automatic fitting procedure described in Hyndman and Khandakar (2008). Finally, we test the normality of the residuals by means of Neyman (1937) smooth tests as recently proposed by Ducharme and Lafaye de Micheaux (2004) and Duchesne, Lafaye de Micheaux and Tagne Tatsinkou (2016). The results from Table 2 before any data transformation suggest the rejection of the normality hypothesis.

As pointed out elsewhere (Emery, 1981; Pindado and Vico, 1996), a possible explanation for non-normality could be the presence of abnormally high values or heavy tails. Thus, we repeated the Shapiro-Wilk, the Lilliefors (Kolmogorov-Simirnov), and the Neyman tests for normality, but now using a trimmed version of the net cash flow time-series by deleting observations greater or lower than three times the sample standard deviation. No difference in the results of the tests was observed, confirming the non-normality hypothesis beyond the conjectures of Emery (1981) and Pindado and Vico (1996).

Non-normal residuals may be problematic in the estimation process when using linear models. Data transformations such as the Box and Cox (1964) transformation to normality represent a possible solution. Forecasts are then calculated on the transformed data, but we must reverse the transformation to obtain forecasts on the original data, resulting in two additional steps. However, these transformations are not always the solution to the non-normality problem. Using both the original observations and the trimmed version of our data sets, we proceeded to transform the data using a Box-Cox transformation of the type:

$$y^{(\lambda)} = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \text{if} \quad \lambda_1 \neq 0, \\ \log(y+\lambda_2) & \text{if} \quad \lambda_1 = 0, \end{cases} \tag{1}$$

where y is the original time-series, and λ_1 and λ_2 are parameters. In these experiments, we first set λ_2 to minus two times the minimum value of the time-series to avoid problems with negative and zero observations. Box and Cox (1964) provided the profile likelihood function for λ_1 and suggested to use this function as a way to tune this parameter. Then, we follow the recommendations in Venables and Ripley (2013) to compute the profile likelihood function for λ_1, and we later select the value that maximizes the log-likelihood function when applying a linear regression model of the time-series based on day-of-month and day-of-week dummy variables. After a Box-Cox transformation on the trimmed time-series, we repeated the Shapiro-Wilk, the Lilliefors (Kolmogorov-Smirnov), and the Neyman smooth tests for normality obtaining again negative results as shown in Table 2. A possible explanation of these results is that the correlational structure of a transformed time-series closely depends on the original. A special case of this feature for a logarithmic transformation can be found in Moriña, Puig and Valero (2015). As a result, we must conclude that, even after Box-Cox transformation, the normality hypothesis does not hold.

Table 2: *Normality tests. SW: Shapiro-Wilk; LKS: Lilliefors (Kolmogorov-Smirnov).*

	Before data transformation						After trimming and Box-Cox transformation					
Id	SW	p-value	LKS	p-value	Neyman	p-value	SW	p-value	LKS	p-value	Neyman	p-value
1	0.20	< 0.05	0.31	< 0.05	13165.72	< 0.05	0.76	< 0.05	0.23	< 0.05	2125.70	< 0.05
2	0.36	< 0.05	0.35	< 0.05	6539.22	< 0.05	0.55	< 0.05	0.35	< 0.05	3884.23	< 0.05
3	0.42	< 0.05	0.26	< 0.05	6452.81	< 0.05	0.76	< 0.05	0.26	< 0.05	839.58	< 0.05
4	0.45	< 0.05	0.32	< 0.05	10249.73	< 0.05	0.69	< 0.05	0.28	< 0.05	4463.36	< 0.05
5	0.54	< 0.05	0.24	< 0.05	5103.79	< 0.05	0.83	< 0.05	0.23	< 0.05	949.19	< 0.05
6	0.59	< 0.05	0.27	< 0.05	3694.65	< 0.05	0.78	< 0.05	0.26	< 0.05	768.83	< 0.05
7	0.48	< 0.05	0.32	< 0.05	4768.18	< 0.05	0.74	< 0.05	0.29	< 0.05	1223.53	< 0.05
8	0.64	< 0.05	0.33	< 0.05	1902.93	< 0.05	0.75	< 0.05	0.28	< 0.05	1724.23	< 0.05
9	0.54	< 0.05	0.39	< 0.05	2558.89	< 0.05	0.63	< 0.05	0.38	< 0.05	1563.74	< 0.05
10	0.41	< 0.05	0.24	< 0.05	6437.00	< 0.05	0.78	< 0.05	0.25	< 0.05	1715.64	< 0.05
11	0.17	< 0.05	0.39	< 0.05	10234.39	< 0.05	0.49	< 0.05	0.35	< 0.05	3164.46	< 0.05
12	0.11	< 0.05	0.37	< 0.05	9828.61	< 0.05	0.76	< 0.05	0.28	< 0.05	1171.58	< 0.05
13	0.59	< 0.05	0.33	< 0.05	2229.85	< 0.05	0.65	< 0.05	0.31	< 0.05	1111.39	< 0.05
14	0.84	< 0.05	0.17	< 0.05	353.45	< 0.05	0.89	< 0.05	0.16	< 0.05	127.69	< 0.05
15	0.89	< 0.05	0.16	< 0.05	216.99	< 0.05	0.91	< 0.05	0.16	< 0.05	148.40	< 0.05
16	0.53	< 0.05	0.22	< 0.05	3145.88	< 0.05	0.88	< 0.05	0.20	< 0.05	175.85	< 0.05
17	0.24	< 0.05	0.34	< 0.05	14724.84	< 0.05	0.69	< 0.05	0.31	< 0.05	2295.82	< 0.05
18	0.55	< 0.05	0.19	< 0.05	2660.82	< 0.05	0.89	< 0.05	0.19	< 0.05	242.65	< 0.05
19	0.92	< 0.05	0.10	< 0.05	456.82	< 0.05	0.97	< 0.05	0.10	< 0.05	54.93	< 0.05
20	0.53	< 0.05	0.24	< 0.05	4530.45	< 0.05	0.84	< 0.05	0.24	< 0.05	578.91	< 0.05
21	0.69	< 0.05	0.23	< 0.05	2015.07	< 0.05	0.84	< 0.05	0.22	< 0.05	588.95	< 0.05
22	0.67	< 0.05	0.22	< 0.05	1263.64	< 0.05	0.82	< 0.05	0.21	< 0.05	285.96	< 0.05
23	0.48	< 0.05	0.27	< 0.05	3365.92	< 0.05	0.77	< 0.05	0.27	< 0.05	500.78	< 0.05
24	0.49	< 0.05	0.29	< 0.05	5986.28	< 0.05	0.71	< 0.05	0.31	< 0.05	1345.56	< 0.05
25	0.36	< 0.05	0.28	< 0.05	5084.02	< 0.05	0.74	< 0.05	0.28	< 0.05	805.87	< 0.05
26	0.75	< 0.05	0.21	< 0.05	1271.00	< 0.05	0.84	< 0.05	0.21	< 0.05	476.22	< 0.05
27	0.81	< 0.05	0.17	< 0.05	3832.19	< 0.05	0.92	< 0.05	0.14	< 0.05	1034.84	< 0.05
28	0.77	< 0.05	0.21	< 0.05	1622.42	< 0.05	0.85	< 0.05	0.19	< 0.05	486.48	< 0.05
29	0.31	< 0.05	0.39	< 0.05	6782.20	< 0.05	0.55	< 0.05	0.37	< 0.05	3712.34	< 0.05
30	0.62	< 0.05	0.20	< 0.05	1988.27	< 0.05	0.86	< 0.05	0.21	< 0.05	196.55	< 0.05
31	0.69	< 0.05	0.18	< 0.05	1707.62	< 0.05	0.86	< 0.05	0.19	< 0.05	580.67	< 0.05
32	0.73	< 0.05	0.22	< 0.05	1838.10	< 0.05	0.80	< 0.05	0.22	< 0.05	741.53	< 0.05
33	0.24	< 0.05	0.37	< 0.05	10442.42	< 0.05	0.44	< 0.05	0.38	< 0.05	6581.93	< 0.05
34	0.68	< 0.05	0.27	< 0.05	2529.16	< 0.05	0.77	< 0.05	0.23	< 0.05	1048.52	< 0.05
35	0.74	< 0.05	0.24	< 0.05	1027.85	< 0.05	0.83	< 0.05	0.23	< 0.05	356.87	< 0.05
36	0.67	< 0.05	0.26	< 0.05	4331.31	< 0.05	0.80	< 0.05	0.23	< 0.05	1949.37	< 0.05
37	0.10	< 0.05	0.46	< 0.05	6768.47	< 0.05	0.18	< 0.05	0.46	< 0.05	5398.45	< 0.05
38	0.41	< 0.05	0.30	< 0.05	1695.03	< 0.05	0.68	< 0.05	0.21	< 0.05	371.92	< 0.05
39	0.82	< 0.05	0.24	< 0.05	497.26	< 0.05	0.87	< 0.05	0.23	< 0.05	191.56	< 0.05
40	0.89	< 0.05	0.21	< 0.05	50.00	< 0.05	0.95	< 0.05	0.15	< 0.05	2155.90	< 0.05
41	0.66	< 0.05	0.28	< 0.05	1117.07	< 0.05	0.73	< 0.05	0.27	< 0.05	507.15	< 0.05
42	0.21	< 0.05	0.36	< 0.05	8189.26	< 0.05	0.68	< 0.05	0.24	< 0.05	725.85	< 0.05
43	0.84	< 0.05	0.16	< 0.05	686.05	< 0.05	0.90	< 0.05	0.17	< 0.05	281.96	< 0.05
44	0.37	< 0.05	0.30	< 0.05	2611.31	< 0.05	0.71	< 0.05	0.30	< 0.05	427.38	< 0.05
45	0.42	< 0.05	0.36	< 0.05	11122.67	< 0.05	0.57	< 0.05	0.35	< 0.05	6084.71	< 0.05
46	0.34	< 0.05	0.37	< 0.05	5194.50	< 0.05	0.51	< 0.05	0.37	< 0.05	2620.56	< 0.05
47	0.40	< 0.05	0.30	< 0.05	3979.90	< 0.05	0.64	< 0.05	0.30	< 0.05	1869.51	< 0.05
48	0.52	< 0.05	0.34	< 0.05	2851.85	< 0.05	0.66	< 0.05	0.33	< 0.05	1247.03	< 0.05
49	0.38	< 0.05	0.46	< 0.05	6891.41	< 0.05	0.51	< 0.05	0.45	< 0.05	2739.27	< 0.05
50	0.30	< 0.05	0.39	< 0.05	6871.79	< 0.05	0.40	< 0.05	0.38	< 0.05	4855.75	< 0.05
51	0.93	< 0.05	0.11	< 0.05	320.22	< 0.05	0.94	< 0.05	0.12	< 0.05	293.46	< 0.05
52	0.80	< 0.05	0.15	< 0.05	1538.44	< 0.05	0.93	< 0.05	0.13	< 0.05	293.37	< 0.05
53	0.35	< 0.05	0.33	< 0.05	9770.54	< 0.05	0.62	< 0.05	0.29	< 0.05	3792.57	< 0.05
54	0.30	< 0.05	0.37	< 0.05	11389.52	< 0.05	0.45	< 0.05	0.37	< 0.05	7568.90	< 0.05

2.2. Correlation and seasonality

In what follows, we test the correlation of cash flows and we also explore if seasonality is present. Autoregressive Integrated Moving Average (ARIMA) models by Box and Jenkins (1976), have been extensively used for time-series analysis and forecasting. When dealing with time-series, the autocorrelation coefficient, r_k, describes the relationship between observations that are lagged k time periods (Makridakis et al., 2008). We say that a time-series is not autocorrelated when the r_k values for different lags are close to zero. An example of an independent time-series is the so-called white-noise model where each observation is made by adding a random component to a certain level.

An intuitive plot to assess correlation is the Poincaré map (Kantz and Schreiber, 2004), which is a scatter plot of the original time-series and a k-periods lagged time-series as in Figure 2, which shows a lag of 1 day for time-series 1 and 2 from Table 1. As a reference, we also include the Poincaré map for a white-noise and for a sinusoidal time-series. A cloud of points suggests lack of correlation, as for time-series 1 and white-noise, and the presence of any form suggests a more complex relationship, as for time-series 2 and the sinusoidal. For comparative purposes, we present in Figure 3 the classical plots showing autocorrelation and partial autocorrelation functions for different lags within the range 1-20 with dashed horizontal lines representing 95% confidence intervals. From the analysis of Figure 3, we note correlation for time-seres 1 and 2 at lags 1 and 15, respectively.

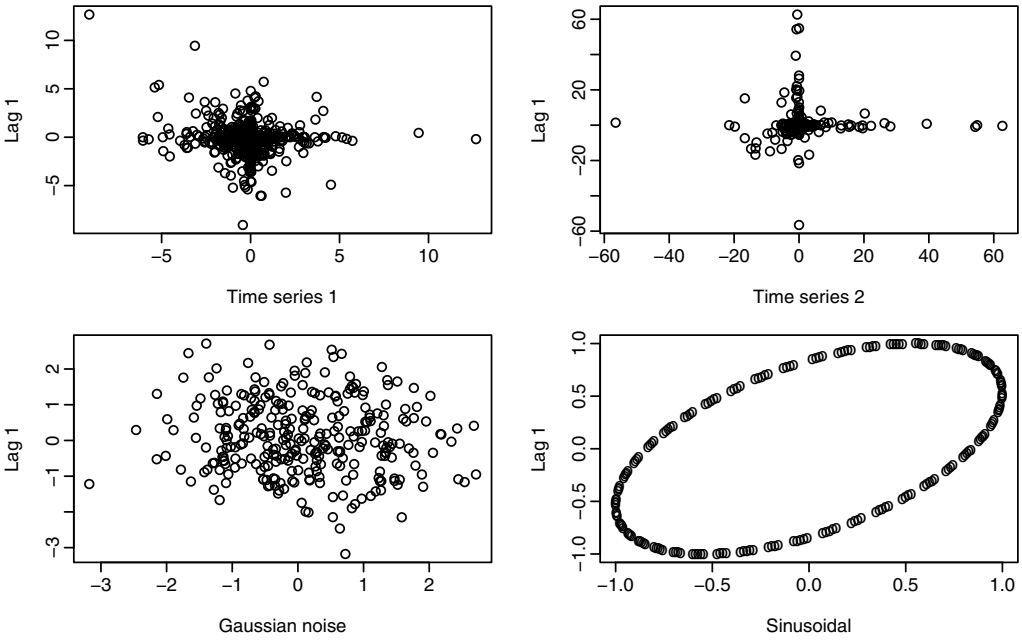

Figure 2: *Poincaré map with lag 1 for time-series 1 and 2.*

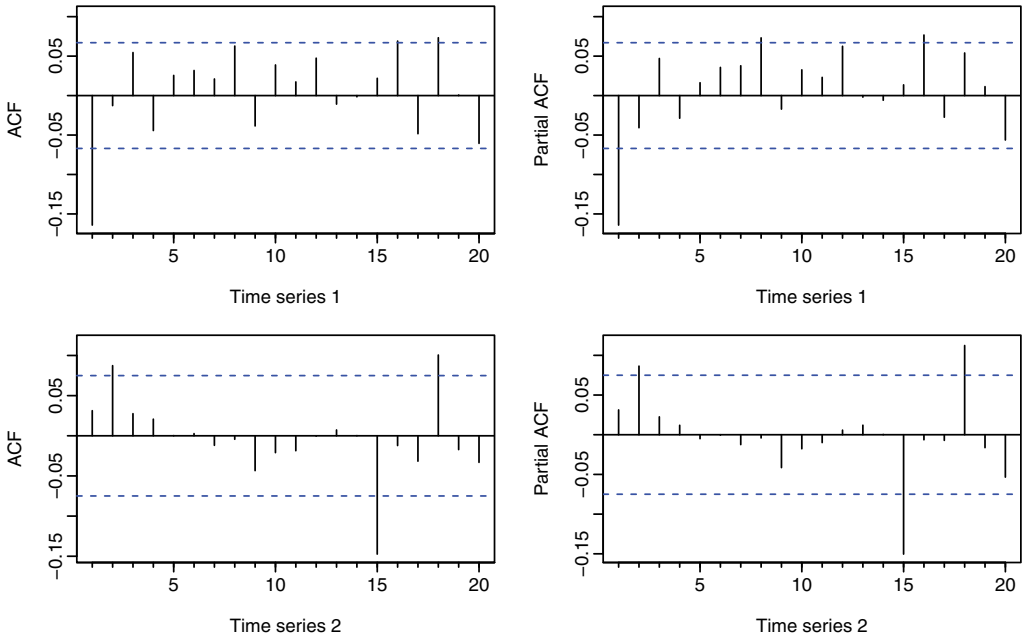

Figure 3: *Autocorrelation plots for time-series 1 and 2.*

A more general approach is to consider a set of the first r_k values as a whole as in the Ljung and Box (1978) test, which we applied to the original time-series and produced mixed results. More precisely, we found that the null hypothesis of independence could not be rejected in 24 out of 54 companies as summarized in Table 3. These results imply that some kind of serial correlation is likely to be present in the case of companies presenting a certain degree of autocorrelation in the sample. A plausible type of serial correlation is seasonality, that is, the existence of a pattern that repeats itself over fixed time intervals in the data (Makridakis et al., 2008). It can be identified by significant autocorrelation coefficients. Seasonal trend decomposition methods Cleveland et al., 1990), seasonal ARIMA models (Box and Jenkins, 1976; Franses and Van Dijk, 2005) or linear (and non-linear) regression models based on seasonal variables are available options to deal with seasonality. In cash flow forecasting, the distribution approach by Miller and Stone (1985) also deserves to be mentioned.

As mentioned in the introduction, previous works by Emery (1981), Miller and Stone (1985), Stone and Miller (1987), and Pindado and Vico (1996), reported the influence of day-of-month and day-of-week effects on cash flow patterns. Here, we test the presence of seasonality by fitting a regression model on raw daily cash flows using day-of-month and day-of-week dummy variables. To avoid co-linearity issues in regression, we use thirty day-of-month dummy variables from the 2nd to the 31st day of the month and four day-of-week variables from Monday to Thursday up to a total of 34 regression variables. At each time step t, predictor x_{ti} is set to one if the corresponding day-of-month is i, zero

otherwise, and x_{tj} is set to one if the corresponding day-of-week is j, with j ranging from 1 for Monday to 4 for Thursday. Mathematically, the linear regression model used to test seasonality is expressed as follows:

$$y_t = \sum_{i=2}^{31} \beta_i x_{ti} + \sum_{j=1}^{4} \beta_j x_{tj} + \epsilon. \tag{2}$$

Table 3 reports, on the one hand, the Ljung-Box correlation test applied to raw data and, on the other hand, the F-statistic, the p-value and the coefficient of determination R^2, derived from the regression model. One may expect that the rejection of the correlation null hypothesis results in better regressions. Our results, however, show a different behavior. Non-linear patterns, non-periodical temporal correlations, and the effect of outliers become possible explanations as we will see below.

2.3. Stationarity

In this section, we analyse if cash flows from our data set can be labelled as stationary. More precisely, we focus on weak stationarity that considers the change over time of the first (mean) and second moment (variance) of a random process. We can visually assess stationarity by inspecting a time-series plot as the one shown in Figure 1. Virtually, every process we find in nature is non-stationary, since its parameters depend on time (Kantz and Schreiber, 2004). However, a minimum requirement is that basic statistical properties of a distribution, such as mean and variance, remain constant over time, when measured through appropriately long time windows. It is important to highlight that seasonality is a particular case of non-stationarity, at least, within each periodic fluctuation when we focus on short-term changes in parameters. In what follows, we pay attention to long-term changes (periods longer than a month) as a way to assess stationarity.

Following the recommendations in Kantz and Schreiber (2004), we perform a stationarity test based on the fluctuations of a sample mean and variance. More precisely, we compute the sample mean and variance of each original time-series by months and obtain the standard errors for both. If the observed fluctuations of the running mean and variance are within these errors, then we consider the time-series stationary. The results from this test shows that none of the time-series is stationary. These results are consistent with the fact that most of the p-values of the regression models used for checking seasonality are below 0.05 as summarized in Table 3.

One way of removing non-stationarity is time-series differencing, which is defined as the change between two consecutive observations. Similarly, seasonal differencing is the change between corresponding observations from two consecutive seasonal periods. Since the presence of seasonality is likely (see Table 3), we next explore three alternative

Table 3: *Correlation and seasonality test results.*

Id	Ljung-Box Test	Statistic	p-value	F-statistic	p-value	R^2
1	Non-rejected	11.05	1.00	1.99	< 0.05	0.08
2	Rejected	65.99	< 0.05	1.05	0.39	0.05
3	Non-rejected	34.47	0.72	1.87	< 0.05	0.07
4	Rejected	120.15	< 0.05	1.51	< 0.05	0.04
5	Rejected	120.91	< 0.05	1.85	< 0.05	0.07
6	Non-rejected	46.96	0.21	1.12	0.29	0.05
7	Rejected	166.97	< 0.05	5.47	< 0.05	0.20
8	Rejected	67.15	< 0.05	0.79	0.80	0.04
9	Rejected	97.32	< 0.05	5.30	< 0.05	0.18
10	Rejected	145.00	< 0.05	2.04	< 0.05	0.09
11	Non-rejected	10.57	1.00	0.97	0.51	0.05
12	Non-rejected	3.25	1.00	0.98	0.51	0.07
13	Rejected	139.26	< 0.05	5.21	< 0.05	0.21
14	Rejected	74.58	< 0.05	7.13	< 0.05	0.30
15	Rejected	87.67	< 0.05	1.92	< 0.05	0.10
16	Non-rejected	38.12	0.56	4.31	< 0.05	0.21
17	Non-rejected	14.49	1.00	4.91	< 0.05	0.14
18	Rejected	57.25	< 0.05	2.99	< 0.05	0.16
19	Rejected	75.16	< 0.05	2.58	< 0.05	0.16
20	Non-rejected	43.37	0.33	2.71	< 0.05	0.10
21	Non-rejected	46.65	0.22	1.37	0.08	0.06
22	Non-rejected	33.35	0.76	1.49	< 0.05	0.10
23	Rejected	68.36	< 0.05	5.60	< 0.05	0.25
24	Non-rejected	41.30	0.41	15.41	< 0.05	0.33
25	Non-rejected	33.35	0.76	4.23	< 0.05	0.21
26	Rejected	95.79	< 0.05	1.22	0.18	0.05
27	Non-rejected	44.66	0.28	1.24	0.18	0.12
28	Rejected	112.21	< 0.05	5.64	< 0.05	0.19
29	Non-rejected	42.55	0.36	1.37	0.08	0.08
30	Rejected	107.46	< 0.05	6.18	< 0.05	0.29
31	Non-rejected	47.51	0.19	1.25	0.16	0.11
32	Rejected	105.26	< 0.05	4.81	< 0.05	0.13
33	Rejected	201.50	< 0.05	1.57	< 0.05	0.06
34	Rejected	130.53	< 0.05	11.61	< 0.05	0.21
35	Rejected	66.04	< 0.05	0.99	0.49	0.07
36	Non-rejected	44.66	0.28	1.82	< 0.05	0.16
37	Rejected	96.75	< 0.05	1.58	< 0.05	0.14
38	Non-rejected	45.37	0.26	1.06	0.39	0.21
39	Rejected	192.30	< 0.05	6.11	< 0.05	0.16
40	Rejected	78.81	< 0.05	0.86	0.68	0.15
41	Non-rejected	39.05	0.51	1.72	< 0.05	0.16
42	Non-rejected	22.85	0.99	3.90	< 0.05	0.20
43	Rejected	80.56	< 0.05	2.96	< 0.05	0.12
44	Non-rejected	19.56	1.00	1.89	< 0.05	0.20
45	Rejected	82.69	< 0.05	1.26	0.15	0.03
46	Non-rejected	32.23	0.80	1.32	0.11	0.08
47	Non-rejected	35.67	0.67	0.90	0.63	0.06
48	Non-rejected	42.53	0.36	1.71	< 0.05	0.09
49	Rejected	105.02	< 0.05	26.15	< 0.05	0.48
50	Rejected	135.48	< 0.05	1.24	0.17	0.07
51	Rejected	131.27	< 0.05	16.66	< 0.05	0.32
52	Rejected	66.68	< 0.05	5.01	< 0.05	0.13
53	Non-rejected	18.62	1.00	1.59	< 0.05	0.04
54	Rejected	129.11	< 0.05	0.88	0.67	0.02

seasons (or periods) to apply differencing: 1) one day, equivalent to no seasonality; 2) five days, to account for day-of-week seasonality; and 3) twenty days, to account for day-of-month seasonality. Finally, differencing can be applied only once to data, twice or a number n of times defining the order of differencing. In Table 4, we summarize stationarity results for our data set in terms of the number of time-series that are labelled as stationary in mean and variance according to the test described above. Only a small fraction of time-series can be considered stationary in mean (but not in variance) after first and second-order differencing. From this analysis, we conclude that our cash flow time-series are non-stationary, even after differencing.

Table 4: *Percentage of time-series labelled as stationary in mean and variance.*

Differencing	Zero-order		First-Order		Second-Order	
Seasonality	Mean	Var	Mean	Var	Mean	Var
1	0	0	18.5	0	18.5	0
5	0	0	3.7	0	5.6	0
20	0	0	0	0	0	0

2.4. Discussion

Our results show that the widely extended hypothesis of cash flow normality is not present in our data sets. The presence of high abnormal values does not explain this behavior since non-normality persisted after removing these abnormal values. Non-linearity could be a possible explanation as we will see below. We also reported mixed results on autocorrelation and the influence of day-of-month and day-of-week effects on cash flow along the lines of the literature. We additionally report that common solutions to non-normality and non-stationarity such as data transformation and differencing produced little benefit when applied to our time-series. Since seasonality and serial correlation are also present in our data set, we further explore the usefulness of alternative forecasting models. More precisely, we next study linearity and data transformation as an additional part of our empirical analysis for cash flow forecasting.

3. A simple cross-validated test for non-linearity

Most forecasting models are linear for computational convenience. However, non-linear patterns are likely to be present in finance and business time-series. A time-series linear model is defined as a variable y_t that depends on the additive contribution of a number of explanatory variables in vector $\boldsymbol{x_t}$ for any time t as follows:

$$y_t = \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{x_t} + e_t \tag{3}$$

where $\boldsymbol{\beta}^{\mathsf{T}}$ is a transposed vector of coefficients, and e_t is the error or the residual component. An alternative and more general model can also be considered:

$$y_t = g(\boldsymbol{x_t}) + \epsilon_t \tag{4}$$

where $g(\boldsymbol{x_t})$ is any function that aims to describe the underlying time-series. By considering non-linear relationships between the set of predictors and the cash flow dependent variable, more complex patterns such as interactions between the day-of-week and the day-of-month may be captured.

Different tests of linearity can be found in Ramsey (1969), Keenan (1985), Lee, White and Granger (1993), and Castle and Hendry (2010). Basically, all of them follow a common approach: first, they choose a function $g(\boldsymbol{x_t})$ in equation (4) including linear and non-linear terms and, second, they test for the significance of the non-linear terms. However, these approaches are not suitable for forecasting purposes owing to the following reasons: (i) the assumption of a specific form $g(\boldsymbol{x_t})$ for the regression equation such as quadratic, cubic or exponential forms; (ii) cross-validation is neglected.

If we relax the assumption of linearity, different non-linear models such as random forests (Breiman, 2001), neural networks (Hornik, Stinchcombe and White, 1989; Zhang, Patuwo and Hu, 1998), or radial basis functions (Broomhead and Lowe, 1988), could also be considered. However, the consideration of non-linear functions may lead to overfitting to the original time-series. To prevent this problem, we propose the use of time-series cross-validation. Cross-validation is a method to assess the predictive performance of a forecasting model that circumvents the problem of overfitting the data by testing the accuracy of the model on subset of data not used in the estimation (Hyndman and Athanasopoulos, 2013). As a result, we here propose a simple cross-validated test for non-linearity based on the following steps:

1. Estimate two alternative forecasting models, one linear and another one non-linear.
2. Cross-validate the predictive accuracy of both models with respect to a baseline.
3. Label as trivial[2] if both models are significantly worse than the baseline.
4. Label as non-linear if the error of the non-linear model is significantly lower than that of the linear model. Otherwise, label as linear as described in Figure 4.

Since we do not assume any distribution for the forecasting results, we use the two-sided Wilcoxon rank-based for statistically significant differences in performance between models. More precisely, we test the null hypothesis that the distribution of the difference is symmetric about zero with a 95% confidence interval (Wilcoxon, Katti and Wilcox, 1970). Approximate p-values are computed based on the asymptotic distribution of the two-sided Wilcoxon test statistic and used to label data sets as detailed in Algorithm 1.

2. Trivial is here used with the meaning of very little value with respect to a basic standard.

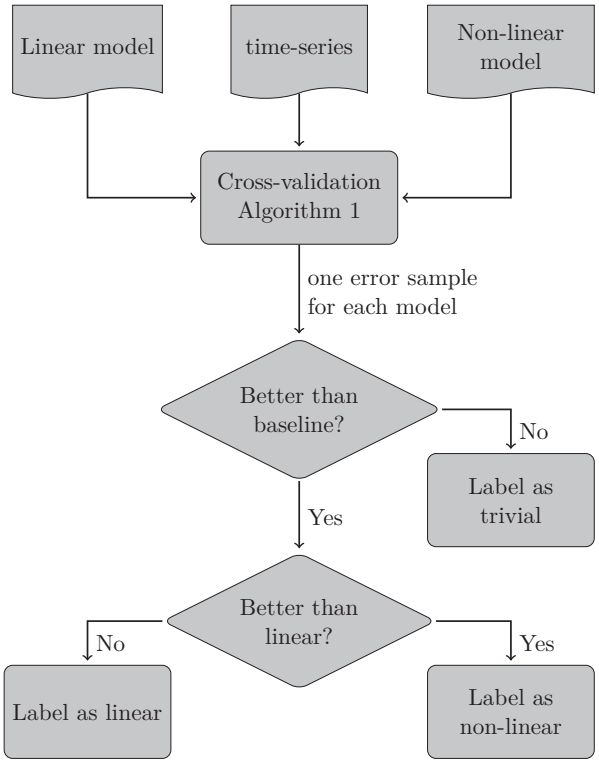

Figure 4: *Simplified flow chart for our cross-validated test for non-linearity.*

Algorithm 1 Algorithm for a simple cross-validation test for non-linearity

Input: Cash flow data set of T instances, minimum number k of instances to estimate a model, baseline m_0, linear model m_1, non-linear model m_2, prediction horizon h, level of significance α.
Output: Average prediction error, statistic for the difference in mean errors, confidence interval.
for $i = 1,2,\ldots,T-k-h+1$ **do**
    Select the instances from time $k+i$ to $k+h+i-1$, for the test set;
    Estimate m_0 with instances at times $1,2,\ldots,k+i-1$;
    Estimate m_1 with instances at times $1,2,\ldots,k+i-1$;
    Estimate m_2 with instances at times $1,2,\ldots,k+i-1$;
    Compute test errors ε_0, ε_1, ε_2 from time $k+i$ to $k+h+i-1$;
Compute average h-step errors $\varepsilon_0(h)$, $\varepsilon_1(h)$, $\varepsilon_2(h)$;
Test for α significant differences between $\varepsilon_0(h)$, $\varepsilon_1(h)$, $\varepsilon_2(h)$;
if $\varepsilon_0(h) < \varepsilon_1(h)$ **and** $\varepsilon_0(h) < \varepsilon_2(h)$ **then**
    Label as trivial;
else if $\varepsilon_2(h) < \varepsilon_1(h)$ **then**
    Label as non-linear;
else
    Label as linear.

A common practice to assess the usefulness of forecasts derived from any model is to compare its accuracy to that of a baseline forecasting model. The use of a baseline model allows us to label our data sets as trivial if neither the linear model nor the non-linear model are able to improve the accuracy of the baseline. We here report accuracy results with respect to a mean forecaster, meaning that forecasts are always the average of all past observations. We also tried with an additional baseline forecaster using the last observed value as a forecast (persistence model) with much worse results in comparison to the mean forecaster.

We consider the minimum length k to estimate a model as the 80% of the oldest instances forming the training set. The remaining 20% of the instances form the test set for cross-validation. Initially, both the linear and the non-linear model are estimated using the first 80% of the instances. Then, forecasts for a prediction horizon up to 20 days are computed using the estimated models and squared errors are recorded. Then, forecasting accuracy is evaluated on a rolling basis, since both the last observation of the training set and the first observation of the test set roll forward in time. As a result, forecasting errors are recorded for each remaining observation in the test set resulting into two paired error samples, one for the linear model and one for non-linear model.

A critical point when using our cross-validated test for non-linearity is the selection of both the linear and the non-linear forecast model. In essence, our test is a comparative tool based on forecasting accuracy as a proxy for non-linearity. Given a set of explanatory variables, a linear label result from our test implies that the non-linear model is not able to capture non-linearity. However, chances are that alternative non-linear models might perform differently. In this sense, if the time-series is not a white-noise process, then the search for a more informative set of features is meant to play a key role. As a result, multiple runs of our test are necessary to discard/assess non-linearity by using alternative linear and non-linear models

For illustrative purposes, we here restrict ourselves to a linear regression model and a non-linear random forest model, both using day-of-month and day-of-week variables as predictors. Salas-Molina et al. (2017) report that these two models perform significantly better than autoregressive models when producing forecasts for usual prediction horizons up to one hundred days. Here, we are interested in comparing forecasting models that perform well for a wide range of planning horizons from the information available at some point in time. Thus, we expect that forecasting models based on seasonal variables capture patterns for common prediction horizons better than time-series models based on previous observations due to lack of relevant information as reported by Salas-Molina et al. (2017).

In the case of the linear regression model, each instance contains 34 dummy predictor variables, 30 for day-of-month and 4 for day-of week, and a cash flow observation. This linear regression model is the same that we used in Section 2.2 to check seasonality. In the case of random forests, each instance contains two categorical variables, one for day-of-month and one for day-of-week. Random forests are ensembles of slightly

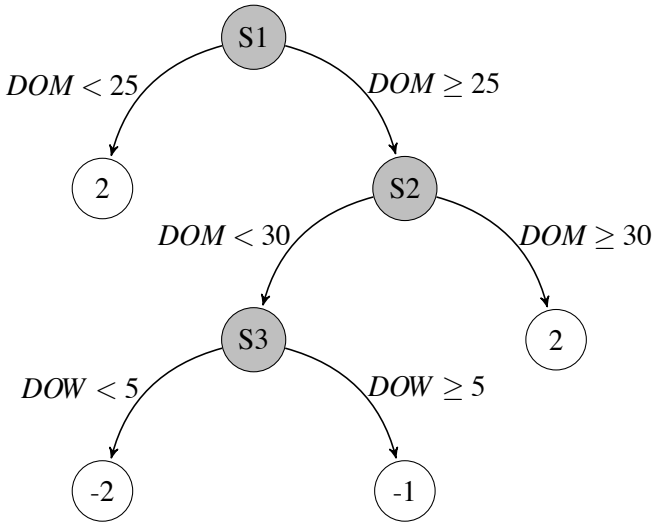

Figure 5: *A basic decision tree. DOM = Day-of-month; DOW = Day-of-week.*

different decision trees (Ho, 1998; Breiman, 2001). An ensemble methodology is able to construct a predictive model by integrating multiple trees in what is called a decision forest (Dietterich, 2000). Decision trees split the input space in subsets based on the value of features such as the day-of-month and day-of-month. In the example in Figure 5, for days comprised between the 25th (node S1) and the 29th of each month (node S2) occurring on Friday (node S3), the predicted cash flow is -1.

Recent examples of time-series forecasting using random forests can be found in Booth, Gerding and Mcgroarty (2014), Zagorecki (2015) and Salas-Molina et al. (2017). Summarizing, random forests are used to forecast variables based on an ensemble of different trees. Unlike linear regression, random forests allow to capture (if any) more complex relationships between predictor variables allowing us to identify possible non-linearities in the underlying cash flow process represented by our sample data sets.

In Table 5, we summarize results only for data sets that can be labelled as trivial because neither the linear model nor the non-linear model were able to significantly beat the baseline forecaster. As described in Algorithm 1, we label time-series as linear when lower normalized squared errors are obtained using the regression model. Similarly, we label time-series as non-linear when lower errors are obtained using the random forest model. In addition, we test the significance of the difference in performance between regression and random forest models. When p-values from these tests are below 0.05, we consider that sample errors for the linear and the non-linear model are significantly different.

From those time-series in which the absence of correlation could no be rejected (see Ljung-Box test at Table 3), 20 out of 24 were labelled as trivial. On the other hand, only 6 of them were labelled as non-linear according to our cross-validated definition. As mentioned above, these results depend on the selected forecasting models. Instead

Table 5: *Results of the test for non-linearity. Reg NSE = Regression normalized squared error; RF NSE = Random forest normalized squared error.*

Id	Reg NSE	RF NSE	Statistic	p-value	Triviality	Linearity
1	0.99	1.00	26	< 0.05	Non-Trivial	Linear
3	0.99	1.01	8	< 0.05	Non-Trivial	Linear
4	1.00	1.01	0	< 0.05	Non-Trivial	Linear
7	0.81	0.83	0	< 0.05	Non-Trivial	Linear
9	0.90	0.93	3	< 0.05	Non-Trivial	Linear
13	0.86	0.88	13	< 0.05	Non-Trivial	Linear
14	0.76	0.77	45	< 0.05	Non-Trivial	Linear
16	0.85	0.86	64	0.13	Non-Trivial	Linear
18	0.86	0.88	63	0.12	Non-Trivial	Linear
19	0.96	0.94	182	< 0.05	Non-Trivial	Non-linear
20	0.99	0.98	209	< 0.05	Non-Trivial	Non-linear
23	0.78	0.79	78	0.33	Non-Trivial	Linear
24	0.73	0.79	0	< 0.05	Non-Trivial	Linear
25	0.77	0.81	21	< 0.05	Non-Trivial	Linear
28	0.84	0.90	0	< 0.05	Non-Trivial	Linear
29	0.99	0.99	30	< 0.05	Non-Trivial	Linear
30	0.73	0.80	5	< 0.05	Non-Trivial	Linear
33	0.94	0.93	166	< 0.05	Non-Trivial	Non-linear
34	0.97	0.95	172	< 0.05	Non-Trivial	Non-linear
39	0.96	0.96	36	< 0.05	Non-Trivial	Linear
42	0.88	0.87	149	0.11	Non-Trivial	Linear
43	0.99	0.96	210	< 0.05	Non-Trivial	Non-linear
48	1.01	0.99	191	< 0.05	Non-Trivial	Non-linear
49	0.63	0.65	7	< 0.05	Non-Trivial	Linear
51	0.77	0.80	0	< 0.05	Non-Trivial	Linear
52	0.94	0.94	116	0.70	Non-Trivial	Linear

of claiming that random forests are able to better capture non-linear patterns than alternative models, we encourage practitioners to consider additional combinations of both linear and non-linear models.

One may assume either linearity or non-linearity from the results of our non-linearity test, but it is important to analyse the robustness of these results to both the presence of outliers and the impact of other data transformations.

4. The impact of data transformations

In this section, we aim to analyse the impact of outlier treatments on noise reduction, as intended, and on information loss, as an undesirable effect. We also study the influence of Box-Cox data transformations on the results of our cross-validated non-linearity test. Detection and treatment of outliers is an ongoing issue in data mining (Rousseeuw and Leroy, 1987; Hodge and Austin, 2004). An outlier is an observation that appears to

significantly deviate from other members of the sample in which it occurs (Grubbs, 1969). Outliers arise due to changes in systems, measurement errors or simply due to deviations from average activity. It is also important to note that an outlier may also be the most interesting part of the data.

On the one hand, from the set of cash flow time-series labelled as trivial, some of them may be labelled as non-trivial after removing outliers as a way of noise reduction. On the other hand, from those data sets labelled as non-trivial, some of them may be labelled as trivial due to the information loss produced by the treatment. We here measure the effect of removing outliers on the prediction error using time-series cross validation for different thresholds of outlier replacement. For each data set, we progressively identify as outliers cash flow observations greater than 5, 4, and 3 times the standard deviation in a training set with the 80% oldest observations. We replace outliers with a linear interpolation of the previous and the posterior observation and we proceed as detailed in Algorithm 1 to cross-validate triviality and linearity. The results from this analysis are summarized in Table 6, where global performance after treatments is assessed by averaging noise (error) reduction. Note that some time-series in Table 6 are not present in Table 5 because outlier treatment and Box-Cox transformation produced a improvement in accuracy.

By following this procedure, we identify data sets 5, 10, 17, 32, 44 and 54 (6 out of 28), initially labelled as trivial that, after outlier treatment, can be labelled as non-trivial due to noise reduction. Similarly, data sets 4 and 48 that were initially labelled as non-trivial can be labelled as trivial after outlier treatment due to information loss. If we measure noise reduction by the error reduction and information loss by the error increase, then we can assess the impact of outlier treatment. Following this approach, we obtained mixed results for non-trivial data sets after outlier treatment: an average noise reduction of 22%, and an average information loss of 14%. It is important to recall that unexpected observations are often the most interesting part of the data to predict, e.g., when the goal is to forecast unusual but genuine cash flows.

Non-linearity and outliers are closely linked. Indeed, Castle and Hendry (2012) hypothesized that non-linear functions can align with outliers, causing functions to be considered relevant spuriously, which can be detrimental for generalizing and forecasting. If this hypothesis is correct, the relative forecasting ability of a linear model in comparison to a non-linear model would increase as the presence of outliers in a training set is reduced. From the set of time-series finally labelled as non-trivial, data sets 33, 34 and 54, initially labelled as non-linear changed their labels to linear. Surprisingly, data sets 17, 18, 23, 25, 39, 44 and 49 (7 out of 30), could be labelled as non-linear after outlier treatment. Except for data sets 17 and 44, in all cases there was information loss, i.e., error increase, suggesting that non-linear models can deal better with information loss.

We also considered a Box-Cox transformation to analyse if this kind of data transformation may influence the results from our cross-validated non-linearity test. From the set of non-trivial data sets we compare linearity labels, first, after outlier treatment,

Table 6: *Results of the test for non-linearity after outlier treatment and Box-Cox transformation. Changes in labels are marked with* *.

Id	Triviality	After outliers		After outliers and Box-Cox	
		Linearity	Noise reduction	Linearity	Noise reduction
1	Non-Trivial	Linear	0.00	Non-linear*	−0.01
3	Non-Trivial	Linear	0.02	Non-linear*	0.00
5	Non-Trivial	Non-linear	0.40	Non-linear	0.41
7	Non-Trivial	Linear	−0.10	Linear	−0.13
9	Non-Trivial	Linear	−0.04	Linear	−0.04
10	Non-Trivial	Non-linear	0.46	Non-linear	0.47
13	Non-Trivial	Linear	−0.18	Linear	−0.21
14	Non-Trivial	Linear	−0.05	Linear	−0.07
16	Non-Trivial	Linear	−0.18	Linear	−0.17
17	Non-Trivial	Non-linear*	0.71	Non-linear	0.71
18	Non-Trivial	Non-linear*	−0.20	Non-linear	−0.20
19	Non-Trivial	Non-linear	−0.03	Non-linear	−0.04
20	Non-Trivial	Non-linear	−0.02	Non-linear	−0.02
23	Non-Trivial	Non-linear*	−0.22	Non-linear	−0.22
24	Non-Trivial	Linear	−0.20	Linear	−0.06
25	Non-Trivial	Non-linear*	−0.26	Non-linear	−0.25
28	Non-Trivial	Linear	−0.05	Linear	−0.04
29	Non-Trivial	Linear	0.07	Non-linear*	0.00
30	Non-Trivial	Linear	−0.06	Linear	−0.04
32	Non-Trivial	Non-linear	0.18	Non-linear	0.21
33	Non-Trivial	Linear*	−0.12	Linear	−0.11
34	Non-Trivial	Linear*	0.12	Linear	0.09
39	Non-Trivial	Non-linear*	−0.02	Linear*	−0.01
42	Non-Trivial	Linear	−0.23	Linear	−0.14
43	Non-Trivial	Non-linear	0.04	Non-linear	0.03
44	Non-Trivial	Non-linear*	0.48	Non-linear	0.82
49	Non-Trivial	Non-linear*	−0.56	Non-linear	−0.61
51	Non-Trivial	Linear	−0.03	Linear	−0.03
52	Non-Trivial	Linear	0.01	Linear	0.03
54	Non-Trivial	Linear*	0.17	Linear	0.17
Average performance			0.00		0.02

and second, after outlier treatment and Box-Cox transformation as described in equation (1). In addition, we compare information loss computed as the difference between the sum of errors of the linear and non-linear forecasting models before and after the outlier treatment. A positive value means noise reduction or error reduction while a negative value means information loss or error increase. Results from Table 6 show a similar performance after Box-Cox transformation since the change in labels occurs in data sets with similar linear and non-linear noise reduction.

Table 7 shows the impact of outlier treatment and data transformation on the classification of time-series derived from our cross-validated non-linearity summarized in

Table 7: *Number of time-series data sets and their labels after transformation. OT=Outlier treatment; DT=Data transformation.*

Label	Raw data	After OT	After OT and DT
Trivial	28	24	24
Non-trivial	26	30	30
-Linear	20	17	15
-Non-linear	6	13	15

Table 6. The high number of trivial data sets may be caused by the general inherent randomness of cash flows. In addition, an increase in the number of time-series classified as non-trivial after treatments suggests a positive impact. However, non-linear models seem to obtain a higher benefit from treatments. First, outlier treatment produced a small improvement in non-triviality but also an outstanding increase in non-linearity. Second, after both outlier treatment and Box-Cox data transformation, resulted in similar results but with better performance for non-linear models.

It is worth mentioning that global performance in terms of error reduction remained unchanged after outlier treatment and slightly improved after data transformation (see Table 6). Thus, we conclude that: (i) common data transformations had little impact on our time-series in terms of linearity and accuracy; and (ii) outlier treatment and Box-Cox transformation were unable to transform non-linear into linear cash flows.

5. Concluding remarks

Small and medium companies contribute to a high percentage of all enterprises, value added and employment in Europe. In this paper, we provide a complete empirical study of the statistical properties of daily cash flows based on 54 real-world time-series for small and medium companies. To the best of our knowledge, this work is the most comprehensive empirical study on daily cash flows so far in terms of the range of statistical properties considered, and also in terms of the number and the length of the data sets. Particularly, we focus on the implications of our analysis for forecasting due to its key role in cash management. An additional contribution of this work is to make all data publicly available online for further research.

5.1. Summary of findings

Our results show that the extended hypotheses of normal, stationary and uncorrelated cash flows are hardly present in our cash flow data set. Thus, we conclude that the standard assumptions of normality, stationarity and uncorrelatedness that have been extensively used in the cash management literature must be verified before the deployment of any cash management model based on them. We do not claim that these results can be generalized to all small and medium companies. Indeed, we hypothesize that companies

with a larger number of daily cash flows may be closer to satisfy these usual assumptions than small and medium companies. This hypothesis represents an interesting subject of future research and we here set the path to this research by providing the methods to verify such hypothesis. We also highlight that common solutions to non-normality and non-stationarity such as data transformation and differencing produce little benefit when applied to our data sets, with the risk of losing important information on extreme cash flows. Alternative and more complex data transformations are nevertheless an option to consider in further research to achieve Gaussian cash flows.

In an attempt to discover the attributes of actual-world cash flows, we also studied the presence of non-linearity. To this end, we proposed a new simple test for non-linearity with two main advantages in comparison to alternative approaches. First, our test does not assume any non-linear function. Second, it is based on time-series cross validation to increase robustness and to avoid overfitting. It is important to note that our cross-validated definition of non-linearity depends on the alternative models considered, one linear and another one non-linear.

Our cross-validated non-linearity test labelled as either trivial, linear or non-linear our cash flow data set after outlier treatment resulting in an important increase in the number of data sets labelled as non-linear. After both outlier treatment and Box-Cox transformation, linearity could not be achieved and non-linear models showed more robust. However, the overall impact of data transformations on forecasting performance was limited. The application of our test to provide further evidence on these topics when using alternative cash flow data sets represents a natural extension of our work.

5.2. Implications

Our results raise questions about two common assumptions in cash flow time-series since we found that: (i) the usual assumption of normality, absence of correlation and stationarity is hardly present; and (ii) common data transformations such as outlier treatment and Box-Cox transformation have little impact on normality and linearity. Contrary to the rather common assumption in the literature, these results imply that neither it is always possible to achieve a Gaussian, white-noise and linear time-series through data transformation nor it is always desirable due to information loss. In this paper, we are interested in models that produce forecasts for a wide range of planning horizons. Thus, autoregressive and linear models should be considered as an initial step towards more realistic ones which are better adapted to real cash flow situations. The results from our cross-validated test for non-linearity suggest that non-linear models represent a justifiable alternative for time-series forecasting. Moreover, since our test is both model and outlier dependent, a promising line of future work is the integration of outlier treatment in the test itself in an attempt to assess noise reduction or information loss.

We claim that a number of preliminary steps are necessary in cash flow forecasting before model selection: (i) statistical summary including normality, correlation and

stationarity; (ii) impact of data transformations such as outlier treatment and Box-Cox transformation; (iii) non-linearity test to determine the type of model which is expected to deliver a better performance. This process is not limited to daily cash flow, since it can also be applied to any other time-series data set when cross-validation is required.

Acknowledgements

Work partially funded by projects Collectiveware TIN2015-66863-C2-1-R (MINECO/ FEDER) and 2014 SGR 118.

References

Arratia, A., Cabana, A. and Cabana, E.M. (2016). A construction of continuous-time arma models by iterations of ornstein-uhlenbeck processes. *SORT-Statistics and Operations Research Transactions*, 1, 267–302.

Baccarin, S. (2009). Optimal impulse control for a multidimensional cash management system with generalized cost functions. *European Journal of Operational Research*, 196, 198–206.

Baumol, W.J. (1952). The transactions demand for cash: An inventory theoretic approach. *The Quarterly Journal of Economics*, 66, 545–556.

Booth, A., Gerding, E. and Mcgroarty, F. (2014). Automated trading with performance weighted random forests and seasonality. *Expert Systems with Applications*, 41, 3651–3661.

Box, G.E. and D.R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211–252.

Box, G.E. and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*, revised ed. Holden-Day.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

Broomhead, D. and Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2, 321–355.

Castle, J.L. and Hendry, D.F. (2010). A low-dimension portmanteau test for non-linearity. *Journal of Econometrics*, 158, 231–245.

Castle, J.L. and Hendry, D.F. (2012). Automatic selection for non-linear models. In *System Identification, Environmental Modelling, and Control System Design*, pp. 229–250. Springer.

Cavaliere, G. and Xu, F. (2014). Testing for unit roots in bounded time series. *Journal of Econometrics*, 178, 259–272.

Cleveland, R.B., Cleveland, W.S., McRae, J.E. and Terpenning, I. (1990). Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6, 3–73.

Constantinides, G.M. and Richard, S.F. (1978). Existence of optimal simple policies for discounted-cost inventory and cash management in continuous time. *Operations Research*, 26, 620–636.

da Costa Moraes, M.B., Nagano, M.S. and Sobreiro, V.A. (2015). Stochastic cash flow management models: A literature review since the 1980s. In *Decision Models in Engineering and Management*, pp. 11–28. Springer International Publishing.

Dietterich, T.G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, pp. 1–15. Springer.

Ducharme, G.R. and Lafaye de Micheaux, P. (2004). Goodness-of-fit tests of normality for the innovations in arma models. *Journal of Time Series Analysis*, 25, 373–395.

Duchesne, P., Lafaye de Micheaux, P. and Tagne Tatsinkou, J. (2016). Estimating the mean and its effects on neyman smooth tests of normality for arma models. *Canadian Journal of Statistics*, 44, 241–270.

Emery, G.W. (1981). Some empirical evidence on the properties of daily cash flow. *Financial Management*, 10, 21–28.

Ewing, B.T. and Thompson, M.A. (2007). Symmetric and asymmetric stationarity tests of productivity. *The Engineering Economist*, 52, 355–365.

Franses, P.H. and Van Dijk, D. (2005). The forecasting performance of various models for seasonality and nonlinearity for quarterly industrial production. *International Journal of Forecasting*, 21, 87–102.

Gormley, F.M. and Meade, N. (2007). The utility of cash flow forecasts in the management of corporate cash balances. *European Journal of Operational Research*, 182, 923–935.

Gregory, G. (1976). Cash flow models: a review. *Omega*, 4, 643–656.

Grubbs, F.E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11, 1–21.

Ho, T.K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 832–844.

Hodge, V.J. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22, 85–126.

Homonoff, R. and Mullins, D.W. (1975). *Cash Management: an Inventory Control Limit Approach*. Lexington Books.

Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2, 359–366.

Horváth, L., Kokoszka, P. and Rice, G. (2014). Testing stationarity of functional time series. *Journal of Econometrics*, 179, 66–82.

Hyndman, R. and Athanasopoulos, G. (2013). *Forecasting: Principles and Practice*. http://otexts.org/fpp/. Last accessed: 2017-05-01.

Hyndman, R.J. and Khandakar, Y. (2008). *Automatic Time Series for Forecasting: the Forecast Package for R*.

Kantz, H. and Schreiber, T. (2004). *Nonlinear Time Series Analysis*, Volume 7. Cambridge University Press.

Keenan, D.M. (1985). A tukey nonadditivity-type test for time series nonlinearity. *Biometrika*, 72, 39–44.

Lee, T.-H., White, H. and Granger, C.W. (1993). Testing for neglected non-linearity in time series models: A comparison of neural network methods and alternative tests. *Journal of Econometrics*, 56, 269–290.

Lilliefors, H.W. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62, 399–402.

Ljung, G.M. and Box, G.E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297–303.

Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (2008). *Forecasting Methods and Applications*. John Wiley & Sons.

Marathe, R.R. and Ryan, S.M. (2005). On the validity of the geometric brownian motion assumption. *The Engineering Economist*, 50, 159–192.

Miller, M.H. and Orr, D. (1966). A model of the demand for money by firms. *The Quarterly Journal of Economics*, 80, 413–435.

Miller, T.W. and Stone, B.K. (1985). Daily cash forecasting and seasonal resolution: Alternative models and techniques for using the distribution approach. *Journal of Financial and Quantitative Analysis*, 20, 335–351.

Moriña, D., Puig, P. and Valero, J. (2015). A characterization of the innovations of first order autoregressive models. *Metrika*, 78, 219–225.

Muller, P., Caliandro, C., Peycheva, V., Gagliardi, D., Marzocchi, C., Ramlogan, R. and Cox, D. (2015). *Annual Report on European SMEs*.

Nedović, L. and Devedžić, V. (2002). Expert systems in finance: a cross-section of the field. *Expert Systems with Applications*, 23, 49–66.

Neyman, J. (1937). "Smooth test" for goodness of fit. *Scandinavian Actuarial Journal*, 1937, 149–199.

Penttinen, M.J. (1991). Myopic and stationary solutions for stochastic cash balance problems. *European Journal of Operational Research*, 52, 155–166.

Pindado, J. and Vico, J. (1996). Evidencia empírica sobre los flujos de caja. Un nuevo enfoque en su tratamiento. *Revista Española de Financiación y Contabilidad*, 25, 497–517.

Premachandra, I. (2004). A diffusion approximation model for managing cash in firms: An alternative approach to the miller-orr model. *European Journal of Operational Research*, 157, 218–226.

Ramsey, J.B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31, 350–371.

Ross, S.A., Westerfield, R. and Jordan, B.D. (2002). *Fundamentals of Corporate Finance* (Sixth ed.). McGraw-Hill.

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*, Volume 589. John Wiley & Sons.

Royston, J. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31, 115–124.

Salas-Molina, F., Martin, F.J., Rodríguez-Aguilar, J.A., Serrà, J. and Arcos, J.L. (2017). Empowering cash managers to achieve cost savings by improving predictive accuracy. *International Journal of Forecasting*, 33, 403–415.

Srinivasan, V. and Kim, Y.H. (1986). Deterministic cash flow management: state of the art and research directions. *Omega*, 14, 145–166.

Stone, B.K. and Miller, T.W. (1987). Daily cash forecasting with multiplicative models of cash flow patterns. *Financial Management*, 16, 45–54.

Stone, B.K. and Wood, R.A. (1977). Daily cash forecasting: a simple method for implementing the distribution approach. *Financial Management*, 6, 40–50.

Torabi, H., Montazeri, N.H. and Grané, A. (2016). A test for normality based on the empirical distribution function. *SORT-Statistics and Operations Research Transactions*, 1, 55–88.

Venables, W.N. and Ripley, B.D. (2013). *Modern Applied Statistics with S-PLUS*. Springer Science & Business Media.

Wilcoxon, F., Katti, S. and Wilcox, R.A. (1970). Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected Tables in Mathematical Statistics*, 1, 171–259.

Zagorecki, A. (2015). Prediction of methane outbreaks in coal mines from mulivariate time series using random forest. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, pp. 494–500. Springer.

Zhang, G., Patuwo,B.E. and Hu, M.Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14, 35–62.