The Barcelona International Conference on Advances in Statistics

Abstracts
Barcelona, June 18-22, 2012

Edited by Vladimir Zaiats, Digital Technologies Group

The Barcelona International Conference on Advances in Statistics (BAS2012)

Abstracts of communications

Barcelona, June 18–22, 2012

The Barcelona International Conference on Advances in Statistics has been organised by: Escola Politècnica Superior, Universitat de Vic

In collaboration with:
Servei d'Estadística, Universitat Autònoma de Barcelona
CosmoCaixa Barcelona

With the support of Ministerio de Economía y Competitividad

Primera edició: juny de 2012

Edició: Vladimir Zaiats, Servei de Publicacions de la UVic i Eumogràfic
Disseny de la coberta: Eumogràfic
Impressió: Artyplan (versió paper)

Universitat de Vic, c/. Sagrada Família, 7, 08500 Vic (Barcelona) Spain
ISBN: 978-84-940081-5-3 (versió digital)
DL: B.23403-2012

Welcome to BAS2012

We are pleased to present the third edition of the BAS International Conference. The previous conferences go back to 2003 and 2008 under the name of the Barcelona Conference on Asymptotic Statistics.
This time we have aimed at expanding the scope of the conference and we have called it The Barcelona International Conference on Advances in Statistics, maintaining the same acronym BAS. The three main topics the conference would like to focus on are:

- theoretical advances in statistics;
- applications of statistics;
- statistical software and its use.

Of course, these topics are intrinsically related to each other and it is sometimes difficult to clearly delimit them. A recent breakthrough in biomedical and financial applications of statistics calls for creating new tools, often requiring refined numerical techniques and simulating for further development of probability fundamentals of statistics. Therefore the BAS conference is intended as a forum for discussion of a wide range of ideas bringing together researchers whose contribution to statistics is already world-wide recognized, as well as young scientists whose professional career in statistics is making first steps.
The BAS venue at CosmoCaixa Barcelona is an excellent place for this type of events. The infrastructure offered by CosmoCaixa and, which is important, the atmosphere created for everybody coming to this center is stimulating and encourages for a fruitful work.
The BAS conference would have been impossible without financial support from the Ministerio de Economía y Competitividad. We gratefully acknowledge this support.
We would like to thank everybody involved in preparation of BAS2012 and hope that all BAS2012 participants will enjoy their stay in Barcelona.

Vladimir Zaiats
BAS2012 Coordinator

INVITED PAPERS

Cox proportional hazards model with measurement error

Alexander Kukush[1] and Elena Usoltseva[1]

[1] National Taras Shevchenko University of Kyiv, 01601 Kyiv, Volodymyrska st. 64, Ukraine

E-mail for correspondence: alexander_kukush@univ.kiev.ua

Abstract: Cox proportional hazards model under covariate measurement error is considered. We investigate a simultaneous estimation method for the baseline hazard and covariate parameter. The strong consistency of the considered estimators is obtained. We also estimate rate of convergence of the estimators in terms of Kullback-Leibler distance.

Keywords: Cox proportional hazards; measurement error; censored observations; baseline hazard function.

1 Model

We consider Cox semiparametric proportional hazards model with censored observations in the presence of measurement errors. The baseline hazard function is not parametrized and belongs to a infinite-dimensional compact set of continuous positive functions. We use the partial log-likelihood function and correct it for censoring and measurement error following the ideas of Augustin (2004). In the Cox proportional hazards model, the intensity of failure at time point t of an individual covariate vector X is specified as

$$\Lambda(t|X;\lambda,\beta) := \lambda(t)\exp(\beta^T X). \tag{1}$$

Here β is k-dimensional parameter, $\beta \in \Theta_\beta \subset R^k$, $\lambda(t) \in \Theta_\lambda \subset C[0,\tau], \tau > 0$. The pdf of the survival time T is equal to

$$f_T(t|X;\lambda,\beta) = \Lambda(t|X;\lambda,\beta)\exp\left(-\int_0^t \Lambda(s|X;\lambda,\beta)ds\right),$$

$$\int_0^\infty \Lambda(t|X;\lambda,\beta)dt = \infty,$$

where

$$\Lambda(t|X;\lambda,\beta) = \frac{f_T(t|X;\lambda,\beta)}{G_T(t|X;\lambda,\beta)}, \quad G_T(t|X;\lambda,\beta) := 1 - F_T(t|X;\lambda,\beta).$$

2 Estimator

First suppose that we observe independent triples (Y_i, X_i, Δ_i), $i = \overline{1,n}$, $Y_i = \min(T_i, C_i)$ are censored lifetimes T_i, and Δ_i are censorship indicators $\Delta_i = I(T_i \le C_i)$. We use the partial log-likelihood function for estimating $\lambda(t)$ and β,

$$Q_n(t) := \frac{1}{n}\sum_{i=1}^{n} q(Y_i, \Delta_i, X_i; \lambda, \beta),$$

$$\text{where} \quad q(Y, \Delta, X; \lambda, \beta) := \Delta\left(\log\lambda(Y) + \beta^T X\right) - e^{\beta^T X}\int_0^Y \lambda(u)du.$$

In the case of the presence of the measurement error we observe instead of X_i the surrogate data

$$W_i = X_i + U_i, \tag{2}$$

$i = \overline{1,n}$. The errors U_i are independent copies of a k-dimensional random vector U with known moment generating function $M_U(\beta) := \mathbf{E}e^{\beta^T U}$, ,and they are independent of $\{X_i, T_i, C_i\}$. In Augustin (2004) the corrected objective function was proposed,

$$Q_n^{cor}(\lambda, \beta) := \frac{1}{n}\sum_{i=1}^{n} q^{cor}(Y_i, \Delta_i, W_i; \lambda, \beta),$$

with

$$q^{cor}(Y, \Delta, W; \lambda, \beta) := \Delta\left(\log\lambda(Y) + \beta^T W\right) - \frac{e^{\beta^T W}}{M_U(\beta)}\int_0^Y \lambda(u)du.$$

The estimators $\left(\widehat{\lambda}_n, \widehat{\beta}_n\right)$ are defined as

$$\left(\widehat{\lambda}_n, \widehat{\beta}_n\right) := \arg\max_{(\lambda, beta)\in\Theta} Q_n^{cor}(\lambda, \beta), \quad \Theta := \Theta_\lambda \times \Theta_\beta.$$

3 Consistency of estimator

Our main results are obtained under the following assumptions.

1. $\Theta_\lambda \subset C[0,\tau]$ is the compact convex set of such positive functions $f : [0,\tau] \to R$ that $f(t) > a, \forall \in [0,\tau]$, $|f(t) - f(s)| \le L|t-s|, \forall t, s \in [0,\tau]$, where $a > 0$ and $L > 0$ are fixed constants.
2. $\Theta_\beta \subset R^k$ is compact and convex.
3. $\mathbf{E}U = 0$; for a fixed $\varepsilon > 0$, $\mathbf{E}e^{D\|U\|} < \infty$ where $D := \max_{\beta\in\Theta_\beta}\|\beta\| + \varepsilon$.
4. $\mathbf{E}e^{D\|X\|} < \infty$, where the positive constant D was defined above.

5. τ is the right endpoint of the distribution of censor C, i.e. $P(C > \tau) = 0$ and for all $\varepsilon > 0$ we have $P(C > \tau - \varepsilon) > 0$.

6. The covariance matrix S_X of the random vector X is positive definite.

Theorem 1 *In the Cox proportional hazards model under measurement error (1), (2) with true parameters λ_0 and β_0, assume that conditions (1)-(6) are satisfied. Then $\left(\widehat{\lambda}_n, \widehat{\beta}_n\right)$ are strongly consistent estimators, that is*

$$\sup_{t\in[0,\tau]} |\widehat{\lambda}_n(t) - \lambda_0| \to 0 \quad \textit{and} \quad \widehat{\beta}_n \to \beta_0 \quad \textit{a.s.} \quad \textit{as} \quad n \to \infty.$$

4 Rate of convergence

We give the rate of convergence of the estimator $\left(\widehat{\lambda}_n, \widehat{\beta}_n\right)$ defined above to the true parameter values (λ_0, β_0) under the following condition:

7. X has a density, $\mathbf{E}\left(e^{2D_\beta \|X\|} + e^{2D_\beta \|U\|}\right) < \infty$, $D_\beta := \max\limits_{\beta\in\Theta_\beta} \|\beta\| > 0$.

Theorem 2 *In the Cox proportional hazards model under measurement error (1), (2) with true parameters λ_0 and β_0, assume that conditions (1), (2), (4)-(7) hold. Then*

$$D\left(f\left(Y, \Delta, X; \lambda_0, \beta_0\right), f\left(Y, \Delta, X; \widehat{\lambda}_n, \widehat{\beta}_n\right)\right) = \frac{O_p(1)}{\sqrt{n}},$$

where f stands for the joint pdf of the triple (Y, Δ, X) w.r.t. a standard measure μ (here μ is a product of two Lebesgue measures and counting measure), and for densities f_1 and f_2 with respect to the measure μ,

$$D(f_1, f_2) := \int f_1(x) \log \frac{f_1(x)}{f_2(x)} d\mu(x)$$

denotes the Kullback-Leibler distance between f_1 and f_2.

The results are joint with Prof. I. Fazekas and Dr. S. Baran (Hungary) and published in Kukush et al. (2011).

References

Augustin, T. (2004). An exact corrected log-likelihood function for Coxs proportional hazards model under measurement error and some extensions. *Scandinavian Journal of Statistics*, **31**, 4350.

Kukush, A., Baran, S., Fazekas, I., Usoltseva, E. (2011). Simultaneous estimation of baseline hazard rate and regression parameters in Cox proportional hazards model with measurement error. *Journal of Statistical Research*, **45**, N2.

On Identification of Threshold Models for Time Series and Diffusion Processes

Yury A. Kutoyants[1]

[1] University of Maine, av. O. Messiaen, Le Mans, 72085, FRANCE

Abstract: We consider the problem of threshold estimation for autoregressive time series with a "space switching" in the situation, when the regression is nonlinear and the innovations have a smooth, possibly non Gaussian, probability density. Assuming that the unknown threshold parameter is sampled from a continuous positive density, we find the asymptotic distribution of the Bayes estimator and show that these estimators are asymptotically efficient. The similar problems for threshold diffusion processes are discussed too.

Keywords: TAR time series; threshold estimation; singular estimation.

1 Introduction

The simplest threshold autoregressive (TAR) process is the time series, generated by the recursion

$$X_{j+1} = \rho_1 X_j \mathbb{1}_{\{X_j<\vartheta\}} + \rho_2 X_j \mathbb{1}_{\{X_j\geq\vartheta\}} + \varepsilon_{j+1}, \qquad j = 0, \dots, n-1,$$

where $\varepsilon_j \sim \mathcal{N}\left(0, \sigma^2\right)$ are i.i.d. random variables, $\rho_1 \neq \rho_2, |\rho_i| < 1$ and σ^2 are known constants. The unknown *threshold* parameter $\vartheta \in \Theta = (\alpha, \beta)$ is to be estimated from the data $X^n = (X_0, X_1, \dots, X_n)$. This model and some of its generalizations has been extensively studied during the last decades (see Tong(2011)). Particularly, much attention focused on the properties of the least squares estimator (Chan (1993)).
This talk gives a review of some recent results concerning the study of the Bayes estimator for the TAR models. We present the results obtained by Chan and Kutoyants (2009), (2010), Chigansky and Kutoyants (2012), Kutoyants (2012). The case of colored noise (Chigansky and Kutoyants (2011)) will be presented separately by Chigansky.
We consider the following nonlinear TAR(1) model

$$X_{j+1} = h\left(X_j\right) \mathbb{1}_{\{X_j<\vartheta\}} + g\left(X_j\right) \mathbb{1}_{\{X_j\geq\vartheta\}} + \varepsilon_{j+1}, \qquad j = 0, \dots, n-1, \quad (1)$$

where $h(x)$ and $g(x)$ are known functions, (ε_j) are i.i.d. random variables with a known density function $f\left(x\right) > 0, x \in R$ and the initial condition X_0 is independent of (ε_j) and has a probability density $f_0(x)$. We suppose

that the condition of strong mixing are fulfilled and the time series has invariant density $\varphi(\vartheta, x)$.
The likelihood function of the sample X^n is given by

$$L(\vartheta, X^n) = f_0(X_0) \prod_{j=0}^{n-1} f\Big(X_{j+1} - h(X_j)\ \mathbb{1}_{\{X_j<\vartheta\}} - g(X_j)\ \mathbb{1}_{\{X_j\geq\vartheta\}}\Big),$$

and the Bayes estimator $\tilde{\vartheta}_n$ with respect to the mean square risk is the conditional expectation

$$\tilde{\vartheta}_n = \mathbf{E}(\vartheta|X^n) = \frac{\int_\Theta \theta\, p(\theta)\ L(\theta, X^n)\ \mathrm{d}\theta}{\int_\Theta p(\theta)\ L(\theta, X^n)\ \mathrm{d}\theta}.$$

Since the likelihood $L(\vartheta, X^n)$ is piecewise constant in ϑ, the estimate can be computed efficiently (see Chan and Kutoyants (2010)).
The asymptotic properties of $(\tilde{\vartheta}_n)$ are formulated in terms of the following compound Poisson process

$$Z(u) = \begin{cases} \exp\left(\sum_{l=1}^{N_+(u)} \ln \frac{f\left(\varepsilon_l^+ + \delta(\vartheta_0)\right)}{f\left(\varepsilon_l^+\right)}\right), & u \geq 0, \\ \exp\left(\sum_{l=1}^{N_-(-u)} \ln \frac{f\left(\varepsilon_l^- - \delta(\vartheta_0)\right)}{f\left(\varepsilon_l^-\right)}\right), & u < 0. \end{cases}$$

Here ϑ_0 is the true value of the patrameter, $\epsilon_l^\pm$ are independent random variables with the density function $f(x)$, $N_+(\cdot)$, $N_-(\cdot)$ are independent Poisson processes with the same intensity $\lambda = \varphi(\vartheta_0, \vartheta_0)$ ($Z(u) := 1$ on the sets $\{N_\pm(u) = 0\}$).
Define the random variable

$$\tilde{u} = \frac{\int_R u\, Z(u)\ \mathrm{d}u}{\int_R Z(u)\ \mathrm{d}u}.$$

We have the following lower bound on the mean square risk of an arbitrary sequence of estimators $(\bar{\vartheta}_n)$:

$$\varliminf_{\delta\to 0} \varliminf_{n\to\infty} \sup_{|\vartheta-\vartheta_0|<\delta} n^2 \mathbf{E}_\vartheta \left(\bar{\vartheta}_n - \vartheta\right)^2 \geq \mathbf{E}_{\vartheta_0} \tilde{u}^2,$$

and the Bayes estimates $(\widetilde{\vartheta}_n)$ are *efficient*, attaining this lower bound asymptotically. Our typical result is the following

Theorem 1 *The sequence of estimates $(\tilde{\vartheta}_n)$ is consistent, the convergence in distribution*

$$n\left(\tilde{\vartheta}_n - \vartheta_0\right) \Longrightarrow \tilde{u}$$

holds and the moments converge:

$$\lim_{n\to\infty} n^p \mathbf{E}_{\vartheta_0} \left|\tilde{\vartheta}_n - \vartheta_0\right|^p = \mathbf{E}_{\vartheta_0} |\tilde{u}|^p, \quad p > 0.$$

This result is generalized in several directions (many thresholds, TAR(p) process, misspesification models etc. At particularly, we consider diffusion process

$$\mathrm{d}X_t = \sum_{j=1}^{k+1} S_j\left(X_t\right)\ \mathbb{1}_{\{\vartheta_{j-1}<X_t\leq\vartheta_j\}}\mathrm{d}t + \sigma\left(X_t\right)\mathrm{d}W_t,$$

where $\vartheta_0 = -\infty$, $\vartheta_j \in \Theta_j = (\alpha_j, \beta_j), j = 1, \ldots, k$, $\vartheta_{k+1} = \infty$, $\beta_j < \alpha_{j+1}$. The functions $S_j(x)$ and $\sigma(x)$ are such that the process X_t is ergodic with invariant density $f(\vartheta, x)$.
Problem: *how to estimate ϑ by observations $X^T = (X_t, 0 \leq t \leq T)$ and what are the properties of estimators as $T \to \infty$?*
We describe the properties of the MLE and BE of the parameter ϑ (Kutoyants(2012))

References

Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Ann. Statist.*, **21**(1):520–533.

Chan,N.H. and Kutoyants Yu. A. (2009) On parameter estimation of threshold autoregressive models. to appear in *Statist. Inference Stoch. Process.*.

Chan,N.H. and Kutoyants Yu. A. (2010) Recent developments of threshold estimation for nonlinear time series. *Journal of the Japan Statistical Society*, **40**(2):277–308.

Chigansky, P. and Kutoyants, Yu. A. (2011) Estimation in threshold autoregressive models with correlated innovations, submitted.

Chigansky, P. and Kutoyants, Yu. A. (2012) On nonlinear TAR processes and threshold estimation, submitted.

Dachian, S. and Negri, I. (2011) On compound Poisson processes arising in change-point type statistical models as limiting likelihood ratio. *Statist. Inference Stoch. Process.*, **14**(3):255–271, .

Ibragimov, I.A. and Has'minskii, R. Z. (1981) *Statistical Estimation: Asymptotic Theory*. New York.

Kutoyants, Yu. A. (2012) On identification of the threshold diffusion processes. *Annals of the Institute of Statistical Mathematics*, **64**, 2, 383-413.

Tong, H. (2011) Threshold models in time series analysis - 30 years on. *Statistics and Its Interface*, 4(2):107–118.

On maximum integrated likelihood estimators

Aleksander Zaigrajew[1]

[1] N.Copernicus University, Chopin str. 12/18, 87-100 Toruń, Poland

E-mail for correspondence: alzaig@mat.uni.torun.pl

Abstract: The problem of parameter estimation in the presence of a nuisance parameter is considered. We concentrate on the situation when the nuisance parameter is either location or scale and compare two estimators: maximum likelihood estimator and maximum integrated likelihood estimator.

Keywords: Maximum likelihood estimator; Maximum integrated likelihood estimator; Scale parameter; Location parameter.

Let a sample $x = (x_1, x_2, \ldots, x_n)$ be drawn from an absolutely continuous distribution with the density $p(\cdot; \theta, \lambda)$, where θ is a parameter of interest and λ is a nuisance parameter. We do not assume the orthogonality of these parameters.
To estimate θ, one can adopt the well-known maximum likelihood (ML) method, that consists in taking the so-called likelihood function

$$L(x; \theta, \lambda) = \prod_{j=1}^{n} p(x_j; \theta, \lambda)$$

and searching for

$$(\theta^*, \lambda^*) \in \operatorname{Arg} \sup_{\theta, \lambda} \; L(x; \theta, \lambda).$$

Throughout the talk we assume that

$$\sup_{\theta, \lambda} \; L(x; \theta, \lambda) = \sup_{\theta} \sup_{\lambda} \; L(x; \theta, \lambda). \tag{1}$$

The function

$$\widehat{L}(x; \theta) = \sup_{\lambda} \; L(x; \theta, \lambda) = L(x; \theta, \widehat{\lambda}(x; \theta)),$$

where $\widehat{\lambda}(x; \theta) \in \operatorname{Arg} \sup_{\lambda} \; L(x; \theta, \lambda)$ is the ML estimator (MLE) of λ given θ, is known as the profile likelihood function. Under condition (1), the MLE θ^* of θ can be obtained by maximizing the profile likelihood function.

The main goal of the talk is to compare the MLE with another estimator of θ obtained by integration of the likelihood function, with a weight function, over the nuisance parameter.
The idea of using the integrated likelihood for estimation of the parameter of interest isn't new: see e.g. Berger et al. (1999) or Severini (2000, 2010).
The integrated likelihood can be written as

$$\widetilde{L}(x;\theta) = \int L(x;\theta,\lambda)w(\lambda)d\lambda,$$

where $w(\cdot)$ is the weight function which can be interpreted, using Bayesian language, as a conditional density *a priori* corresponding to the parameter λ given θ. There are several arguments to choose the weight function as follows: $w(\lambda) = \lambda^{-1}$, if λ is the scale parameter, and $w(\lambda) \equiv 1$, if λ is the location parameter.
Then as an estimator of θ (we call it MILE) we take $\theta^{**} \in \operatorname{Arg}\sup_{\theta} \widetilde{L}(x;\theta)$.
Some properties of the MILE is given in the next theorem.

Theorem 1 *Under usual regularity conditions on the distribution, the MILE θ^{**} is consistent and asymptotically normal with the same asymptotic distribution as the MLE θ^*, i.e. $n^{1/2}(\theta^{**}-\theta) \to \mathcal{N}(0,(H^{-1})_{11})$ in distribution, as $n \to \infty$, where $(H^{-1})_{11}$ is the left upper element of the inverse matrix with respect to the Fisher information matrix $H = H(\theta,\lambda)$.*

To compare the estimators we use two criteria: the bias and the mean square error. As an example of results obtaining, we present here that on comparison the biases of both estimators in the asymptotic case.

Theorem 2 *Under some regularity conditions on the distribution, at least for all sufficiently large values of n: the sign of the difference $|E\theta^* - \theta| - |E\theta^{**} - \theta|$ coincides with the sign of the product*

$$E\frac{b(x;\theta)}{h'(x;\theta)}E\Big(\frac{2nh(x;\theta)}{h'(x;\theta)} + \frac{nh''(x;\theta)h^2(x;\theta)}{(h'(x;\theta))^3} - \frac{b(x;\theta)}{h'(x;\theta)}\Big).$$

Here $b(x;\theta) = \frac{1}{2}\left(\ln(-l''_{\lambda\lambda}(x;\theta,\widehat{\lambda}(x;\theta)))\right)'_{\theta} - \left(\ln w(\widehat{\lambda}(x;\theta))\right)'_{\theta}$, $h(x;\theta) = \frac{1}{n}l'_{\theta}(x;\theta,\widehat{\lambda}(x;\theta)))$, $l(x;\theta,\lambda) = \ln L(x;\theta,\lambda)$.

Further on, we consider different two-parametric families of distributions. The results are supported with numerical calculations.

References

Berger, J. O., Liseo, B., and Wolpert R. (1999). Integrated likelihood functions for eliminating nuisance parameters (with discussion). *Statistical Science*, **14**, 1-28.

Severini T. A. (2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.

Severini T. A. (2010). Likelihood ratio statistics based on an integrated likelihood. *Biometrika*, **97**, 481-496.

Insurance premiums and risk measures: models and estimation

Ričardas Zitikis[1]

[1] Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario N6A 5B7, Canada

E-mail for correspondence: zitikis@stats.uwo.ca

Abstract: Numerous risk measures and premium calculation principles have appeared in the literature. Many of them are special cases of two highly encompassing functionals, distortion and weighted, which are based on two different modifications of probabilities with which insurance (or financial) risks and losses take on their values. Here we discuss these functionals, as well as related statistical inferential methods and results.

Keywords: Insurance; Losses; Risks; Premiums; Inference.

1 Introduction

Insurance losses are non-negative random variables, $X \in \mathbf{R}_+$. The mean of X is the net premium $\mathbf{E}[X]$, to which we add a loading to have a useful premium. For an overview, we refer to Young (2004). Serious issues arise when constructing premiums, including the choice of loading: should it reflect the mean loss, volatility, or something else? The reason is that the loading is not just a reflection of the severity of the loss X but also of the risk perception by those (to be) insured, as well as of many other factors. Naturally, decision theory under risk and uncertainty plays a pivotal role in defining risk measures and premiums. In summary, given the net premium, we want to modify it in such a way that the resulting premium would reflect:

1. The distribution of the loss X as well as with it associated collateral loss, which can be expenses associated with the claim processing time, human resources required, etc. Hence, mathematically, we deal with the transformed loss $v(X)$ for a function v. We may view v as a utility function (think of classical economic theory) or perhaps as a value function (think of behavioural economics).

2. The potential distortion of loss probabilities, due to various factors (natural and artificial). Mathematically, this means modifying the

underlying de-cumulative distribution function (ddf) or, alternatively, the probability density function (pdf) of the loss X.

We shall next discuss these topics in more detail.

2 Distortion premiums and risk measures

To begin, we write the net premium $\mathbf{E}[X]$ as the integral

$$\int_{\mathbf{R}_+} S(x)dx, \tag{1}$$

where $S(x) = \mathbf{P}[X > x]$ is the ddf of X, also known as the survival function. Modifying the values of X means using a 'value' function v as the integrator in (1). Distorting probabilities means integrating the function $g(S(x))$ instead of the ddf $S(x)$ in (1) for a 'distortion' function g. Since we want $g(S(x))$ to be a ddf, we use a non-decreasing $g : [0,1] \to [0,1]$ such that $g(0) = 0$, $g(1) = 1$, and $g(t) \geq t$ for all $t \in [0,1]$. In summary, we have the distortion risk measure

$$\Delta_{g,v}[X] := \int_{\mathbf{R}_+} g(S(x))dv(x). \tag{2}$$

The role of this risk measure in insurance was discussed and explored in a series of pioneering papers by Shaun Wang (e.g., Wang (1998) and references therein). A thorough mathematical treatment of integral (2) was given by Denneberg (1994). For a role of $\Delta_{g,v}[X]$ in economics, we refer to Quiggin (1982), Schmeidler (1986), Yaari (1987), Quiggin (1993), and references therein. An impetus for developing statistical inferential results in the area was given by Jones and Zitikis (2003), who noted a close relationship between $\Delta_{g,v}[X]$ and L-statistics. Necir et al. (2007) explained how to handle heavy-tailed losses when estimating $\Delta_{g,v}[X]$. Numerous articles dealing with statistical inferential results have appeared during the last five years or so (cf., e.g., Brazauskas et al. (2008), Greselin et al. (2009), Zitikis et al. (2010), and references therein).

3 Weighted premiums and risk measures

This time, we write the net premium $\mathbf{E}[X]$ as the integral

$$\int_{\mathbf{R}_+} xdF(x), \tag{3}$$

where $F(x) = \mathbf{P}[X \leq x]$ is the cumulative distribution function (cdf) of the loss X. Modifying the values of X with a function v means integrating $v(x)$

instead of x, whereas 'weighing' the cdf F with a function $w : \mathbf{R}_+ \to \mathbf{R}_+$ means replacing F with its weighted version F_w, which is defined by the equation

$$dF_w(x) = \frac{w(x)dF(x)}{\int_{\mathbf{R}_+} w(z)dF(z)}. \tag{4}$$

In summary, from (3) and (4) we have the weighted premium (cf. Furman and Zitikis, 2008a)

$$\Pi_{v,w}[X] := \int_{\mathbf{R}_+} v(x)dF_w(x) = \frac{\mathbf{E}[v(X)w(X)]}{\mathbf{E}[w(X)]}.$$

Various special cases of the weighted premium have appeared in the literature, but in full generality, the premium was discussed and explored by Furman and Zitikis (2008a, 2009, 2010). For related econometric, insurance, and financial insights, we refer to Schechtman et al. (2008), Furman and Zitikis (2008b, 2009), and references therein. Statistical inferential results have been explored by Necir and Zitikis (2012), who concentrate on the case of heavy-tailed losses. As we see from equation (4), the weighted premium $\Pi_{v,w}[X]$ relies on the notion of weighted distributions, which have been extensively discussed in the statistical literature (cf., e.g., Rao (1997), Patil (2002), and references therein).

Acknowledgments: The research has been partially supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Brazauskas, V., Jones, B.L., Puri, M.L. and Zitikis, R. (2008). Estimating conditional tail expectation with actuarial applications in view. *Journal of Statistical Planning and Inference*, **138** (11, Special Issue in Honor of Junjiro Ogawa: Design of Experiments, Multivariate Analysis and Statistical Inference), 3590–3604.

Denneberg, D. (1994). *Non-additive Measure and Integral*. Dordrecht: Kluwer.

Furman, E. and Zitikis, R. (2008a). Weighted premium calculation principles. *Insurance: Mathematics and Economics*, **42**, 459–465.

Furman, E., and Zitikis, R. (2008b). Weighted risk capital allocations. *Insurance: Mathematics and Economics*, **43**, 263–269.

Furman, E. and Zitikis, R. (2009). Weighted pricing functionals with applications to insurance: an overview. *North American Actuarial Journal*, **13**, 1–14.

Furman, E. and Zitikis, R. (2010). Weighted pricing functionals: a new method and its manifold implications. In: *Proceedings of the International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management*, edited by I. Frenkel, I. Gertsbakh, L. Khvatskin, Z. Laslo, and A. Lisnianski), 283–289, Shamoon College of Engineering, Beer Sheva, Israel.

Greselin, F., Puri, M.L. and Zitikis, R. (2009). *L*-functions, processes, and statistics in measuring economic inequality and actuarial risks. *Statistics and Its Interface*, **2**, 227–245.

Jones, B.L. and Zitikis, R. (2003). Empirical estimation of risk measures and related quantities. *North American Actuarial Journal*, **7**, 44–54.

Necir, A., Meraghni, D., and Meddi, F. (2007). Statistical estimate of the proportional hazard premium of loss. *Scandinavian Actuarial Journal*, **2007**, 147–161.

Necir, A. and Zitikis, R. (2012). Coupled risk measures and their empirical estimation when losses follow heavy-tailed distributions. (Submitted for publication.)

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, **3**, 323–343.

Quiggin, J. (1993). *Generalized Expected Utility Theory.* Dordrecht: Kluwer.

Patil, G.P. (2002). Weighted distributions. In: *Encyclopedia of Environmetrics*, edited by A.H. El-Shaarawi and W.W. Piegorsch. 2369–2377, Chichester: Wiley.

Rao, C.R. (1997). *Statistics and Truth: Putting Chance to Work.* (Second edition.) Singapore: World Scientific.

Schechtman, E., Shelef, A., Yitzhaki, S., and Zitikis, R. (2008). Testing hypotheses about absolute concentration curves and marginal conditional stochastic dominance. *Econometric Theory*, **24**, 1044–1062.

Schmeidler, D. (1986). Integral representation without additivity. *Proceedings of the American Mathematical Society*, **97**, 255-61.

Wang, S. (1998). An actuarial index of the right-tail risk. *North American Actuarial Journal*, **2**, 88–101.

Yaari, M.E. (1987). The dual theory of choice under risk. *Econometrica*, **55**, 95–115.

Young, V.R. (2004). Premium principles. In: *Encyclopedia of Actuarial Science,* edited by J. Teugels and B. Sundt. New York: Wiley.

Zitikis, R., et al. (2010). Special Issue on "Actuarial and Financial Risks: Models, Statistical Inference, and Case Studies." *Journal of Probability and Statistics*, **2010**.

CONTRIBUTED PAPERS

Forecasting the number of occupational accidents in Turkey by using a hybrid forecasting technique

Cagdas Hakan Aladag[1] and Sibel Aladag[2]

[1] Hacettepe University/Department of Statistics, Ankara, Turkey
[2] Republic of Turkey Social Security Institution/General Director of Service Provision, Ankara, Turkey

E-mail for correspondence: aladag@hacettepe.edu.tr

Abstract: Artificial neural networks (ANN) have been widely used in various time series forecasting problems in recent years because of the ability to model both the linear and the nonlinear parts of time series. Although ANN produces accurate forecasts in many time series implementations, there are still some problems with using ANN. ANN method consists of some main components such as architecture structure, learning algorithm and activation function. It is a well-known fact that these components directly affect the forecasting performance of ANN. Since artificial neural networks approach is data-driven method, determining the elements of ANN issue should be carefully considered due to the data examined. An important decision is the selection of optimum architecture of neural network that consists of determining the numbers of neurons in the layers of the network. Therefore, various approaches have been proposed to find the best ANN architecture in the literature. On the other hand, trial and error method have been still the most preferred method to find a good architecture when ANN method is utilized to forecast time series. In this study, occupational accidents in Turkey is forecasted by using a hybrid heuristic method proposed by Aladag (2011) which is based on feed forward neural networks and Tabu search algorithm. Data was collected from Republic of Turkey Social Security Institution from January 2003 to December 2011. In Aladag's (2011) forecasting approach, tabu search heuristic method is utilized to determine the best neural network architecture which gives the most accurate forecasts. Occupational accidents issue is very important for every country in the world. Hence, forecasting the number of occupational accidents accurately is a crucial problem. As a result of the implementation, it is observed that the hybrid forecasting approach produces accurate forecasts for the number of occupational accidents in Turkey.

Keywords: Artificial neural networks; Forecasting; Occupational accidents; Tabu search; Time series.

1 Introduction

ANN approach has been successfully used in time series forecasting in recent years. In the literature, there have been various forecasting studies in which it is observed that ANN method produces very accurate forecasts (Aladag and Aladag, 2011). In spite of the fact that ANN is an effective forecasting tool for time series, there are still some problems with using ANN (Aladag et al., 2010b). ANN is a data driven method so the components of the method should be determined due to the data. Main components of the method can be given as architecture structure, learning algorithm and activation function (Egrioglu, 2008). Especially, choosing the best architecture is an important decision in order to reach accurate forecasts (Aladag, 2011). In the literature, there have been some systematic approaches to find the best architectures but trial and error method is the most preferred method (Aladag et al., 2010a). In this study, occupational accidents in Turkey is forecasted by using a hybrid heuristic method proposed by Aladag (2011) which is based on feed forward neural networks and tabu search algorithm. To forecast same time series, trial and error method was also utilized. The results obtained from tabu search algorithm and trial and error methods were obtained with each other and it was observed that the most accurate forecasts are obtained when tabu search algorithm was used to determine the best architecture.

2 The implementation

In this study, occupational accidents in Turkey is forecasted by using a hybrid heuristic method proposed by Aladag (2011) which is based on feed forward neural networks and Tabu search algorithm. Data was collected from Republic of Turkey Social Security Institution from January 2003 to December 2011. In Aladag's (2011) forecasting approach, tabu search heuristic method is utilized to determine the best neural network architecture which gives the most accurate forecasts. The time series is also forecasted with FFNN in which trial and error method is employed to find best architecture. The time series has 121 observations. The first 110 and the last 12 observations are used for training and test sets, respectively.
The best feed forward neural network architecture, which has the minimum objective function value, was tried to be found by using the hybrid forecasting approach. Before searching process was started, other elements of the networks are fixed like in Aladag (2011). The logistic activation function is used in all of the neurons of networks. Levenberg Marquardt algorithm is employed as training algorithm. In all computations, Matlab 2010 computer package is utilized. Besides, same parameters which were used in Aladag (2011) are employed for tabu search algorithm. (See Aladag [10] for detailed information about the parameters of the used tabu search algorithm).

Observed	Tabu Search	Trial&Error
3963	4226.48	5174.79
5400	4317.09	5281.53
5633	5639.32	5517.52
6523	6125.12	5420.91
6193	6351.20	5805.44
6721	6650.38	6570.22
5647	6690.41	5301.75
6483	6541.01	7263.70
5577	6350.07	6505.20
6568	6439.02	5851.06
5204	6331.70	5801.76
5954	6367.37	6964.25
hline TBA	2-4-1	12-1-1
RMSE	617.74	728.32

TABLE 1. The obtained forecasting results

As mentioned above, when the time series is forecasted with FFNN, to determine the best architecture, both the trial and error, and the tabu search algorithm are exploited. The obtained results are summarized in Table 1. In the table, observed and forecast values produced by trial and error, and tabu search methods are presented. Also, the best architectures (TBA) found and corresponding root mean square error (RMSE) values are shown in Table 1. When the tabu search algorithm is employed in the architecture selection, the architecture 2-4-1, which contains 2, 4, and 1 neurons in the input, the hidden, and the output layers, respectively, is determined as the best architecture with 617.74 RMSE. The obtained results calculated over the test set are also visually examined. As a result of the implementation, it can be said that using tabu search algorithm proposed by Aladag (2011) to find the best architecture produces accurate forecasts for the number of occupational accidents in Turkey.

3 Conclusions

Occupational accidents issue is very important for every country in the world (Dimitrov, 2011). Hence, forecasting the number of occupational accidents accurately is a crucial problem (Aidoo and Eshun, 2012). In this study, a hybrid forecasting method combines feed forward neural networks (FFNN) and tabu search algorithm is employed to forecast occupational accidents in Turkey in order to reach high forecasting accuracy level.
The hybrid forecasting approaches proposed by Aladag (2001) is employed in this study. In this method, tabu search algorithm is utilized to solve

architecture selection problem. It is aimed to obtain accurate forecasts for the number of occupational accidents in Turkey. It is a well-known fact that ANN approach can produce more accurate forecasts that those produced by conventional time series (Zhang et al., 1998). Therefore, the time series are analyzed with FFNN and to find the best architecture, which gives the most accurate forecasts, both tabu search and trial and error methods are employed. Then, the obtained forecasting results are compared with each other. As a result of the comparison, it is observed that the best architecture which was found by tabu search produces better forecasts than those obtained from the best architecture determined by trial and error method.

References

Aidoo, S.J., and Eshun, P.A. (2012). Time series model of occupational injuries analysis in Ghanaian mines-a case study. *Research Journal of Environmental and Earth Sciences*, **4**, 162-165.

Aladag, C.H. (2011). A new architecture selection method based on tabu search for artificial neural networks. *Expert Systems with Applications*, **38**, 3287-3293..

Aladag, C.H., and Aladag, S. (2011). Forecasting total health expenditures with a hybrid heuristic method. In: Proceedings of 12th IEEE International Symposium on Computational Intelligence and Informatics 243-246, Budapest, Hungary.

Aladag, C.H., Egrioglu, E., and Kadilar, C. (2010a). Modeling brain wave data by using artificial neural networks *Hacettepe Journal of Mathematics and Statistics*, **39**, 81-88.

Aladag, C.H., et al. (2010b). Improving weighted information criterion by using optimization. *Journal of Computational and Applied Mathematics*, **233**, 2683-2687.

Dimitrov, P. M. (2011). Forecasting the number of occupational accidents in bulgaria through exponential smoothing. *International Journal of Contemporary Economics and Administrative Sciences*, **1**, 208-221.

Egrioglu, E., Aladag, C.H., and Gunay, S. (2008). A new model selection strategy in artificial neural network. *Applied Mathematics and Computation*, **195**, 591-597.

Oraee, S.K., Yazdani-Chamzini, A., and Basiri, M.H (2011). Forecasting the number of fatal injuries in underground coal mines. In: Proceedings of SME Annual Meeting 11-049, Denver, CO.

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

On the Stability of Estimation in Autoregressive Models with Exponential Innovations

Lynda Atil[1], Hocine Fellag[2], Cherifa Belkacem[1]

[1] Laboratory of Pure and Applied Mathematics, Mouloud Mammeri University, Tizi-Ouzou, Algeria and Mouloud Mammeri University, Tizi-Ouzou, Algeria
[2] Laboratory of Pure and Applied Mathematics, Mouloud Mammeri University, Tizi-Ouzou, Algeria

E-mail for correspondence: atillynda@yahoo.fr

Abstract: The aim of our paper is to present an exhaustive study of the estimation of first order autoregressive models with exponential innovations under various kinds of contamination. Some theoretical aspects and Monte Carlo results are presented to study the behavior of this estimator.

Keywords: Autoregressive; exponential; estimation; stability.

1 The model

Consider the first order autoregressive process of the form

$$X_t = \rho X_{t-1} + \varepsilon_t, \quad t = 0, 1, ..., n. \tag{1}$$

Where the ε_t's are independently distributed according to an exponential distribution $Exp(1)$, i.e., the density of ε_t is

$$f_{\varepsilon_t}(y) = e^{-y}, \quad y > 0.$$

Suppose that all what we observe is a segment of the process

$$X_1, X_2, ..., X_n, \quad n \quad \text{fixed}, \tag{2}$$

and ρ is unknown and is to be estimated.
Assuming that X_0 is distributed according to $Exp(1-\rho)$, the maximum likelihood estimator for ρ is (see Andel (1988))

$$\widehat{\rho} = \min_{2 \le t \le n} (X_t / X_{t-1}) = \rho + \min_{2 \le t \le n} \frac{\varepsilon_t}{X_{t-1}} \tag{3}$$

Andel (1988) considered the following modification. Since the process is stationary with mean $m = 1/(1-\rho)$, he proposed a new estimator which was

$$\widehat{\rho}^{*} = \rho + \frac{1}{m} \min_{2 \leq t \leq n} \varepsilon_t \tag{4}$$

He found then the distribution for this new estimator of ρ.
The aim of our paper is to present some results concerning the bias and the MSE of $\widehat{\rho}^{*}$ when the innovations are contaminated following the Sinha model (see Sinha and Kale, (1969)).

References

Andel, J. (1988). On AR(1) processes with exponential white noise. *Commun. Statist. Theory. Meth.* **17**, 1481-1495.

Fellag, H. (2001). Testing on the first order autoregressive model with contaminated exponential white noise finite sample case. *Discussiones Mathematicae Probability and Statistics.* **21**, 11-20.

Nielsen, B., and Shephard, N. (2003). Likelihood analysis of a first order autoregressive model with exponential innovations *Journal of Time Series Analysis.* **24**, 337-344.

Provost, S. B., and Sanjel, D. (2005). Inference about the first order autoregressive coefficient. *Commun. Statist. Theory. Meth.* **34**, 1183-1201.

Gene filtering with optimal threshold selection

Josep Bau-Macià[1], Jordi Solé-Casals[1], Cesar F. Caiafa[2,3] and Sergio Lew[4]

[1] University of Vic, Sagrada Família 7, 08500, Vic, SPAIN
[2] IAR-CONICET, C.C.5, (1894) Villa Elisa, Buenos Aires, ARGENTINA
[3] FIUBA, Av. Paseo Colón 850 (1063), Capital Federal, ARGENTINA
[4] IIBM-FIUBA, Av. Paseo Colón 850 (1063), Capital Federal, ARGENTINA

E-mail for correspondence: josep.bau@uvic.cat

Abstract: Gene filtering is a useful preprocessing technique often applied to microarray datasets. However, it is no common practice because clear guidelines are lacking and it bears the risk of excluding some potentially relevant genes. In this work, we propose to model microarray data as a mixture of two Gaussian distributions that will allow us to obtain an optimal filter threshold in terms of the gene expression level.

Keywords: Gene expression; gene filtering; microarray data; MOG model.

1 Introduction

Non-specific gene filtering is a very useful technique as it increases the sensitivity of the microarray data analysis and reduces the dimensionality of the dataset. Thus, correct and stringent filtering will substantially reduce the problem of overfitting in classification problems. Several filtering approaches exist, some of them often used in combination. The most used are based on (i) filtering by expression level and (ii) filtering by gene variance across samples. These techniques involve the use of more or less subjective thresholds. The drawback of data-independent thresholds is that gene expression distributions are very variable between different microarray datasets and can result in too stringent or too loose filtering conditions. In this work we develop a data-driven selection of a threshold based on the minimization of the classification error.

2 Materials and Methods

ALL Dataset: The Acute Lymphoblastic Leukemia (ALL) data were reported by Chiaretti et al [Chiaretti et al, S. 2004]. We consider the comparison of the 37 samples from patients with the BCR/ABL fusion gene

resulting from a chromosomal translocation (9;22) with the 42 samples from the NEG group. They are available in the R package ALL. The comparisons conducted in this work maintain the criteria of removing the genes with inter-quartile range (IQR) below 0.5 used by Scholtens and Heydebreck [Scholtens D. and Von Heydebreck A., 2005] but using an optimized threshold for intensity filtering instead of $6.64 = \log_2(100)$.

Mixture Of Gaussian (MOG) model: We assume that each gene belongs to one of the following two classes: class C_1 (un-expressed genes), class C_2 (expressed genes), and each one of these classes can be well modeled using a Gaussian distribution with specific means μ_1, μ_2, and standard deviations σ_1, σ_2. Then, the probability density functions (pdfs) for gene expression values conditioned to a particular class is: $p(x|C_p) = \Phi\left(\frac{x-\mu_p}{\sigma_p}\right)$, $(p = 1, 2)$ with $\Phi(x) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}x^2\right)$ being the zero-mean and unit-norm Gaussian pdf. Then, the pdf for the variable x (gene expression value) is: $p(x) = \alpha_1 p(x|C_1) + \alpha_2 p(x|C_2)$, where $\alpha_p = P(C_p)$ $(p = 1, 2)$ are the probabilities of each class.

Fitting the MOG model: We use the Maximum Likelihood (ML) criterion to fit the model. Since it is difficult to obtain a closed form of the likelihood for a MOG model, a well-known solution is to use the Expectation-Maximization (EM) algorithm. **Optimal threshold selection:** Once the MOG model is fitted to the available data we need to determine the optimal threshold h to classify samples as belonging to C_1 $(x < h)$ or C_2 $(x \geq h)$. Our objective is to choose h such that the error of classification is minimized. It is easy to show that such a value of h must satisfy $\alpha_1\Phi\left(\frac{h-\mu_1}{\sigma_1}\right) = \alpha_2\Phi\left(\frac{h-\mu_2}{\sigma_2}\right)$, which can be explicitly solved.

Differential expression analysis: NEG samples were compared to BCR/ABL samples applying a Welch t-test for equality of the mean expression levels in the two groups in order to obtain the differential expression p-value for each gene.

3 Results

For ALL dataset, after 151 iterations the MOG model converged to an optimal intensity threshold (OIT) value of 4.17 on a log2 scale (Figure 1). This value is clearly lower than the 6.64 arbitrary intensity threshold (AIT) used by Scholtens and Heydebreck. Table 1 shows how the selection of the threshold affects to the number of significant genes ($p_{val} < 0.05$) discarded by the filtering process. Using an arbitrary threshold set up at 6.64 the total number of discarded significant genes is 101, which represents the 61.6% of significant genes of the whole dataset. On the other hand, using our optimal threshold at 4.17 determined by the MOG model, the total number of discarded significant genes is 37 which represents the 22.6% of significant genes of the whole dataset. Clearly our method increases the number of significant genes not discarded by the filtering process.

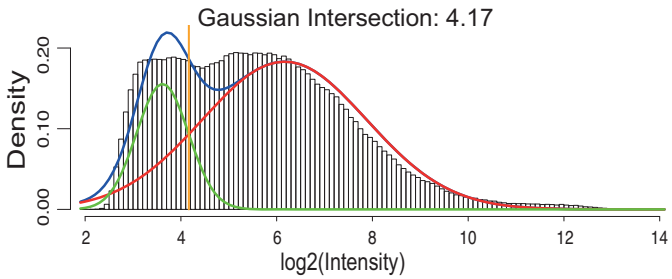

FIGURE 1. Gaussian mix model of the ALL data.

TABLE 1. Discarded genes depending on the selected threshold.

	#Genes	% Genes	#Genes $p_{val} < 0.05$	%Genes $p_{val} < 0.05$
Total	12625	100%	164	100%
Discarded with AIT ($h = 6.64$)	10231	81%	101	61.6%
Discarded with OIT ($h = 4.17$)	8599	68.1%	37	22.6%

4 Conclusions

A new method for the automatic selection of a threshold for filtering genes in microarray datasets has been proposed and compared to classical filtering techniques. Our experimental results on the ALL dataset demonstrates the advantage of using the proposed technique.

Acknowledgments: This work has been in part supported by the MINCYT-MICINN Research Program 2010-2011 (Ref. AR2009-0010) and by the University of Vic under the grants R0904 and R0901.

References

Scholtens D. and Von Heydebreck A. (2005). *Analysis of Differential Gene Expression Studies. Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Statistics for Biology and Health.* **Part III**, 229-248.

Chiaretti S., Li X., Gentleman R., Vitale A., Vignetti M., Mandelli F., Ritz J., Foa R. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood.* **103**, 2771-2778.

Bivariate risk models in insurance: ruin and survival probabilities

Castañer, A.[1], Claramunt, M.M.[1], Lefèvre, C.[2]

[1] Universitat de Barcelona, Departament de Matemàtica Econòmica, Financera i Actuarial, Av. Diagonal,690, 08034 Barcelona.
[2] Université Libre de Bruxelles, Département de Mathématique, Boulevard du Triomphe, Campus de la Plaine C.P. 210, 1050 Bruxelles.

E-mail for correspondence: mmclaramunt@ub.edu

Abstract: This paper deals with an insurance portfolio that covers two interdependent risks. The model is a discrete-time bivariate risk process with independent claim increments.Our main purpose is to develop a numerical method for determining non-ruin probabilities over a finite-time horizon. The approach relies on, and exploits, the existence of a special algebraic structure of Appell type.

Keywords: Multirisks model; Discrete time; Finite-time ruin probability; Recursive methods

1 Introduction

The present work is concerned with the evaluation of non-ruin probabilities for a two-dimensional risk process. The extension to multivariate cases is rather simple. The model is a discrete-time process with independent claim increments. Our model is a variant of the multivariate model studied by Picard et al. (2003) and is an extension of the univariate model investigated by Castañer et al. (2011).
Let $\{U_1(t), U_2(t)\}$ be the bivariate surplus process for the two risks. Following e.g. Cai and Li (2007), we consider three possible definitions.
(i) Ruin occurs at time T_{or} as soon as one of the two surpluses becomes negative. Thus, $T_{or} = \min(T_1, T_2)$ where $T_i = \inf\{t \geq 1 : U_i(t) < 0\}$ is the ruin time for risk i.
(ii) Ruin occurs at time T_{and} when the two surpluses become negative, not necessarily at the same time. Thus, $T_{and} = \max(T_1, T_2)$.
(iii) Ruin occurs at time T_{sim} when the two surpluses become negative simultaneously, i.e. at the same time. Thus, $T_{sim} = \inf\{t \geq 1 : U_1(t) < 0 \text{ and } U_2(t) < 0\}$.
Our main purpose is to derive, for these three definitions of ruin, a simple formula that enables us to calculate survival probabilities over a finite-time horizon.

2 A discrete-time bivariate risk model

The risk process is a discrete-time model for two dependent risks, labelled 1 and 2. Let $t \in N = \{0, 1, \ldots\}$ be the time scale. The initial reserves are u_1 and u_2. During each period $(t-1, t]$, $t \geq 1$, the company receives a total premium income of $c_{1,t}$ for risk 1 and $c_{2,t}$ for risk 2. These premiums are collected at the beginning of the period, i.e. at time $(t-1)^+$, as often (other cases might be considered).
The total claim amounts during $(t-1, t]$, $t \geq 1$, are non-negative random variables $X_{1,t}$ for risk 1 and $X_{2,t}$ for risk 2. These amounts are registered at the end of the period, i.e. at time t. Claim amounts have any continuous distribution, possibly time-dependent, but with also an atom at 0 to allow the possibility of no claim. The two risk processes have independent increments, i.e. the random vectors $(X_{1,t}, X_{2,t})$, $t \geq 1$, are independent.
For each risk $i = 1, 2$, during the first t periods, the aggregate premiums with the initial reserves are

$$h_i(t) = u_i + c_{i,1} + \ldots + c_{i,t},$$

and the aggregate claim amounts to be covered are

$$S_i(t) = \sum_{j=1}^{t} X_{i,j}, \quad t \geq 1.$$

Let $F_t(s_1, s_2)$ denote the joint distribution function of $[S_1(t), S_2(t)]$. We recall that, for instance, $F_t(\infty, 0) = P[S_2(t) = 0] > 0$ by hypothesis. The two surpluses $U_i(t)$ are then given by

$$U_i(t) = h_i(t) - S_i(t), \quad t \geq 1.$$

We are going to propose a method to evaluate the non-ruin probabilities over any finite horizon, for the three definitions of ruin indicated in the introduction. The key case will be concerned with the definition (i).

Finite-time survival probabilities. Consider any time $t \geq 1$ and define

$$\phi_{or}(t, x_1, x_2) = P[T_{or} > t,\, U_1(t) \geq x_1,\, U_2(t) \geq x_2].$$

This is the probability that ruin for each risk does not occur until t and the two surpluses at t are at least equal to x_1 and x_2. By construction, $0 \leq x_1 \leq h_1(t)$, $0 \leq x_2 \leq h_2(t)$. Note also that

$$\phi_{or}(t, 0, 0) \equiv \phi_{or}(t) = P(T_{or} > t) = P(T_1 > t \text{ and } T_2 > t).$$

Theorem 1 (Proposition) *For the definition (i),*

$$\phi_{or}(t,x_1,x_2) = F_t(0,0) + \int_{w_1=0}^{h_1(t)-x_1} b(w_1,0)\, F_t[h_1(t)-x_1-w_1,0]\, dw_1$$
$$+ \int_{w_2=0}^{h_2(t)-x_2} b(0,w_2)\, F_t[0,h_2(t)-x_2-w_2]\, dw_2$$
$$+ \int_{w_1=0}^{h_1(t)-x_1} \int_{w_2=0}^{h_2(t)-x_2} b(w_1,w_2) F_t[h_1(t)-x_1-w_1,h_2(t)-x_2-w_2] dw_1 dw_2,$$

where $b(w_1,w_2)$ is a real function satisfying the equations

$$\begin{aligned} 0 &= \int_{w_1=0}^{s_1} b(s_1-w_1,0)\, dF_{v(s_1,0)}(w_1,0), \quad s_1>0, \\ 0 &= \int_{w_2=0}^{s_2} b(0,s_2-w_2)\, dF_{v(0,s_2)}(0,w_2), \quad s_2>0, \\ 0 &= \int_{w_1=0}^{s_1}\int_{w_2=0}^{s_2} b(s_1-w_1,s_2-w_2)\, dF_{v(s_1,s_2)}(w_1,w_2), \quad s_1,s_2>0. \end{aligned}$$

being $v(s_1,s_2)$ a family of integers, for any non-negative reals s_1, s_2 defined as follows. If $s_1 \le h_1(1)$ and $s_2 \le h_2(1)$, then $v(s_1,s_2)=0$; otherwise, put

$$v(s_1,s_2) = \sup\{t \ge 1 : h_1(t) < s_1 \text{ or } h_2(t) < s_2\}.$$

The marginal survival probabilities are easily deduced. For instance, to obtain $\phi_1(t,X_1) = P(T_1 > t, U_1(t) \ge x_1)$, it suffices to put above $x_2 = 0$ and $X_{2,t} = 0$ a.s. for all t. So, for each risk $i = 1,2$, let $F_{i,t}(s)$ be the distribution function of $S_i(t)$, and define $v_i(s) = 0$ if $s \le u_i$, otherwise

$$v_i(s) = \sup\{t \in N : h_i(t) < s_i\}.$$

We then get the result that can be found in Castañer et al. (2011).

Acknowledgments: The authors thank the Ministerio de Educación y Ciencia of Spain (MICINN) for support under grant ECO2010-22065-C03-03. Part of this work has been done using the resources at the Centre de Serveis Científics i Acadèmics de Catalunya (CESCA). C. Lefèvre is grateful to the Banque Nationale de Belgique for support.

References

Cai, J., and Li, H. (2007). Dependence properties and bounds for ruin probabilities in multivariate compound risk models. *Journal of Multivariate Analysis*, **98**, 757-773.

Castañer, A., Claramunt, M.M., Gathy, M. Lefèvre, C., and Mármol, M. (2011). Ruin problems for a discrete time risk model with non-homogeneous conditions. *Scandinavian Actuarial Journal*, (forthcoming).

Picard, P., Lefèvre, C., and Coulibaly, I. (2003). Multirisks model and finite-time ruin probabilities. *Methodology and Computing in Applied Probability*, **5**, 337-353.

Detection of distributions with bounded support

Joan del Castillo[1], Maria Padilla[1]

[1] Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain (castillo@mat.uab.cat, mpadilla@mat.uab.cat)

Abstract: The peaks over thresholds method is used for estimating high quantiles of distributions. The main goal is select the point from which the generalized Pareto distribution (GPD) may be used. Here we are interested in this problem when the tail distribution has compact support, as happens in hydrology and other environmental sciences.

Keywords: Extreme values theory; Peaks over thresholds.

1 Introduction

The generalized Pareto distribution (GPD) is closely related to the extreme values theory. Since Pickands (1975) is well known that the residual distribution of any random variable follows approximately a GPD. The cumulative distribution function for the GPD is given by

$$F(x) = 1 - (1 + \xi x/\psi)^{-1/\xi} \tag{1}$$

where $\psi > 0$ and ξ are scale and shape parameters. For $\xi > 0$ the range of x is $x > 0$ and the GPD is the Pareto distribution. For $\xi > 0$ the range of x is $0 < x < -\psi/\xi$; then the GPD have bounded support. The limit case $\xi = 0$ corresponds to the exponential distribution.
Castillo, Daoudi and Lockhart (2011) have developed new methods to distinguish between polynomial and exponential tails of a distribution. In this work the applicability of the above methods to distinguish between tails with exponential decrease and tails with finite support is studied. This allows us to distinguish between tails that behave asymptotically as GPD with $\xi < 0$ and $\xi = 0$. These distributions have been used to model exceedances in fields such as hydrology, see Castillo and Hadi (1997).

2 The residual coefficient of variation

The coefficient of variation (CV) of a GPD plays an important role in this work. It determines the behaviour of the likelihood function, see Castillo and Daoudi (2009), and a constant residual CV characterizes the GPD, see

Gupta and Kirmani (2000).
The CV of the residual distribution for a GPD from a threshold u takes the following form

$$CV(u) = \sqrt{1/(1-2\xi)}, \tag{2}$$

which is always defined for the case of finite variance ($\xi < 0.5$) in particular when there are compact support (the case $\xi < 0$), it takes values smaller than 1. In the extreme case $\xi = 0$ we have that CV is exactly 1.
The methods of Castillo, Daoudi and Lockhart (2011) are based on the study of the process that assigns each threshold the coefficient of variation of the residual distribution.
Given that this value remains constant for any residual distribution from a threshold u, we can make the graphic analysis of a sample of data drawn for each threshold value of CV (u), so if there is a trend around a constant value smaller than 1 may be thought to follow the distribution of these data is compactly supported. This chart is called CV-plot.
We have applied the graphics methods and we have made different contrasts on data from Bilbao waves, see Castillo and Hadi (1997), and we have compared our results with theirs. Figure 1 shows the CV-plot of these data set having first subtracted the value 7, of the initial threshold. All graphics have been made ââin R language, see R Development Core Team (2010).

3 The mixture of distributions

In this work we study the mixture of GDP with finite support, which is also a distribution with finite support, but with a much larger coefficient of variation of the distributions involved in the mixture.
Using the parametritzation $\beta = -1/\xi > 0$, $\sigma = -\psi/\xi$, the probability density function of the GPD with compact support is

$$p(x;\beta,\sigma) = \beta\sigma^{-1}(1 - x/\sigma)^{\beta-1}, \tag{3}$$

for $0 < x < \sigma$, therefore, the mixture of GPD with finite support takes the following form

$$f(x) = \lambda p_1(x) + (1-\lambda)p_2(x) \tag{4}$$

where p_i(x)=p(x; β_i, σ_i).
Given λ, σ_1 and σ_2, taking β_1=β_2= β from the following expression

$$\beta = \frac{-\sigma_1^2\lambda + 4\sigma_1\lambda\sigma_2 - 4\sigma_1\lambda^2\sigma_2 + 2\sigma_2^2\lambda^2 - 3\sigma_2^2\lambda + 2\sigma_1^2\lambda + \sigma_2^2}{\lambda(2\sigma_1\lambda\sigma_2 - \sigma_2^2\lambda + \sigma_1^2 - 2\sigma_1\sigma_2 + \sigma_2^2 - \sigma_1^2\lambda)} \tag{5}$$

gives a mixture distribution with CV(0)=1. Hence, the samples from the corresponding distribution looks like and exponential distribution and we face the problem of select the threshold from which the residual distribution belongs to GPD. A test based on the CV could accept a mixture as exponential, while our tests based on different thresholds, are able to detect

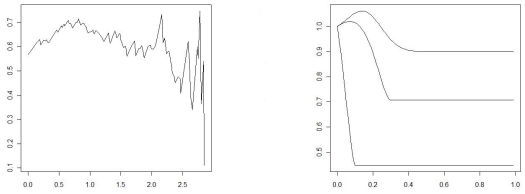

FIGURE 1. In left CV-plot from Bilbao waves. In right theoretical CV-plot for different mixtures.

cases as we have just outlined.
Figure 1 shows the theoretical CV-plot of three GPD mixtures with compact support; parameters given by the Table 1, which shows also the CV of the mixture and the two GPD appearing in it.

TABLE 1. Parameters selected to represent CV-plot.

λ	β_1	σ_1	β_2	σ_2	CV(0)	$CV_1\,(0)$	$CV_2\,(0)$
0.9	12.76	0.7	8.53	1	1	0.93	0.9
0.6	1.92	0.3	2	1	1	0.70	0.707
0.45	1.69	0.1	0.5	1	1	0.94	0.45

References

Castillo, J., Daoudi, J. (2009). Estimation of the generalized Pareto distribution. *Statistics and Probability Letters*, **79**, 684-688.

Castillo, J., Daoudi, J., Lockhart, R. (2011). Methods to distinguish between polynomial and exponential tails. *arXiv:*1112.0514.

Castillo, E., Hadi, S. (1997). Fitting the Generalized Pareto Distribution to Data. *Journal of the American Statistical Association*, **92**, 1609-1620.

Gupta, R., Kirmani, S. (2000). Residual coefficient of variation and some characterization results. *Journal of Statistical Planning and Inference*, **91:1**, 23-31.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, **3**, 119-131.

R Development Core Team (2012). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria.*.

Threshold estimation in autoregressive models driven by colored noise

P. Chigansky[1], Yu. Kutoyants[2], R. Liptser[3]

[1] Department of Statistics, The Hebrew University, Mount Scopus, Jerusalem, Israel
[2] Laboratoire de Statistique et Processus, Universite du Maine, France
[3] School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel

E-mail for correspondence: pchiga@mscc.huji.ac.il

Abstract: Large sample statistical analysis of threshold autoregressive (TAR) models is usually based on the assumption that the underlying driving noise is white. In this paper, we consider a model, driven by Gaussian colored noise with geometric correlation tail and derive a complete characterization of the asymptotic distribution for the Bayes estimator of the threshold parameter.

Keywords: Threshold autoregression (TAR) models; Parameter estimation; Hidden Markov models.

1 Introduction

Let (X_j) be the sequence generated by the recursion

$$X_j = \Big(\rho_+ 1_{\{X_{j-1}\geq\theta\}} + \rho_- 1_{\{X_{j-1}<\theta\}}\Big)X_{j-1} + \epsilon_j, \quad j \geq 1 \tag{1}$$

where (ϵ_j) is a random process with known distribution, ρ_+ and ρ_- are known constants and θ is the unknown threshold parameter to be estimated from the sample $X^n := (X_1, ..., X_n)$. The equation (1) is a basic instance of the threshold autoregression (TAR) models, which play a considerable role in the theory and practice of time series. This type of models have been studied by statisticians for already more than three decades, producing interesting theory and finding many important applications, some of which can be traced in the early and more recent surveys Tong (1983), Tong (2011), Tsay (1989), Hansen (2011), Chan and Kutoyants (2010).
When it comes to the asymptotic analysis of the estimators, the standard conditions imposed on the models such as (1) is strong ergodicity of the observed process (X_j) and independence of ϵ_j's. Departure from these assumptions often poses challenging problems. While the former assumption has been relaxed by a number of authors Pham et al (1991), Caner and

Hansen (2001), Liu et al (2011) the independence assumption of the driving noise sequence has not yet been addressed.

As we shall shortly see, in the dependent case the problem falls into the framework of statistical inference of hidden Markov models (HMM), where the driving noise plays the role of the hidden signal (see Ch. 10-12 in Cappe et al (2005) and the references therein). However, most of the HMM literature deals with locally asymptotically normal (LAN) experiments and, to the best of our knowledge, non-LAN models with partial observations have not yet been studied systematically.

In this paper, we consider the model (1) in which (ϵ_j) is a sequence with geometrically decaying correlation. More precisely, let $X = (X_j)$ be generated by the recursion

$$X_j = \Big(\rho_+ 1_{\{X_{j-1}\geq\theta\}} + \rho_- 1_{\{X_{j-1}<\theta\}}\Big)X_{j-1} + \xi_{j-1} + \varepsilon_j, \tag{2}$$

subject to $X_0 \sim N(0,1)$, where $\rho := |\rho_+| \vee |\rho_-| < 1$ and the unknown parameter θ takes values in an open bounded subset of the real line Θ. We shall consider the problem with discontinuous drift function $f(x,\theta) := \big(\rho_+ 1_{\{x\geq\theta\}} + \rho_- 1_{\{x<\theta\}}\big)x$, and thus assume $\rho_+ \neq \rho_-$ and $0 \notin \Theta$. The driving noises (ε_j) and (ξ_j) are independent: the *white noise* component (ε_j) is a sequence of i.i.d. $N(0,1)$ random variables and the *colored noise* (ξ_j) is the Gaussian process, generated by the linear recursion

$$\xi_j = a\xi_{j-1} + \zeta_j, \quad j \geq 1, \tag{3}$$

where (ζ_j) are i.i.d. $N(0,1)$ random variables and a is a known constant $|a| < 1$, controlling the bandwidth of the noise.

As we shall see below, the threshold θ can be estimated at the rate of n and hence the asymptotic analysis of the estimators for the other parameters, such as ρ_+, ρ_- and a, can be essentially carried out within the LAN framework (see Chan (1993)).

The recursions (2) and (3) form a conditionally Gaussian system, which means that the conditional law of ξ_n given X^n is Gaussian, and by Theorem 13.5 in Liptser and Shiryaev (2001)

$$X_j = f(X_{j-1},\theta) + \widehat{\xi}_{j-1}(\theta) + \sqrt{1+\gamma}\,\widehat{\varepsilon}_j, \tag{4}$$

where $(\widehat{\varepsilon}_j)$ is the innovation sequence of i.i.d. $N(0,1)$ random variables. The process $\widehat{\xi}_j(\theta) := E_\theta\big(\xi_j|\mathcal{F}_j^X\big)$ satisfies the Kalman filter equation

$$\widehat{\xi}_j(\theta) = a\widehat{\xi}_{j-1}(\theta) + \frac{a\gamma}{1+\gamma}\Big(X_j - f(X_{j-1},\theta) - \widehat{\xi}_{j-1}(\theta)\Big), \tag{5}$$

subject to $\widehat{\xi}_0 = 0$, where $\gamma := E_\theta\big(\xi_j(\theta) - \widehat{\xi}_j\big)^2 = E_\theta\xi_0^2$ is the unique positive root of

$$\gamma = a^2\gamma + 1 - \frac{a^2\gamma^2}{1+\gamma}.$$

The the innovation representation (4) implies that the likelihood of the data X^n is given by

$$L_n(X^n;\theta) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}X_0^2\Big) \times$$
$$\left(\frac{1}{\sqrt{2\pi(1+\gamma)}}\right)^n \exp\left(-\frac{1}{2}\frac{1}{1+\gamma}\sum_{j=1}^{n}\Big(X_j - f(X_{j-1},\theta) - \widehat{\xi}_{j-1}(\theta)\Big)^2\right).$$

The likelihood function is discontinuous in θ and hence we are faced with an irregular statistical experiment. In such problems, the maximum likelihood estimator is often asymptotically inferior to the Bayes estimator $\widetilde{\theta}_n$ and the latter is typically efficient for arbitrary positive priors in the asymptotic minimax sense (see Theorem 9.1, Ibragimov and Hasminskii (1981)). Since the likelihood function is piecewise constant in θ and has at most n jumps at $\{X_0, ..., X_{n-1}\}$, the Bayes estimator for the problem at hand has relatively low computational complexity.

2 The main result

The main result of the presented paper Chigansky et al (2011) is the following characterization of the asymptotic distribution of the sequence of Bayes estimators:

Theorem 1 *Let $(\widetilde{\theta}_n)$ be the sequence of the Bayes estimators with respect to the quadratic loss function and a prior with continuous positive density π. Then for any continuous function ϕ with at most polynomial growth*

$$\lim_n E_{\theta_0}\phi\Big(n\big(\widetilde{\theta}_n - \theta_0\big)\Big) = E_{\theta_0}\phi(\tilde{u}),$$

uniformly on compacts from Θ, where

$$\tilde{u} = \frac{\int_R uZ(u)du}{\int_R Z(u)du}$$

and $\ln Z(u)$, $u \in R$ is the following two sided compound Poisson process:

$$\ln Z(u) = \begin{cases} \sum_{j=1}^{\Pi^+(u)}\Big(\beta\varepsilon_j^+ - \frac{1}{2}\beta^2\Big) & u \geq 0 \\ \sum_{j=1}^{\Pi^-(|u|)}\Big(\beta\varepsilon_j^- - \frac{1}{2}\beta^2\Big) & u < 0 \end{cases} \tag{6}$$

Here Π^+, Π^- are i.i.d Poisson processes with the intensity

$$\varpi = \int_R p(\theta_0, y; \theta_0)dy,$$

$p(x,y;\theta_0)$ is the unique invariant probability density of the Markov process (X_j,ξ_j) under P_{θ_0}, $\left(\varepsilon_j^{\pm}\right)$ are i.i.d. $N(0,1)$ random variables, independent of Π^+ and Π^- and

$$\beta^2=\left(\frac{\theta_0(\rho_+-\rho_-)}{\sqrt{1+\gamma}}\right)^2\left(1+\left(\frac{a\gamma}{1+\gamma}\right)^2\sum_{j=0}^{\infty}\left(\frac{a}{1+\gamma}\right)^{2j}\right)=$$
$$\theta_0^2(\rho_+-\rho_-)^2\frac{1+\gamma^3}{(1+\gamma)(1+\gamma^2)}.$$

Acknowledgments: The first author is supported by ISF grant 314/09

References

Caner, M and Hansen, B.E. (2001). Threshold autoregression with a unit root *Econometrica*, 69(6):1555–1596

Cappé, O. Moulines, E. and Rydén, T (2005) *Inference in hidden Markov models*. New York: Springer

Chan, K.S. (1993) Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Ann. Statist.*, 21(1):520–533

Chan, N.H. and Kutoyants, Yu. A. (2010) Recent developments of threshold estimation for nonlinear time series. *Journal of the Japan Statistical Society*, 40(2):277–308.

Chan, N.H. and Kutoyants, Yu. A. (2012) On parameter estimation of threshold autoregressive models, arXiv preprint 1003.3800, to appear in *Statistical Inference for Stochastic Processes*

Chigansky, P and Kutoyants, Y and Liptser, R. Threshold estimation in autoregressive models driven by colored noise, *arXiv preprint 1108.1536*

Hansen, B.E. Threshold autoregression in economics. *Statistics and Its Interface*, 4:123–127, 2011.

Ibragimov, I. A. and Has'minskii, R. Z. (1981) *Statistical Estimation: Asymptotic Theory*. New York: Springer

Liptser, R.S. and Shiryaev, A.N. (2001) *Statistics of random processes, Vol. II*, Berlin: Springer-Verlag

Liu, W. Ling, S. and Shao, Q. (2011) On non-stationary threshold autoregressive models. *Bernoulli*, 17(3):969–986.

Dinh Tuan Pham, K. S. Chan, and Howell Tong (1991) Strong consistency of the least squares estimator for a nonergodic threshold autoregressive model. *Statist. Sinica*, 1(2):361–369.

Tong, H. (1983). *Threshold models in nonlinear time series analysis*. New York: Springer-Verlag.

Tong, H. (2011). Threshold models in time series analysis - 30 years on. *Statistics and Its Interface*, 4(2):107–118

ROC curves in distance-based credit risk models

Costa, T.[1], Boj, E.[1], Fortiana, J.[2]

[1] Departament de Matemàtica Econòmica, Financera i Actuarial, Universitat de Barcelona
[2] Departament de Probabilitat, Lògica i Estadística, Universitat de Barcelona

E-mail for correspondence: tcosta@ub.edu

Abstract: In this work we analyse goodness-of-prediction aspects of distance-based logistic regression in the framework of credit risk, concerning the dependence of predictions on the cut-off point, as shown in the ROC curve. We illustrate these analyses with a real credit risk portfolio. We use our *R* package, *dbstats*, in computations.

Keywords: Distance-based prediction; Logistic regression; Credit risk; *R*.

1 Empirical analysis

In Boj *et al.* (2011a) we apply Distance-Based (DB) logistic regression to the *Australian data set*, using a cut-off point of 0.5. We obtain a competitive result compared with other techniques such as discriminant analysis, ordinary logistic regression, k nearest neighbor, and decision trees (see West, 2000). In the present contribution we study the goodness-of-prediction of the DB logistic model for different cut-off points.
The credit scoring data set concerns credit card applications, *http://archive.ics.uci.edu/ml/datasets/Statlog +(Australian+Credit+Approval)*. Data are quite balanced, with 307 good and 383 bad credit risks, the set of predictors consisting of six continuous and eight categorical nominal variables. We calculate the predictor distance matrix, `D2`, with Gower's similarity index. We fit the model with the *dbstats* package for *R* (Boj *et al.*, 2011b):

```
dbglm.D2(y, D2, family = binomial (link = "logit"), maxiter =
   50, eps1 = 0.05, eps2 = 0.05, rel.gvar = 0.99)
```

We summarize in Table 1 the next quantities for different cut-off points: "Good credit", proportion of applicants that are creditworthy but are classified as a bad credit risk; "Bad credit", proportion of applicants that are not creditworthy but are incorrectly identified as creditworthy; "Overall", proportion of all applicants that are incorrectly classified; "Cost (0.144)"

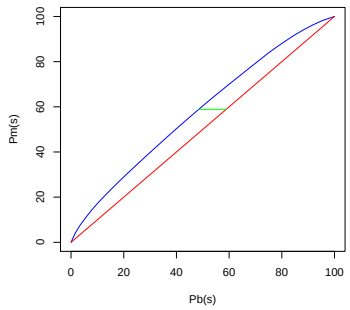

FIGURE 1. ROC curve of the Australian data set fitted with the DB logistic regression.

and "Cost (0.249)", error cost function of formula (13) in West (2000) for $\pi_2 = 0.144$ and 0.249 respectively. We obtain the minimums for "Overall", "Cost (0.144)" and "Cost (0.249)" in the cut-off points 0.49, 0.48 and 0.3. Figure 1 gives the well known ROC curve for the DB logistic model. We obtain that the cut-off must be 0.51 to maximize the K-S coefficient. Then, we can consider that for this data set the cut-off of 0.5 is adequate.

Acknowledgments: Authors have partially been supported by the Spanish Ministerio de Educación y Ciencia under grant MTM2010-17323, and by the Generalitat de Catalunya, AGAUR under grant 2009SGR970.

References

Boj, E., Fortiana, J., Esteve, A., Claramunt, M.M. and Costa, T. (2011a). Aplicación de un modelo de regresión logística basado en distancias en el problema de credit scoring. In: Feria, J. M. y otros (2011). *Investigaciones en Seguros y Gestión de riesgos: RIESGO 2011*. Cuadernos de la Fundación MAPFRE, **171**. Fundación MAPFRE Estudios, Madrid. (293 - 305).

Boj, E., Caballè, A., Delicado, P. and Fortiana, J. (2011b). *dbstats: Distance-based statistics (dbstats)*. R package version 1.0.1, URL *http://CRAN.R-project.org/package=dbstats*.

West, D. (2000). Neural network credit scoring models. *Computer & Operations Research*, **27**, 1131-1152.

TABLE 1. Probabilities and costs of the Australian data set fitted with the DB logistic regression for different cut-off points.

Cut-off	Good credit	Bad credit	Overall	Cost (0.144)	Cost (0.249)
0.05	0.478827	0.005222	0.215942	0.247207	0.224456
0.1	0.348534	0.005222	0.157971	0.194817	0.176993
0.15	0.260586	0.010444	0.121739	0.161661	0.152452
0.2	0.224756	0.010444	0.105797	0.143740	0.136246
0.25	0.179153	0.013055	0.086957	0.122738	0.119614
0.3	0.143322	0.020888	0.075362	0.111145	**0.115617**
0.35	0.133550	0.026110	0.073913	0.110860	0.119483
0.4	0.127036	0.036554	0.076812	0.117568	0.133595
0.41	0.120521	0.039164	0.075362	0.116097	0.134136
0.42	0.107492	0.041775	0.071014	0.110344	0.130647
0.43	0.100977	0.041775	0.068116	0.106125	0.126714
0.44	0.097720	0.041775	0.066667	0.104003	0.124737
0.45	0.097720	0.044386	0.068116	0.106481	0.128884
0.46	0.094463	0.044386	0.066667	0.104337	0.126882
0.47	0.091205	0.046997	0.066667	0.104624	0.128961
0.48	0.087948	0.046997	**0.065217**	**0.102449**	0.126928
0.49	0.081433	0.052219	**0.065217**	0.102837	0.130840
0.5	0.081433	0.054830	0.066667	0.105198	0.134801
0.51	0.081433	0.060052	0.069565	0.109879	0.142651
0.52	0.078176	0.060052	0.068116	0.107618	0.140516
0.53	0.071661	0.070496	0.071014	0.112128	0.151449
0.54	0.068404	0.073107	0.071014	0.112012	0.152963
0.55	0.065147	0.073107	0.069565	0.109653	0.150720
0.56	0.061889	0.075718	0.069565	0.109471	0.152151
0.57	0.061889	0.075718	0.069565	0.109471	0.152151
0.58	0.061889	0.078329	0.071014	0.111645	0.155810
0.59	0.061889	0.080940	0.072464	0.113807	0.159448
0.6	0.061889	0.083551	0.073913	0.115957	0.163065
0.65	0.045603	0.107050	0.079710	0.122046	0.182363
0.7	0.035831	0.127937	0.086957	0.129554	0.200771
0.75	0.029316	0.138381	0.089855	0.131444	0.207930
0.8	0.026059	0.180157	0.111594	0.156267	0.252096
0.85	0.019544	0.227154	0.134783	0.178450	0.294083
0.9	0.013029	0.292428	0.168116	0.206764	0.346924
0.95	0.000000	0.407311	0.226087	0.242592	0.419482

A model of a random binary process reconstruction from truncated Walsh-Hadamard Spectrum

S.Dolev[1], S.Frenkel[2]

[1] Department of Computer Science, Ben-Gurion University of the Negev, Israel
[2] Institute of Informatics Problems, Russian Academy of Sciences

E-mail for correspondence: fsergei@mail.ru

Abstract: Walsh-Hadamard Tramsformation (WHT) (and WHT-based codes) of digital random sequences is used widely in many computer science and data transmission areas. In some cases, in order to enhance performance and to restrict memory resources as well, a truncated set of the WHT coefficients is used. In this presentation we suggest a model for accuracy of reconstruction a binary sequence from a truncated Walsh-Hadamard series, and analyze possible probabilistic models which could be used in the framework of suggested conceptual model. We consider the number of erroneously reconstructed bits as the accuracy measure.

Keywords: Walsh-Hadamard Transformation, Random Processes, Coding Theory

1 Introduction

Walsh-Hadamard tramsformation (WHT) (and WHT-based codes) of digital random sequences is used widely in many computer science and data transmission areas, for example, for image data transmission (H. Y. Jung, R. Prost, T. Y. Choi (1997)). In some cases, in order to enhance performance and to restrict memory resources as well, many system's designers try to restrict the number of the WHT coefficients used as much as possible. A question is to define a model in order to estimate and predict reconstruction accuracy from this truncated WH series. We will consider the number of erroneously reconstructed bits as the accuracy measure. In contrast to well-known mean square metric used for estimation of Fourier transformation-based reconstruction accuracy, which is an L_2 metric, we consider an L_1 one, as it equal to the Hamming distance between original and reconstructed sequence. One of principal challenges in the using of truncated Walsh-Hadamard series to reconstruct a binary sequence is necessary to apply a threshold rule in the reconstruction process as the values of truncated Walsh-Hadamard (WH) series have not to belong to

the $GF(2)$ field. Therefore, we must decide for each estimation of original value, computed by the truncated WH series, to what of two quantization level (0 or 1) the estimated value belongs. In contrast to traditional signal quantization task, where there are some plausible reasons to consider the "'quantization noise" as a random normal process independent of the reconstructed signal, in our case there is not any explicit reasons for such conclusion. In this presentation we suggest a model for accuracy of reconstruction a binary sequence from a truncated WH series. We analyze possible probabilistic models which could be used in the framework of suggested conceptual model.

2 Waslh-Hadamard Spectrum of a binary sequence

The Walsh-Hadamard transformation is based on a complete set of orthogonal functions. That is, if $b = (b_1, b_2, \ldots, b_m)$ is a binary file (sequence), then n-character encoding of the file b can be represented as $c^T = W^T b$, where $c = (c_1, \ldots, c_n)$, $c = 2^k$, k is an integer, is the Walsh-Hadamard coefficients. These orthogonal functions use only the values 1 or -1. More detailed, the spectral coefficients of WHT are $c_h = (1/n)\sum_{i=0}^{n-1} b_i W(h,i)$, and the inverse transform is $b_i = \sum_{h=0}^{n-1} c_i W(h,i)$.
Let $b = b_0, b_2 \ldots, b_{n-1}$ be an uncorrelated ("'white-noise'-like") sequence of n bits, where n is a power of two integer, and, due to the uncorrelation, $Prob(b_i = 1)$=$Prob(b_i = 0)$=1/2. Note, that these settings fit several applications for example, when a secure data sequence is produced by pseudo-random generator, for example as in S.Dolev and S.Frenkel (2010).
Let us we use for the original sequence reconstruction only $l \ll n$ WHT coefficients $c_1, \ldots, c_l$. In this case, we can estimate each bit b_i of the randomized sequence b by WHT mentioned above as $\hat{b}_i = \tilde{b}_i + e_i(l)$, where $\tilde{b}_i = \sum_{j=0}^{l} c_j W(j,i)$, and $e_i(l) = \sum_{q=l+1}^{n-1} c_q W(q,i)$.
Our goal is to compute a metric that captures the difference of the bits b_i and $\tilde{b}_i$. The result may depend on the coefficients we choose for reconstruction, in dependency on the application requirements. Each coefficient c_i is transmitted/stored with its index i in the WHT matrix, namely the pairs $(c_i; i)$ are stored as the representation of the data. We may consider various ways of the l choice, for example, either random choice of l coefficients (which can be reasonable, say, for distributed communication channels), or using first greatest l coefficients.

3 WHT coefficients choice

Inverse WHT with partial sums may result in non-binary values, that differ from binary domain of original sequence. Therefore, the reconstruction metric should be considered along with a decision rule mapping each value

to a corresponding binary value. We suggest to round the values to the closest value in the field during the decoding process.
We consider a probabilistic model that takes into account the rounding and the mentioned above accuracy metric requirement.
The reconstructed estimation of a bit $b_i = round(\hat{b}_i)$, where $\hat{b}_i$ is the estimation of the i-th value before rounding, computed by a partial sum of inverse WHT, is determined by the following random events:
$e_0 : (b_i = 0), e_1 : (b_i = 1)$, that is the bit b_i of randomized file F is 0 (event e_0) or 1 (event e_1), $v_{i0} : \tilde{b}_i \leq 1/2, v_{i1} : \tilde{b}_i \geq 1/2$, (defined on the space of the rational values $\tilde{b}_i$).
Let $Pr_{err=0}(i)$ be the probability that the actually zero bit b_i was erroneously reconstructed as $b_i = 1$, and $Pr_{err=1}(i)$ be the probability that the bit $b_i = 1$ was erroneously reconstructed as $b_i = 0$.
Both the probabilities $Prob(v_{i0})$, $Prob(v_{i1})$ are the probabilities of the partial sums mentioned above that have a value that can be estimated to be close to 1/2. Formally, in order to estimate error of the sequence reconstruction by truncated number of coefficients we should know both joint and marginal distributions both the sum of l terms of the WHT $S_l = \sum_{j=0}^{l} c_j W(j,i))$ and sum of residue $S_R = \sum_{j=n-l+1}^{N} c_j W(j,i)$. Then, taking into account that the sum $S_l + S_R$ is an exact value $b_i = 0$ or 1, we could compute the error probability as $Prob(S_l \geq Tr/S_l + S_R = 0) + Prob(S_l \leq Tr/S_l + S_R = 1)$. In accordance with Theorem 6.4 in P. A. Morettin (1981), WHT coefficients are distributed (asymptotically) as some independent normal random values with zero mean and dispersion of $n \times f(i)$, where i is the WHT coefficient index and $f(i)$ is the (dyadic) spectral density of b.
If we deal with random chosen of $l \ll n$ coefficients, we may use an assumption about independence of the partial sums from the complete one. In this case we use the following way to estimate the error.
Let l be large enough for using an asymptotic approach for the normal approximation of the WHT sums. Then, taking into account that the WHT coefficients are orthogonal, and that asymptotic distribution of the Inverse WHT sums is also normal, we can get that the error probability can be expressed as a probability that a normal distributed random variable v, taking the values in accordance with the events v_{i0},v_{i1}, falls into an interval $[a, b]$, that is:

$$Prob(a \leq v \leq b) = \Phi((b - E(v))/s_v) - \Phi((a - E(v))/s_v) \tag{1}$$

$$Prob(v_{io}) = Prob(b_{min} \leq v \leq 1/2) \tag{2}$$

where $s_v = sqrt(Var(v))$

$$Prob(v_{i1}) = Prob(1/2 \leq v \leq b_{max}) \tag{3}$$

where b_{min}, b_{max} are lower and upper bounds on v.

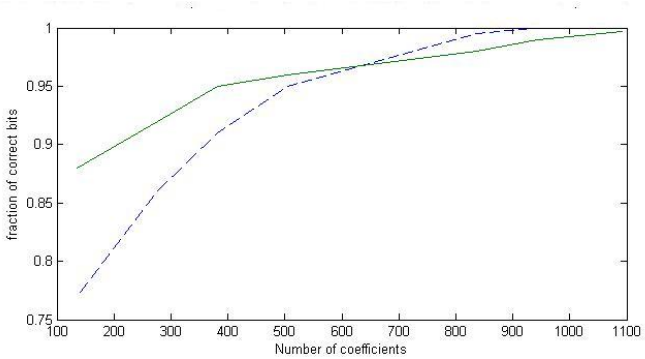

FIGURE 1. Fractions of correctly reconstructed bits (for a 2K bit file) vrs. number of coefficients. The solid line is the theoretical curve (Equations (1)-(3)), the dashed line is the experimental results.

Standard intervals based on $\pm\ \sigma$ can be used for b_{min} and b_{max}. Then we compute the number of correct reconstructed bits as $n(\sum_{i=0}^{n} Pr_{err}(i))$. Thus, using Equations (1) to (3) we may compute the probability of erroneous reconstruction, which depends on the number of transmitted coefficients l, the relationship between l and the length of the file n, and the coefficients values (i.e., the value of the sum $\sum_{j=0}^{l} |c_j|$).

4 Most significant coefficients

Let us consider Walsh-Hadamard series truncation via a choice of most significant coefficients. The most simple way to estimate significance of a coefficients subset is a choice of l largest (over absolute values) coefficients. In this case assumption about independency of partial and total sums is not true totally. First, each next largest coefficients depends on the previously chosen coefficients, secondly, as the sum for b_i is the sum of the largest coefficients, in general there is no reason to suggest that the sum of the truncated Inverse WHT has insignificant influence on the full Inverse WHT sum. Besides, there is an ambiguity in the definition of choice of l largest (over absolute values) coefficients if there exist pair of coefficients c_i, c_j, such that $abs(c_i) = abs(c_j)$. Indeed, there is a question in this case, what of the two coefficients should be included in the l-set. It is possible to use an identification of all WHT coefficients indexes that contribute significantly to the binary sequences energy, that is the sum of the sequence of Boolean ones. Following A. C. Gilbert et al.(2002) there is a polynomial-complex algorithm of the coefficients choice.

Thus, in both ways we keep the coefficients in accordance with the mean square (energy) criterion, but we use an L_1 metric in order to approve them. This is a consequence of the fact, that the result of computations of the Inverse truncated WHT are perform in a real domain whereas the estimation metric is formed for the Hamming distance of $d(x;y) = |b - \tilde{b}|$ that is an object from GF(2). Further, if we consider ordered WHT coefficients as an order statistic, we may reduce our problem of the original file reconstruction by $l \ll n$ coefficients to known method of a linear combination distribution estimation (R. Arellano-Valle and A. Genton (2007)).

Acknowledgments: The first author has partially been supported by the Lynne and William Frankel Center for Computer Science and the Rita Altura Trust Chair in Computer Science. The second author has partially been supported by the Russian Foundation for Basic Research under grant RFBR 12-07-00109.

References

H. Y. Jung, R. Prost, T. Y. Choi (1997), A unified mathematical form of the Walsh-Hadamard transform for lossless image data compression, *Signal Processing*, Vol. 63 pp. 35-43, 1997.

S.Dolev, S. Frenkel (2010), A way of coding and decoding of digital data based on digital holograpphy principles, *Patent of Russian Federation 2010145892/08(066164)* of 11.11.2010.

P. A. Morettin (1981), Walsh Spectral Analysis, *SIAM Review*, vol. 23, pp. 277-291, 1981.

R. Arellano-Valle, A. Genton (2007), On the exact distribution of linear combinations of order statistics from dependent random variables, *Journal of Multivariate Analysis*, vol.98, pp.1876-1894, 2007.

A. C. Gilbert, S. Guhay, P. Indykz, S. Muthukrishnan, M. Strauss (2002), Near Optimal Sparse Fourier Representations via Sampling,*STOC 2002*, May 1921, 2002, Montreal, Quebec, Canada.

Robust estimators of the parameters of Student-distribution

Sándor Fegyverneki[1]

[1] Department of Applied Mathematics, University of Miskolc,
H-3515 Miskolc, Miskolc-Egyetemváros, Hungary

E-mail for correspondence: matfs@gold.uni-miskolc.hu

Abstract: This article gives a construction of a weighted mean which has good robust properties (qualitative robustness, bounded influence function, high breakdown points). This method can be applied to some other statistical problems: the estimation of shape parameter at stable, Student- and Weibull-family.

Keywords: Robust statistics; Student-distribution; M-estimators.

1 Introduction

One of the simplest statistical problems is the location-scale problem on the real line. Given a data set $\{x_1, x_2, \ldots, x_n\}$, we are required to specify two numbers T_n and s_n, together with upper and lower bounds, which describe the location and the scale, respectively, of the data with given probability. In spite of its apparent simplicity, the problem has as yet no satisfactory solution. Most approaches including robust ones are based on a central model G_0 which is assumed to be true or to contain the truth within some small metric ball. Data rarely come accompanied by a central model and when analyzing large numbers of data sets in an automatic manner, such an approach is unwarranted. The estimation of the parameters is a well discussed problem. Usually, the location parameter μ and the scale parameter σ are estimated by the maximum likelihood (ML) estimators $\hat{\mu}$ and $\hat{\sigma}$, respectively, which are asymptotically the best.
Our location and scale problem is the following: Let us assume that $\xi = \sigma\eta + \mu$, where the distribution of the random variable η is $G_0(x)$. Given the sample $\xi_1, \xi_2, \ldots, \xi_n$ and the type of distribution G_0, the distribution of the random variable ξ_i is $G_0\left(\dfrac{x-\mu}{\sigma}\right)$ with estimate the location ($\mu \in \mathbf{R}$) and scale ($\sigma > 0$) parameters from the sample.
We can solve this statistical problem, e.g., by the maximum likelihood or by the method of moments. However, if g_0 is the density function and the derivative of $-\ln g_0$ is not bounded then the maximum likelihood estima-

tion is very sensitive to outliers in the sample and the moments are also sensitive.

The system of equations for the parameters, using Huber's (1981) notations, is given by $\sum_{i=1}^{n} \psi\left(\frac{\xi_i - \mu}{\sigma}\right) = 0, \quad \sum_{i=1}^{n} \chi\left(\frac{\xi_i - \mu}{\sigma}\right) = 0,$

where $\psi(x) = G_0(x) - 0.5$, $\chi(x) = \psi^2(x) - \frac{1}{12}$.

Therefore,

$$\sum_{i=1}^{n} \left(G_0 \left(\frac{\xi_i - T_n}{s_n} \right) - \frac{1}{2} \right) = 0,$$

$$\sum_{i=1}^{n} \left(\left(G_0 \left(\frac{\xi_i - T_n}{s_n} \right) - \frac{1}{2} \right)^2 - \frac{1}{12} \right) = 0. \tag{1.1}$$

If the solutions T_n and s_n of this system of equations exist, T_n and s_n are called the probability integral transformation (PT)-estimators of the location and the scale parameters, respectively.

2 Main results, numerical algorithm

Theorem 1 *Assume that G_0 is differentiable, strictly monotone increasing and $G_0(0) = 0.5$, then T_n and s_n are well defined, that is, (1.1) has a unique solution with $s_n > 0$.*

Theorem 2 *The two dimensional joint distribution of (T_n, s_n), under the conditions of Theorem 2.1, converges to a normal one:*

$$\sqrt{n}\,((T_n, s_n) - (\mu, \sigma)) \xrightarrow{d} N(0, \Sigma),$$

where the covariance matrix Σ is given by $\Sigma = C^{-1} S [C^{-1}]^T$. The matrices C and S are given by

$$C = \begin{matrix} E\left(\frac{\partial}{\partial\mu}\psi\left(\frac{\xi-\mu}{\sigma}\right)\right) & E\left(\frac{\partial}{\partial\sigma}\psi\left(\frac{\xi-\mu}{\sigma}\right)\right) \\ E\left(\frac{\partial}{\partial\mu}\chi\left(\frac{\xi-\mu}{\sigma}\right)\right) & E\left(\frac{\partial}{\partial\sigma}\chi\left(\frac{\xi-\mu}{\sigma}\right)\right) \end{matrix},$$

and

$$S = \begin{matrix} E(\psi^2(\eta)) & E(\psi(\eta)\chi(\eta)) \\ E(\psi(\eta)\chi(\eta)) & E(\chi^2(\eta)) \end{matrix} = \begin{matrix} \frac{1}{12} & 0 \\ 0 & \frac{1}{180} \end{matrix},$$

where $\eta \sim G_0$.

Theorem 3 *The (PT)-estimators, under the conditions of Theorem 2.1, are B-robust, V-robust, qualitatively robust and their breakdown points*

$$\epsilon^*(T_n) = \frac{\delta}{1+\delta} = 0.5, \quad \text{where} \quad \delta = \min\left\{-\frac{\psi(-\infty)}{\psi(+\infty)}, -\frac{\psi(+\infty)}{\psi(-\infty)}\right\},$$

and

$$\epsilon^*(s_n) = \frac{-\chi(0)}{\chi(-\infty) - \chi(0)} = \frac{1}{3},$$

We propose an algorithm for estimating the location and the scale simultaneously. Let the general system be

$$\sum_{i=1}^{n} \psi\left(\frac{\xi_i - T}{s}\right) = 0, \quad \sum_{i=1}^{n} \chi\left(\frac{\xi_i - T}{s}\right) = 0,$$

where the functions ψ and χ are given by (1.1).
Step 1: Preestimation of location and scale by median (med) and median absolute deviation (MAD), i.e.,

$$T_n{}^{(0)} = med\{\xi_i\} \quad \text{and} \quad s_n{}^{(0)} = MAD\{\xi_i\}.$$

Step 2: Estimation of location by

$$T_n{}^{(m+1)} = T_n{}^{(m)} + \frac{s_n^{(m)} \sum_{i=1}^{n} \psi\left(\dfrac{\xi_i - T_n{}^{(m)}}{s_n^{(m)}}\right)}{n}.$$

Step 3: Estimation of scale by

$$[s_n{}^{(m+1)}]^2 = \frac{12}{(n-1)} \sum_{i=1}^{n} {\psi_b}^2\left(\frac{\xi_i - T_n{}^{(m+1)}}{s_n{}^{(m)}}\right)[s_n{}^{(m)}]^2.$$

Step 4: Stop or goto step 2.

3 Estimators of the parameters of *t*-distribution

Consider the standard t-distribution written in general form of the density function

$$f_a(x) = \frac{\Gamma\left(\dfrac{a+1}{2}\right)}{\sqrt{a\pi}\,\Gamma\left(\dfrac{a}{2}\right)} \left(1 + \frac{x^2}{a}\right)^{-\frac{a+1}{2}},$$

where the shape parameter a is a positive real number and denote its probability distribution function $F_a(x)$.

PROBLEM: Given a sample $\xi_1, \xi_2, \ldots, \xi_n$ with common probability distribution function $F_a\left(\frac{x-T}{s}\right)$. Estimate the parameters T, s, a by the sample.

The statistical analysis of classical methods (maximum likelihood estimation, method of moments) was performed by Cramer. Since the fourth moment does not exist for $a \leq 4$, these methods lead to an estimate of a which is greater than 4. This is a subtantial defect. Moreover, the variance of the estimator of a is infinite if $a \leq 8$. Because our major interest is in small a we turn to the use of robust estimators.

Let

$$\beta_1(a) = \int_{-\infty}^{+\infty} (F_1(x) - 0.5)^2 dF_a(x),$$

$$\beta_\infty(a) = \int_{-\infty}^{+\infty} (\Phi(x) - 0.5)^2 dF_a(x),$$

where $\Phi(x)$ is the standard normal distribution function.

By our theory of M-estimators we can give the following systems of equations

$$\sum_{i=1}^{n} F_1\left(\frac{\xi_i - T_1}{s_1(a)}\right) = \frac{n}{2},$$

$$\sum_{i=1}^{n} \left(F_1\left(\frac{\xi_i - T_1}{s_1(a)}\right) - 0.5\right)^2 = n\beta_1(a),$$

$$\sum_{i=1}^{n} \Phi\left(\frac{\xi_i - T_\infty}{s_\infty(a)}\right) = \frac{n}{2},$$

$$\sum_{i=1}^{n} \left(\Phi\left(\frac{\xi_i - T_\infty}{s_\infty(a)}\right) - 0.5\right)^2 = n\beta_\infty(a).$$

The function $d(a) = s_1(a) - s_\infty(a)$ is stricly monotone increasing and $d(a) = 0$ iff a is the shape parameter of the distribution of the sample elements.

We can give an iterative algorithm for estimators of parameters T, s, a with the cut-and-try method and the above mentioned recursive algorithm.

Acknowledgments: This research has benn supported by TAMOP- 4.2.1.B-10/2/KONV-2010-0001 project with support by the European Union, cofinance by the European Social Fund.

References

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust statistics, the approach based on influence functions*. New York: Wiley.

Fegyverneki, S. (2004). Robust estimators and probability integral transformation *Math. Comput. Modelling*, **38**, 803-814.

Huber, P.J. (1981). *Robust statistics*. New York: Wiley.

Empirical Estimation for the Stationary Distribution of a Discrete-Time Semi-Markov Process

Stylianos Georgiadis[1], Nikolaos Limnios[1]

[1] Université de Technologie de Compiègne, Laboratoire de Mathématiques Appliquées de Compiègne, Centre de Recherches de Royallieu, BP 20529, 60205 Compiègne, Cedex, France

Abstract: We consider a discrete-time semi-Markov process with finite state space and an observation censored at a fixed time. Some results on the nonparametric empirical estimation of the stationary distribution of the embedded Markov chain and the mean sojourn times are given. We propose two empirical estimators for the stationary distribution of the semi-Markov process and study their asymptotic properties, as strong consistency and asymptotic normality, as the length of the observation tends to infinity. Finally, a numerical application is presented to illustrate the comparison of the two estimators.

Keywords: Semi-Markov Chains; Nonparametric Estimation; Stationary Distribution; Empirical Estimators.

1 Introduction and Preliminaries

Semi-Markov chains are a class of discrete-time stochastic processes that generalize Markov chains and discrete-time renewal processes. These processes are very useful in stochastic modeling. A study on the semi-Markov chains is given toward applications by Barbu and Limnios (2008).
Let $\boldsymbol{Z} := (Z_k)_{k\in\mathbb{N}}$ be a semi-Markov chain with finite state space E. Let $(\boldsymbol{J}, \boldsymbol{S}) := (J_n, S_n)_{n\in\mathbb{N}}$ be the corresponding homogeneous Markov renewal chain, where $\boldsymbol{J} := (J_n)_{n\in\mathbb{N}}$ is the embedded Markov chain (EMC) of the successive visited states with state space E and $\boldsymbol{S} := (S_n)_{n\in\mathbb{N}}$ are the jump times with values in $\mathbb{N}$. Also, we denote by $X_n := S_n - S_{n-1}$, $n \in \mathbb{N}^*$, the sojourn times in these states with values in $\mathbb{N}$. Let us now define the jump counting process of the jump times $\boldsymbol{S}$, as $N(k) = \max\{n \geq 0 : S_n \leq k\}$. That is, the semi-Markov chain is defined as

$$Z_k := J_{N(k)}, \quad k \in \mathbb{N}.$$

We denote by $\boldsymbol{\nu} = (\nu_i; i \in E)$ the stationary distribution of the EMC $\boldsymbol{J}$ and by $\boldsymbol{m} := (m_i; i \in E)$ the mean sojourn times of $\boldsymbol{Z}$ in each state. The

stationary distribution $\boldsymbol{\pi} = (\pi_i; i \in E)$ of the SMC $\boldsymbol{Z}$ is given by

$$\pi_i := \frac{\nu_i m_i}{\sum_{k \in E} \nu_k m_k}.$$

From now on, we assume that the EMC $\boldsymbol{J}$ is irreducible with finite mean sojourn times.
Let $\boldsymbol{U} := (U_k)_{k \in \mathbb{N}}$ be the sequence of the backward recurrence times, where $U_k := k - S_{N(k)}$. The coupled process $(\boldsymbol{Z}, \boldsymbol{U}) := (Z_k, U_k)_{k \in \mathbb{N}}$ is a Markov chain with values in $E \times \mathbb{N}$. In our case, where $S_0 = 0$, we get that $U_0 = 0$. Let $\boldsymbol{R}_0^\sharp := (R^\sharp(i,u)(j,v); (i,u)(j,v) \in (E \times \mathbb{N}))$ be the fundamental matrix of $(\boldsymbol{Z}, \boldsymbol{U})$ defined as

$$\boldsymbol{R}_0^\sharp := (\boldsymbol{\Pi}^\sharp - \boldsymbol{P}^\sharp + \boldsymbol{I}^\sharp)^{-1} - \boldsymbol{\Pi}^\sharp,$$

where $\boldsymbol{\pi}^\sharp := (\pi^\sharp(i,u); (i,u) \in E \times \mathbb{N})$ and $\boldsymbol{P}^\sharp := (P^\sharp(i,u)(j,v); (i,u),(j,v) \in E \times \mathbb{N})$ are the stationary distribution and the transition kernel of $(\boldsymbol{Z}, \boldsymbol{U})$, respectively, $\boldsymbol{\Pi}^\sharp = \mathbf{1}^\top \boldsymbol{\pi}^\sharp$ is the limiting matrix of the sequence $(P^\sharp)^n$, $n \in \mathbb{N}$, $\mathbf{1}$ a column-array with all entries equal to 1, and $\boldsymbol{I}^\sharp$ equals to 1, if $i = j$ and $u = v$, and 0 otherwise.

2 Nonparametric Empirical Estimation

We observe a SMC in the interval $[0, M]$, where $M \in \mathbb{N}^*$ a fixed censoring time. The observation of the SMC $\boldsymbol{Z}$ censored at time $M \in \mathbb{N}^*$ is defined

$$\mathcal{H}_M := \{Z_u;\, 0 \le u \le M\} := \{J_0, X_1, J_1, \ldots, X_{N(M)}, J_{N(M)}, U_M\},$$

where $U_M := M - S_{N(M)}$.
First, the empirical estimators $\hat{\boldsymbol{\nu}}(M) := (\hat{\nu}_i(M);\, i \in E)$ and $\hat{\boldsymbol{m}}(M) := (\hat{m}_i(M);\, i \in E)$, $M \in \mathbb{N}^*$, for the stationary distribution $\boldsymbol{\nu}$ of $\boldsymbol{J}$ and the mean sojourn times are given as

$$\hat{\nu}_i(M) := \frac{N_i(M)}{N(M)} \quad \text{and} \quad \hat{m}_i(M) = \frac{1}{N_i(M)} \sum_{r=1}^{N_i(M)} X_{i,r},$$

where $N_i(k)$ is the number of visits of $\boldsymbol{Z}$ to state $i \in E$ up to time k, and $X_{i,r}$ the r-th sojourn time in state $i \in E$. Both estimators are proved to be strongly consistent and asymptotically normally distributed, as M tends to infinity.
Now, we propose two nonparametric empirical estimators for the stationary distribution of a semi-Markov chain and give their asymptotic properties. The empirical estimators $\tilde{\boldsymbol{\pi}}(M) := (\tilde{\pi}_i(M);\, i \in E)$ and $\hat{\pi}(M) := (\hat{\boldsymbol{\pi}}_i(M);\, i \in E)$ of the stationary distribution $\boldsymbol{\pi}$ of $\boldsymbol{Z}$ are defined as follows

$$\tilde{\pi}_i(M) := \frac{1}{M} \sum_{k=1}^{M} \mathbf{1}_{\{Z_{k-1} = i\}}, \tag{1}$$

$$\hat{\pi}_i(M) := \frac{\hat{\nu}_i(M)\hat{m}_i(M)}{\sum_{k\in E}\hat{\nu}_k(M)\hat{m}_k(M)}. \tag{2}$$

Limnios et al. (2005) have studied the properties of the second estimator in continuous time case. Our main results follow.

Theorem 1 *The proposed estimator (1) of the stationary distribution for a SMC satisfies the following properties :*

1. *Strong consistency*

$$\max_i |\tilde{\pi}_i(M) - \pi_i| \xrightarrow{a.s.} 0, \quad M \to \infty.$$

2. *Asymptotic normality*

$$\sqrt{M}(\tilde{\pi}_i(M) - \pi_i) \xrightarrow{D} \mathcal{N}(0, \sigma^2_{\pi_i}), \quad M \to \infty,$$

where

$$\sigma^2_{\pi_i} = 2\sum_{u,v\geq 0} \pi^\sharp(i,u)R^\sharp(i,u)(i,v) - \pi_i(1-\pi_i).$$

Theorem 2 *The proposed estimator (2) of the stationary distribution for a SMC satisfies the following properties :*

1. *Strong consistency*

$$\max_i |\hat{\pi}_i(M) - \pi_i| \xrightarrow{a.s.} 0, \quad M \to \infty.$$

2. *Asymptotic normality*

$$\sqrt{M}(\hat{\pi}_i(M) - \pi_i) \xrightarrow{D} \mathcal{N}(0, \sigma^2_{\pi_i}), \quad M \to \infty,$$

with $\sigma^2_{\pi_i}$ as given in Theorem 1.

Finally, a numerical application on a three state semi-Markov system is presented.

References

Barbu, V. S. and Limnios, N. (2008) *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications*. New York: Springer.

Limnios, N., Ouhbi, B. and Sadek, A. (2005) Empirical estimator of stationary distribution for semi-Markov processes. *Communications in Statistics - Theory and Methods*, **34**(4), 987-995.

Regression models prediction of air pollution concentrations in Moscow city

Evgeny Ivin [1]

[1] Moscow School of Economics, Lomonosov Moscow State University

E-mail for correspondence: evg.ivin@gmail.com

Abstract: Nowadays there is a great collection of scientific evidence about air pollution exposure assessment. Nevertheless, in a global community, little is known about Moscow and Russian Federation in general. This information bias occurs because the absolute majority of relevant studies have been published in Russian language. This research presents several air pollution modeling methods in application to Moscow data. Those methods are: kriging, land-use regression and machine learning methods.

Keywords: Air pollution; Land-use regression; Kriging; Machine learning.

1 Description of the study

Air pollution is an important problem for Moscow city and its surroundings. Moscow is a busy city; it counts about 12 million inhabitants, and almost 4 million vehicles are registered in the area. Traffic is the main source of air pollution, and thus Moscow has a strong necessity in monitoring, evaluation and reduction of contamination burden.

This research presents several exposure assessment models. One of them is geostatistical interpolation method, that is kriging. Simple and ordinary kriging has been performed for GIS monitoring data. A land-user regression model has also been developed, taking into consideration traffic density and other land-use information.

Finally, a newly developed machine learning method called conformal predictors has been used (see Vovk and Gammerman and Shafer (2005)). By its definition, a conformal predictor always provides valid estimates with confidence, where the level of confidence is chosen. Such a predictor is very flexible, because it can be constructed on the basis of almost any machine learning algorithm. In this research, two regression conformal predictors have been derived: one of them - on the basis of the interpolation model, and another one - on the basis of the land-use model. Overall modeling results have been compared and discussed.

All the data for this study has been kindly provided by “Mosecomonitoring,” an environmental unit of the Government of Moscow.

References

Gilbert, N.L., et al. (2005). Assessing Spatial Variability of Ambient Nitrogen Dioxide in Montreal, Canada, with a Land-Use Regression Model. *Journal of the Air & Waste Management Association*, **55**, 10591063.

Jerrett, M., et al. (2005). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology*, **15**, 185-204.

Vovk, V., Gammerman, A., Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.

Statistical and econometric practicum at the Moscow School of Economics

E. Ivin[1], A. Kurbatskiy[1]

[1] Moscow School of Economics, Lomonosov Moscow State University

E-mail for correspondence: evg.ivin@gmail.com

Keywords: Air pollution; Land-use regression; Kriging; Machine learning.

Moscow School of Economics (MSE) is a joint project of Russian Academy of Science (RAS) and Lomonosov Moscow State University (MSU), launched in order to train and educate researchers who would further analyze and solve the problems of Russia's economy. The head of the School is Prof. Nekipelov, A. D., Academician of RAS. MSE works as follows. The bachelors are taught in one stream (80 persons, 4 years), while the masters (50-60 persons, 2 years) are divided into two programs: "National and World's Economy" and "Financial Strategies". Moreover, there is also a doctorate program called "Mathematical Methods in Economy." General supervision of these programs is realized by Prof. Polterovich, V.M., Academician of RAS. The program "National and World's Economy" is carried out by the chair of Economical Theory. The head of the chair is Prof. Glinkina, S.P., and the deputy director is Prof. Golovnin, M.Yu., both are deputy directors of the Institute of Economy of RAS. The program named "Financial strategies" is run by the chair with the same name. The head of the chair is Prof. Kvint, V.L., Academician of RAS, and the deputy head is Prof. Alimuradov, M.K. Moscow School of Economics assigns a great share of its educational programs to mathematics. All the mathematical courses in bachelor, master and doctorate programs are executed by the chair of Econometrics and Mathematical Methods in Economy. This chair is directed by Prof. Aivazian, S.A., deputy director of CEMI RAS, and the deputy head of the chair is Dr. Ivin, E.A.

When it comes to mathematical disciplines, the major accent is put onto probability theory, statistics, econometrics and time series analysis. All of these courses are accompanied by practical classes with computers. Two practical courses are defined: "Data analysis with computers" (probability theory and statistics) and "Practical work in econometrics and time series analysis." These courses are independent, because they not only accompany the theoretical ones, but they are also aimed to solve unique problems. These tasks are directed to educate students how to solve economical

and socio-economical problems in classical settings, and also to help students cope with their own final projects. We will provide more information regarding these specific tasks in our presentation, and here we will just shortly mention their global features.

Within the framework of these courses, we study basic distribution laws and the beginnings of the estimation theory, together with the analysis of one or two sets, contingency tables, nonparametric methods, and beginnings of factor and dispersion analysis. Moreover, we teach various types of regression analysis and prediction, including the choice of the appropriate model and hypothesis testing. Then, we continue with the problems of binary and multiple choice, systems of regression equations, systems of simultaneous equations, dynamic models, classical time series problems, distributed lags models, cointegration, ARCH and GARCH models, and classical and modern models of financial markets.

As for mathematical software, we make use of Excel, EViews and Stata. The data are taken from the state statistical databases, both Russian and international. We also use regional data and the data from big enterprises. Recently, due to broadening of collaboration between MSE and European universities, we have become very interested in R statistical software. Thus, we will be very thankful if your European colleagues could share their scientific knowledge and practical experience with us.

Kernel methods for geostatistics and their implementation with R

Olga Ivina[1]

[1] University of Girona, Spain

E-mail for correspondence: olga.ivina@udg.edu

Abstract: A covariance function is the cornerstone geostatistical modeling tool. It is well-known that covariance functions are similar to kernels in signal processing. In geostatistics, a vast variety of models are developed under assumption that a spatial process is a) Gaussian, b) second-order stationary. Thus the covariance functions used for geostatistical modeling are functions only of distance between spatial points. This research takes up a classical geostatistical technique, that is ordinary kriging, and opposes it to a newly developed machine learning method, that is ridge regression confidence machine. In the latter, kernels are introduced via "kernel trick" technique. These kernels are of the same analytical form as the corresponding covariance functions. Barcelona air pollution data has been used, and R statistical software has been applied for practical computation.

Keywords: Kriging; Covariance function; Conformal predictor; Kernel trick.

1 Introduction

Covariance function in geostatistics is aimed to model spatial variability of a factor of interest. As being tightly binder with variograms, covariance functions are used in practice to approach empirical variograms for the data with a proper variomodel. Also, they are used for prediction geostatistical models as they help consider spatial dependence between observations. One of the most widely used geostatistical methods, ordinary kriging, uses covariance functions to consider spatial features of data distribution.
Kernel trick is aimed to deal with high-dimensional, and thus computationally difficult, problems. This method consists in mapping the initial input space into another Euclidian space, called *feature space*, where the modeling is performed. The mapping is performed with the use of kernel functions. A kernel function is used to express the dot product of the vectors of feature space in terms of the input space. This approach has been first suggested by V. Vapnik for support vector machines, but it has later been implemented for other model types.
A recently developed method of conformal predictors can be successfully used in geostatistics. A conformal predictor allows to obtain predictions

with high confidence, and it is always valid. Its main strength is that it can be developed upon almost any (not necessarily) regression algorithm, as inheriting the predictive power of the latter, but always providing valid predictions with high confidence. In application to geostatistics, a conformal prediction can be derived on top of classical ordinary kriging. An existing regression conformal predictor called *ridge regression confidence machine* (RRCM) (Vovk et. al (2005)) has been used in this research and adjusted to meet kriging specification. It is noteworthy that this RRCM specification does not require that data follows any specific distribution, apart from being iid.
Kernels can be used for this newly derived confidence interpolation predictor in the same manner as the covariance functions are used in ordinary kriging. The space in geostatistical problems is not high-dimensional, as it is based on only two parameters: longitude and latitude. Nevertheless, (non-linear) variomodels help introduce spatial dependence between observations and thus more precisely adjust data distribution when fitting prediction models. In the same way, (non-linear) kernels can be implemented in regression conformal predictors, such as RRCM.

2 An example

To demonstrate, how kernels match widely used isotropic covariance functions, the following example is provided. Gaussian covariance (correlation) function is represented as:

$$C(h) = \exp\left(-\left(\frac{h}{\sigma}\right)^2\right), \tag{1}$$

where h is a distance between points and σ is a scale parameter. It is a direct analogue of the Gassian radial basis function, otherwise known as RBF-kernel (Schölkopf and Smola (2002)):

$$\mathcal{K}(x^{(1)}, x^{(2)}) = \exp -\frac{||x^{(1)} - x^{(2)}||^2}{2a^2}, \tag{2}$$

where a is a scale parameter. Other covariance function can be used as kernels for a geostatistical conformal predictor, too.

3 Practical computation

Apart from Gaussian, the following covariance functions have been taken up and approached by kernels: exponential, spherical and Maérn. Barcelona air pollution data has been used. This dataset has been kindly provided by XVPCA of the Generalitat of Catalonia. R statistical software has been

employed for computations. Kriging models have been fitted with the *geoR* package (Ribeiro and Diggle (2001)), and RRCM models have been implemented with the derivations of the *PredictiveRegression* package (Vovk et. al (2009)).

Acknowledgments: Many thanks to Ilia Nouretdinov, PhD, and Prof. Alex Gammerman, PhD, from the Royal Holloway University of London for their incredibly valuable counseling on conformal predictors and, generally, machine learning methods.

References

Diggle, P.J., Ribeiro, P.J. (2007). *Model-based geostatistics*. Springer.

Ribeiro P.J., Diggle, P. J. (2001). geoR: a package for geostatistical analysis. In: *R-NEWS*, **1**, 2, 15-18.

Schölkopf, B., Smola, A.J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press.

Vovk, V., Gammerman, A., Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.

Vovk, V. and Nouretdinov, I. and Gammerman, A. (2009). On-Line Predictive Linear Regression. *Annals of Statistics*, **37**, 3, 1566-1590.

The Markov processes under the Kendall convolution and heavy tailed distributions

Barbara H. Jasiulis-Gołdyn[1]

[1] Institute of Mathematics, University of Wrocław, Wrocław, Poland

E-mail for correspondence: Barbara.Jasiulis@math.uni.wroc.pl

Abstract: We give here a construction of discrete time Markov processes based on the Kendall generalized convolutions. For the discrete step distribution such processes can be considered as random walks with respect to these generalized convolutions. We obtain new classes of heavy tailed distributions.

Keywords: generalized convolution; heavy tailed distribution; Markov process; random walks; weakly stable distribution.

1 Weakly stable distribution and generalized convolution

There exists a very interesting and not completely characterized class of weakly stable probability distributions $\mu \in \mathcal{P}$ with the following property:

$$\forall \quad a, b \in \mathbb{R} \quad \exists \lambda \in \mathcal{P} \quad T_a\mu * T_b\mu = \mu \circ \lambda$$

where $T_a\mu(A) = \mu(A/a)$ for every Borel set A when $a \neq 0$, $T_0\mu = \delta_0$, $\circ$ denotes multiplicative convolution and $\star$ deneotes classical convolution (corresponding to the sum of two independent random elements). The measure μ generates a binary operation $\otimes_\mu$ called a weak generalized convolution. One can prove that equivalently, μ is weakly stable if

$$\forall \quad \lambda_1, \lambda_2 \in \mathcal{P} \quad \exists \lambda \in \mathcal{P} \quad \mu \circ \lambda_1 * \mu \circ \lambda_2 = \mu \circ \lambda.$$

The most known examples are symmetric α -stable distributions ($\alpha \in (0, 2]$) and uniform distribution on the unit sphere in $\mathbb{R}^n$. ,

2 Random walks under the Kendall convolution

We construct discrete time Markov processes based on the weak generalized convolution, i.e. such random walks that theirs increments are independent and instead of summation of unit steps we take their cumulation in the weak stability sense. Theorem about asymptotical properties such objects

will be showed. As an example of constructed processes we present random walk under the Kendall convolution having the following probability kernel:

$$\delta_x \otimes_{\mu_\alpha} \delta_1 = |x|^\alpha \pi_{2\alpha} + (1 - |x|^\alpha)\delta_1,$$

for $x \in [-1, 1]$ and $\alpha \in (0, 1]$ where $\pi_{2\alpha}$ is the Pareto distribution with the density $2\alpha y^{-(2\alpha+1)} I_{[1,\infty)}(y)$. Considering the random walks under the Kendall convolution we obtain new classes of heavy tailed distributions containing the Pareto distribution $\pi_{2\alpha}$. We present basic properties of constructed random walks.

References

Jasiulis-Gołdyn, B.H. (2011). *On the random walk under the Kendall convolution*, submitted.

Jasiulis-Gołdyn, B.H., Kula, A. (2012). *The Urbanik generalized convolutions in the non-commutative probability and a forgotten method of constructing generalized convolution*, to appear in Proceedings of the Indian Academy of Science - Math. Sc.

Jasiulis-Gołdyn, B.H., Misiewicz, J.K. (2011). On the uniqueness of the Kendall generalized convolution. *Journ. of Theor. Probab.*, **24(3)**, 746–755.

Kingman, J.F.C. (1963). Random Walks with Spherical Symmetry. *Acta Math.*, **109(1)**, 11–53.

McNeil, A.J., Nešlehová, J. (2009). Multivariate Archimedean Copulas, $d-$ monotone Functions and l_1- norm Symmetric Distributions. *Ann. Statist.*, **37(5B)**, 3059–3097.

Misiewicz, J.K., Oleszkiewicz, K. and Urbanik, K. (2005). Classes of measures closed under mixing and convolution. Weak stability. *Studia Math.*, **167(3)**, 195–213.

Urbanik, K. Generalized convolutions I-V. *Studia Math.*,**23**(1964), 217–245, **45**(1973), 57–70, **80**(1984), 167–189, **83**(1986), 57–95, **91**(1988), 153–178.

Joint asymptotic normality of kernel type density estimator for spatial observations

Zsolt Karácsony[1], István Fazekas[2], Renáta Vas[2]

[1] Department of Applied Mathematics, University of Miskolc, H-3515 Miskolc-Egyetemváros, Hungary
[2] Faculty of Informatics, University of Debrecen, H-4032 Debrecen, Kassai út 26, Hungary

E-mail for correspondence: matkzs@uni-miskolc.hu

Abstract: The Central Limit Theorem is considered for m-dependent random fields. The random field is observed in a sequence of irregular domains. The sequence of domains is increasing and at the same time the locations of the observations become more and more dense in the domains. The Central Limit Theorem is applied to prove asymptotic normality of kernel type density estimators. It turns out that the covariance structure of the limiting normal distribution can be a combination of those of the continuous parametric and the discrete parametric results. Numerical evidence is presented.

Keywords: Asymptotic normality; central limit theorem; random field; kernel.

1 CLT for stationary random fields

Consider a domain D in R^d. We observe a random field $\xi(\cdot)$ in certain points of the domain D and we assume the following setup. Suppose that the random field $\xi(\cdot)$ is observed at finitely many locations i.e. at the elements $\mathbf{s}_{n1}, \ldots, \mathbf{s}_{nn} \in D_n$ lying in the sampling region $D_n \subset D$. We shall use the notion of the mixed (or nearly infill or infill-increasing) domain sampling which means that the sampling region D_n increases and at the same time, the data sites $\{\mathbf{s}_{n1}, \ldots, \mathbf{s}_{nn}\}$ fill in any given sub-region of D_n increasingly densely as $n \to \infty$. This approach was studied e.g. by Lahiri (1999), Fazekas and Chuprunov (2006) and Park, Kim, Park and Hwang (2009) (see also Karácsony and Filzmoser (2010)). It can be useful in geostatistics, environmental sciences etc.
Let $\xi(\cdot)$ be an m-dependent field which means that m is the infimum of the numbers denoted by b such that if $\|\mathbf{s}_1 - \mathbf{s}_2\| > b$ then $\xi(\mathbf{s}_1)$ and $\xi(\mathbf{s}_2)$ are independent. Let $I_{m,n}(\mathbf{u}) = \{\mathbf{s} \in D_n : \|\mathbf{s} - \mathbf{u}\| \leq m\}$ and $\kappa_n = \max_u \sharp\{I_{m,n}(\mathbf{u})\}$. So κ_n denotes the number of elements of the set $I_{m,n}(\mathbf{u})$ with maximal cardinality. We assume that $\kappa_n > 0$ holds for each n. We suppose that the measure κ_n of density satisfies $\kappa_n \sim n^a$ with a constant $0 < a < 1$.

Introduce the notations $\xi_i = \xi_n(\mathbf{s}_{ni})$, $S_n = \sum_{i=1}^n \xi_n(\mathbf{s}_{ni}) = \sum_{i=1}^n \xi_i$, $\nu_n = \mathrm{var}(\xi_n(\mathbf{s}))$, $\mathcal{I}_n = \{(i,j) : 0 < \|\mathbf{s}_{ni} - \mathbf{s}_{nj}\| \leq m\}$, $\tau_n = \frac{1}{n\kappa_n} \sum_{(i,j)\in\mathcal{I}_n} \mathrm{cov}(\xi_n(\mathbf{s}_{ni}), \xi_n(\mathbf{s}_{nj}))$. We know that $\mathrm{var}(S_n) = n\nu_n + n\kappa_n\tau_n$ and τ_n can be negative as well.

Theorem 1 *(Park-Kim-Park-Hwang, (2009)) Let $\{\xi_n\}$ be a sequence of strictly stationary random fields on $D \subset R^d$ with $E\xi_n(\mathbf{s}) = 0$. Assume that $\sup_{\mathbf{s}\in D} |\xi_n(\mathbf{s})|$ is bounded with probability one and $E\left|\prod_{j=1}^l \xi_n(\mathbf{s}_{nj}^{'})\right| = \mathrm{O}\left(\nu_n^l\right)$ holds uniformly for all the different points $\mathbf{s}_{nj}^{'} \in \{\mathbf{s}_{n1}, \ldots, \mathbf{s}_{nn}\}$. If $\nu_n + \kappa_n\tau_n \geq \delta\kappa_n\nu_n^2$ for some $\delta > 0$ then we have $\frac{S_n}{\sqrt{var(S_n)}} \Rightarrow \mathcal{N}(0,1)$ in distribution.*

2 Application to density estimation

Let $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ be a strictly stationary m-dependent random field, $D \subseteq R^d$. For each $z \in R$, let $F(z) = P(Z(\mathbf{s}) \leq z)$. We call the function F marginal distribution function. Assume that there exist the appropriate marginal density function f. Suppose that we observe the values of Z at the points $\mathbf{s}_{n1}, \ldots, \mathbf{s}_{nn}$ in D. In this section we study the nonparametric estimation of the marginal density function. Consider the kernel type density estimator $\hat{f}_n(z) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{z - Z(\mathbf{s}_{ni})}{h_n}\right)$. Here K is a kernel.
Let $f_{s_{ni},s_{nj}}$ be the joint density function of $Z(\mathbf{s}_{ni}), Z(\mathbf{s}_{nj})$. Let $z \in R$ be fixed. Consider the following assumptions.

(1) (a) $f(z) > 0$, f is continuous at z,
   (b) $f_{s_{ni},s_{nj}}$ are equicontinuous at (z,z), i.e. if $(z_1,z_2) \to (z,z)$ then $\sup_{i,j} |f_{s_{ni},s_{nj}}(z_1,z_2) - f_{s_{ni},s_{nj}}(z,z)| \to 0$,
   (c) all finite dimensional densities of $Z(\mathbf{s}_{n1}), Z(\mathbf{s}_{n2}), \ldots$ exist and are bounded and continuous,
   (d) if $n \to \infty$ then $\frac{1}{n\kappa_n} \sum_{(i,j)\in\mathcal{T}_n} \{f_{s_{ni},s_{nj}}(z,z) - f(z)^2\} \to \tau$, where τ is a nonnegative constant depending on z,

(2) The kernel K is nonnegative on R and satisfies $\int_R K = 1$; $|z|K(z) \to 0$ as $|z| \to \infty$.

(3) $h_n > 0$ is a sequence satisfying $h_n \to 0$ and $nh_n \to \infty$, as $n \to \infty$.

(4) There exists a constant $\delta > 0$ such that $f(z)\int_R K^2 + \tau\kappa_n h_n \geq \delta\kappa_n h_n$.

Theorem 2 *(Park-Kim-Park-Hwang, (2009)) Let us suppose that the assumptions (1) – (4) hold.*

1. *Then $\left\{n^{-1}h_n^{-1}f(z)\int_R K^2 + n^{-1}\kappa_n\tau\right\}^{-\frac{1}{2}} \{\hat{f}_n(z) - E\hat{f}_n(z)\} \Rightarrow \mathcal{N}(0,1)$.*

2. Suppose that f is twice differentiable in a neighbourhood of z and $\int uK(u)du = 0$. Moreover assume that f'' is continuous and bounded and $nh_n^5 \to 0$, $n\kappa_n^{-1}h_n^4 \to 0$. Then

$$\left\{n^{-1}h_n^{-1}f(z)\int_R K^2 + n^{-1}\kappa_n\tau\right\}^{-\frac{1}{2}}\{\hat{f}_n(z) - f(z)\} \Rightarrow \mathcal{N}(0,1).$$

3 Joint asymptotic normality for the density estimator

Our aim is to study the multidimensional version of Theorem 2, i.e. the joint asymptotic normality of the density estimator. Let $z_1, z_2, \ldots, z_q$ be given distinct real numbers. We assume that

$$\frac{1}{n\kappa_n}\sum_{i,j\in\mathcal{T}_n}\left(f_{s_{ni},s_{nj}}(z_r,z_t) - f(z_r)f(z_t)\right) \to \tau_{rt} \text{ if } n\to\infty.$$

Then $(\widehat{f_n}(z_i) - f(z_i), i = 1,\ldots,q)$ is asymptotically $\mathcal{N}(0,\Sigma)$ with

$$\Sigma = \frac{1}{nh_n}\begin{bmatrix} f(z_1)\int K^2(t)dt + \tau_{11}\kappa_n h_n & \ldots & \tau_{1q}\kappa_n h_n \\ \tau_{21}\kappa_n h_n & \ldots & \tau_{2q}\kappa_n h_n \\ \vdots & \ddots & \vdots \\ \tau_{q1}\kappa_n h_n & \ldots & f(z_q)\int K^2(t)dt + \tau_{qq}\kappa_n h_n \end{bmatrix}.$$

We also present examples that give numerical evidence for the phenomena described in the above proposition.

Acknowledgments: The first author was supported by TAMOP-4.2.1.B-10/2/KONV-2010-0001 project with support by the European Union, co-financed by the European Social Fund.

References

Fazekas, I. and Chuprunov, A. (2004). A central limit theorem for random fields. *Acta Mathematica Academiae Paedagogicae Nyiregyhaziensis*, **20**, (1), 93-104.

Fazekas, I. and Chuprunov, A. (2006). Asymptotic normality of kernel type density estimators for random fields. *Stat. Inf. Stoch. Proc.*, **9**, 161-178.

Karácsony, Zs. and Filzmoser, P. (2010). Asymptotic normality of kernel type regression estimators for random fields. *Journal of Statistical Planning and Inference*, **140**, 872-886.

Lahiri, S.N. (1999). Asymptotic distribution of the empirical spatial cumulative distribution function predictor and prediction bands based on a subsampling method. *Probab. Theory Related Fields*, **114** (1), 55-84.

B.U. Park, T.Y. Kim, T.-S. Park and S.Y. Hwang (2009). Practically Applicable Central Limit Theorem for Spatial Statistics. *Math. Geosci.* **41**, 555-569.

A copula-based method for synthetic microarray data generation

Sergio Lew[1], Jordi Solé-Casals[2], Cesar F. Caiafa[3,4] and Josep Bau-Macià[2]

[1] IIBM-FIUBA, Av. Paseo Colón 850 (1063), Buenos Aires, ARGENTINA
[2] University of Vic, Sagrada Família 7, 08500, Vic, SPAIN
[3] IAR-CONICET, C.C.5, (1894) Villa Elisa, Buenos Aires, ARGENTINA
[4] FIUBA, Av. Paseo Colón 850 (1063), Capital Federal, ARGENTINA

E-mail for correspondence: jordi.sole@uvic.cat

Abstract: In this work, we propose a copula-based method to generate synthetic gene expression data that account for marginal and joint probability distributions features captured from real data. Our method allows us to implant significant genes in the synthetic dataset in a controlled manner, giving the possibility of testing new detection algorithms under more realistic environments.

Keywords: Copulas; gene expression; microarray data.

1 Introduction

Detection of differentially expressed genes in microarray experiments has been subject of great effort in the bioinformatics community. Optimal detection methods allow to reduce the amount of both, pathological and control experiments and, in consequence, time and costs [Dupuy A. and Simon R. M., 2007]. However, most of the developed algorithms have been tested with synthetic data using simple generative models and assuming incorrect hypothesis about variable statistics and their dependence. The proposed method captures the statistical structure of real datasets allowing us to generate new random samples drawn from a copula-based random generator.

2 Materials and Methods

The proposed method is shown in figure 1. Briefly, we fit real microarray data to a t-copula [Nelsen R. B., 1999] and then we generate random gene expression data sharing marginal and high-order dependence with the original data.
Firstly, original gene expressions dataset (Fig. 1.a) are mapped into a unitary hypercube by means of a monotonically increasing function, i.e.

the inverse cumulative distribution function of the marginal distributions. An approximated Maximum Likelihood (ML) method is used for fitting this transformed data to a t-copula. Once the copula parameters are obtained, random samples are generated according to this copula structure (Fig. 1.b). This new dataset, which have uniform marginal distributions, is then mapped to a $\mathcal{N}(0,1)$ marginal Gaussian distributions (Fig. 1.c). It is important to remark that monotonically increasing transformations do not alter high-order dependence measures like Kendall-τ or Spearman-ρ. At this point, significantly expressed genes are introduced in a controlled manner into the data (Fig. 1.d). By means of the inverse transformations used before, we then back-transform the data to the original space, obtaining a synthetic dataset which preserve the same marginal distributions and high-order variable dependence as the real one (Fig. 1.e-f).

3 Results

When added to the synthetic data, significant genes were recovered by the step-down minP adjusted p-values method [Westfall P. H. and Young S. S., 1993] in all the cases. However, its important to ensure that the algorithm is able to minimize the number of false positive (FP) cases. To prove the robustness of the method against FP generation, we compare the results of our method versus synthetic data generated by multivariable gaussian random process with the same covariance matrices of the original data [Carmona-Saez P. et al, 2006]. We ran 30 experiments for both types of synthetic data with no significant genes added, meaning that the significance test should recover (almost) zero genes differentially expressed between pathological and control groups. Due to the sparseness of significant genes in microarray experiments (less than 1%, under and over-expressed genes) the copula captures the distribution of normally expressed genes. In that sense, our synthetic microarray data produces much less FP genes that the ones generated with a multivariable gaussian process having the same covariance matrix (1.5 ± 0.23 vs 24.76 ± 1.04, $p < 0.0001$, $mean \pm s.e.m.$)

4 Conclusions

In this paper, we propose a new copula-based method for synthetic microarray data generation that allows us to control the number of under and over-expressed genes, preserving the original statistical structure of real data. To our knowledge, this is the first work that overcomes the problem of building synthetic data using simple generative models. Experimental results show the robustness of the method and its usefulness helping researchers to develop new and more powerful algorithms for gene filtering and clustering.

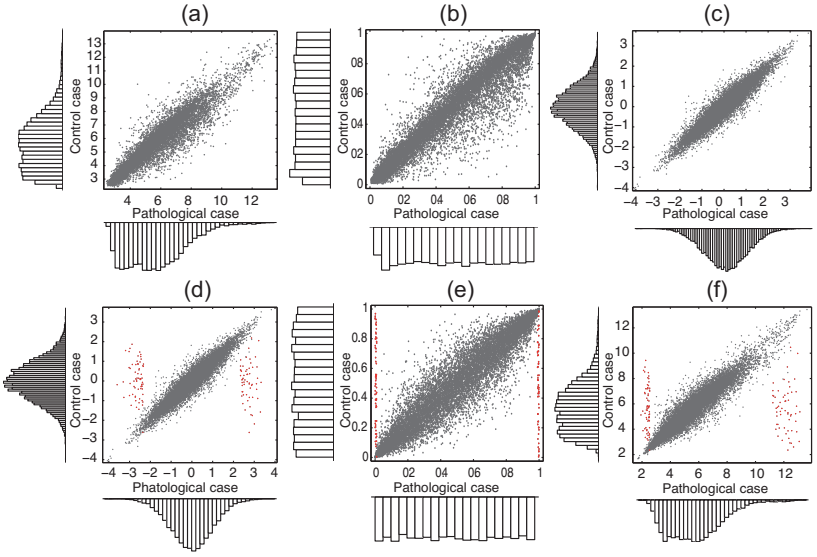

FIGURE 1. Proposed method to generate synthetic data preserving the same marginal distributions and high-order variable dependences of the real data.

Acknowledgments: This work has been in part supported by the MINCYT-MICINN Research Program 2010-2011 (Ref. AR2009-0010) and by the University of Vic under the grants R0904 and R0901.

References

Carmona-Saez P., Pascual-Marqui R. D., Tirado F., Carazo J. M., Pascual-Montano A. (2006). Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization. *BMC Bioinformatics*, **7 (78)**, 1-18.

Dupuy A., Simon R.M. (2007). Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl. Cancer Inst.*, **99 (2)**, 147-157.

Nelsen R. B. (1999). *An Introduction to Copulas*. New York: Springer-Verlag.

Westfall P. H. and Young S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: John Wiley & Sons.

Frequentist or Bayesian Mixed Models? A comparison to provide better estimates of CPUE

Valeria Mamouridis[1], Carmen Cadarso Suarez[2], Germán Aneiros Pérez[3], Francesc Maynou[1]

[1] Institut de Ciències del Mar, CSIC, Psg Marítim de la Barceloneta 37-49, 08003 Barcelona, Spain
[2] Facultade de Medicina e Odontoloxía, R/ de San Francisco, s/n, 15782 Santiago de Compostela, Spain
[3] Facultade de Informática, Campus de Elviña s/n, 15071, A Coruña, Spain

E-mail for correspondence: mamouridis@icm.csic.es

Abstract: Generalized Additive Mixed Models were used to make up a regression analysis applied to red shrimp CPUE. Mixed Models are required when units are not repeatable, such as the case of fishing boats. We also compared the methodologies actually in use: the frequentist REML, the full Bayesian by MCMC tecniques and the empirical Bayesian methods.

Keywords: GAMM; Fisheries; CPUE; REML; MCMC.

1 Introduction

In fisheries research, regression models are often used to analyze CPUE (Catch per Unit Effort), an index of relative abundance of an exploited species. To date CPUE has been mainly analyzed through Generalized Linear Models (GLM), e.g. Goñi et al. (1999), and rarely Generalized Additive Models (GAM) e.g. Damalas et al. (2007). Instead, fishery data usually hold a random nature, being associated to fishing vessels, unrepeatable units. That is not conteplated in such kind of models. In very few occasions random effects has been considered, e.g. using Generalizad Linear Mixed Models (GLMM), Cooper et al. (2004).
On the other side, recent advancements in regression methodologies provide many estimators of random effects in a Generalized Additive Mixed model (GAMM) framework using frequentist (Lin and Zhang, 1999) or Bayesian (Fahrmeir and Lang, 2001) inference.
In this work we present a regression analysis of red shrimp (*Aristeus antennatus*) CPUE from the port of Barcelona (Spain). The last update of red shrimp CPUE modeling in the NW Mediterranean Sea was implementing GLM (Maynou et al, 2003).

The main porpuses of this stady were: 1. improve red shrimp CPUE modeling and 2. compare the Frequentist REstricted Maximum Likelihood, REML (FR), the Empirical Bayesian version of REML (EB) and the Full Bayesian MCMC simulation (FB).

2 Methodology

The frequentist REML was used, implementing the R-package mgcv (Wood, 2006), to obtain the final model, that is, after checking assumptions, the one with the highest deviance explained (DE%). Predictors were selected by a stepwise forward procedure and 2-nd order P-spline was used as smoother., Afterwards, the final model was fitted using the two Bayesian inferences as well, with the impementation of *BayesX* software (Brezger et al., 2005). $J = 21$ subsets, excluding *vessel* j, were used to estimate the model by each method. Then they were compared using the mean square error of predictions, MSEP, calculated on predictions from subset $\{J - j\}$ on subset $\{j\}$. Variables implemented in the model are reported in Table 1.

TABLE 1. Variables used in the study.

Name	Description
cpue	monthly CPUE for each vessel, $i = 1, \ldots, 2314$
time	months from 01-1992 to 12-2008, $t = 1, \ldots, 204$
vessel	a numeric code assigned to each vessel, $j = 1, \ldots, 21$
trips	number of trips performed monthly by each vessel, j during month t
grt	Gross Registered Tonnage of vessels
*nao*3	NAO index of 3 years before the observed *cpue*
period	season variable with 2 levels, $p2$: Jun and Nov; $p1$: otherwise

3 Results

The selected final model belongs to the class of GAMMs:

$$ln(cpue) = \alpha + \beta grt + f(time) + g(trips) + h(nao3) + \gamma p2 + \sum_{j=1}^{J} b_j vessel + \epsilon \tag{1}$$

where $\epsilon \in Gamma(a, b)$.
The partial effects of model 1 are visualized in Figure 1. Effort predictor (*trips*) is the most important sources of variability. The NAO (North Atlantic Oscillation) index is to date the only environmental predictor avail-

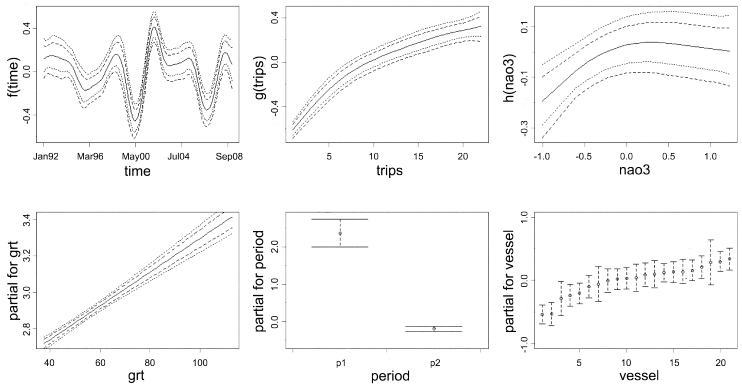

FIGURE 1. Partial effects of model 1 estimated through EB method.

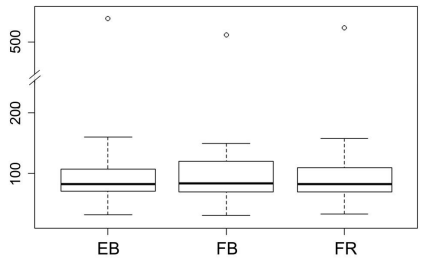

FIGURE 2. Box plot of MSEP estimated for the different methods.

able for deep sea fisheries. Random effects allow to predict for unknown boat effects. Smooth functions increase the explanatory power of the model. Figure 2 shows that there is no difference in predictions between methods, however the EB gives lowest MSEP in almost the 50% of subsets.

4 Conclusion

That study update the red shrimp CPUE modeling through the implementation of effort and environmental predictors and of smooth functions. It also demonstrate that there is no difference in predictions between methods. The use of mixed models permits to infer on the entire population however when units, boats in this case, are not repeatable.

Acknowledgments: Authors thank Dr. T. Kneib for his hints on BayesX usage.

References

Brezger, A., Kneib, T., Lang, S. (2005). BayesX : Analyzing Bayesian Structured Additive Regression Models. *Journal of Statistical Software*, **14** (11), 1-22.

Cooper, A.B., Rosenberg, A.A., Stefánsson, G., Mangel, M. (2004). Examining the importance of consistency in multi-vessel trawl survey design based on the U.S. west coast groundfish bottom trawl survey. *Fisheries Research*, **70**: 239-250.

Damalas,D., Megalofonou, P. and Apostolopoulou, M. (2007). Environmental, spatial, temporal and operational effects on swordfish (Xiphias gladius) catch rates of eastern Mediterranean Sea longline fisheries. *Fisheries Research*, **84**, 233-246.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society C*, **50**, 201-220.

Goñi, R., Alvarez, F. and Adlerstein, S. (1999). Application of generalized linear modeling to catch rate analysis of Western Mediterranean fisheries: the Castellón trawl fleet as a case study. *Fisheries Research*, **42**, 291-302.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models using smoothing splines. *Journal of the Royal Statistical Society, Series B*, **61**, 381-400.

Maynou, F., Demestre, M. and Sánchez, P. (2004). Analysis of catch per unit effort by multivariate analysis and generalised linear models for deep-water crustacean fisheries off Barcelona (NW Mediterranean). *Fisheries Research*, **65**, 257-269.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. CRC Chapman & Hall, Boca Raton, Florida.

Dynamic simulations of food webs with R

Valeria Mamouridis[1], Laurine Burdorf[2], Karline Soetaert[2]

[1] Institut de Ciències del Mar, CSIC, Psg Marítim de la Barceloneta 37-49, 08003 Barcelona, Spain
[2] NIOZ (Royal Netherlands Institute for Sea Research) Yerseke, Korringaweg 7, P.O. Box 140, 4400 AC Yerseke, The Netherlands

E-mail for correspondence: `mamouridis@icm.csic.es`

Abstract: We present a methodology to create and dynamically simulate food webs in the open source software R. This is done in three steps. First a plausible binary food web is generated with a preset number of species (S) and links (L). Then a quantified steady-state foodweb is generated using linear inverse modeling (LIM) techniques. Thirdly, the food web flows are converted into dynamic formulations. The flexibility of this methodology allows to study the stability of these webs and how they react when perturbed.

Keywords: food webs; linear inverse models; dynamic models.

1 Introduction

Food webs describe who eats whom in an ecosystem. For a given number of species S, and links L, a food web can be represented by an $S \times S$ matrix $\mathbf{S}$, where if the species i is a prey of species j then element $s_{i,j} = 1$ while $s_{i,j} = 0$ otherwise. This is a "binary food web". However, species interactions are only feasible if enough energy is transferred to the predator. To assess the energetic feasibility, a foodweb needs to be quantified. This generates a $S \times S$ flow matrix $\mathbf{X}$, whose elements x are estimates of the magnitude of each feeding flow. This is a "quantified food web".
Theoretical ecologists have suggested simple models to generate binary food webs, based on the assumption that $L \in U(0,1)$ (i.e. random and cascade models: Cohen and Newman, 1985) or $L \in B(\alpha, \beta)$, (i.e. niche model: Williams and Martinez, 2000; and nested-hierarchy model: Cattin et al. 2004). The two latter models describe more realistic food webs.
On the other hand, applied ecologists have used Linear Inverse Modeling (LIM) to quantify the flows of real food webs (see van Oevelen et al., 2010), given an incomplete data set. The LIM methodology consists in solving the

following linear problem for the unknown flows x:

$$\mathbf{Ex} = \mathbf{F} \tag{1}$$

$$\mathbf{Ax} \approx \mathbf{B} \tag{2}$$

$$\mathbf{Gx} > \mathbf{H} \tag{3}$$

Here the first and/or second set of equations typically contain the component's mass balance equations and observed data, while the third set of equations holds physiological information and positivity constraints (i.e. the flows have a direction).
A LIM returns a "steady-state" snapshot of a food web, although the behavior of food webs under changing conditions is often of interest. This implies that the food web should be written as a dynamic model and solved by numerical integration.
Recently the R software has been made suitable for solving LIMs and for dynamic simulation thanks to two add-on packages (the R-package limSolve (Soetaert et al., 2009), and deSolve (Soetaert et al., 2010).

2 Methodology

We present how these three approaches can be combined in R:

1. We first generate binary food webs according to a theoretical model. Three functions generate the random, the cascade and the niche binary webs.

2. We then check the (energetic) feasibility, using the LIM methodology and quantify the flows. To do this, we convert the binary matrices into a LIM (1) assuming a minimal "growth efficiency" when consuming a species. If the LIM can be solved, then the problem is feasible and allows to estimate the flows.

3. The stability and long-term behavior of the quantified food web is then studied in dynamic simulations. To generate the dynamic system the species biomasses are needed, to convert the total ingestion and respiration rates into mass-specific rates and second order rates. We assumed allometric scaling of rates according to the trophic level of each species. The Jacobian matrix of the dynamic system allows to check the model's stability properties.

3 Examples

Figure 1 gives an illustration of the three types of simulated food webs. The random model was not feasible and could not be solved given the energetic constraints, so its flows are not represented. The cascade and niche model

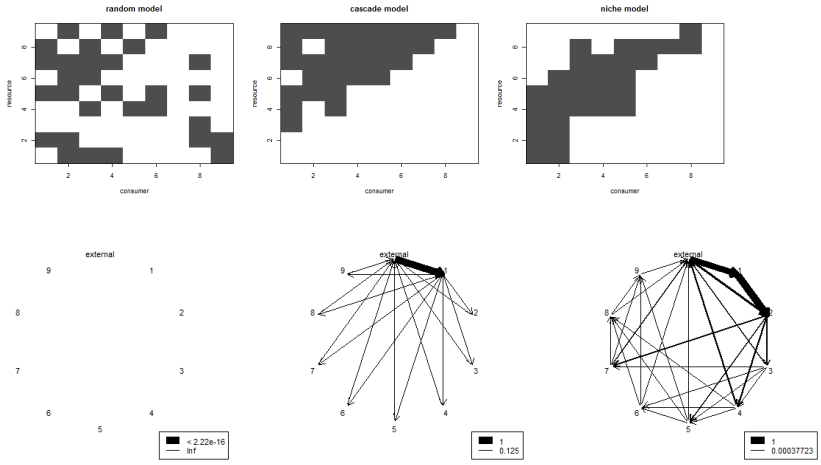

FIGURE 1. The binary food web and the quantitative food web for the three theoretical models.

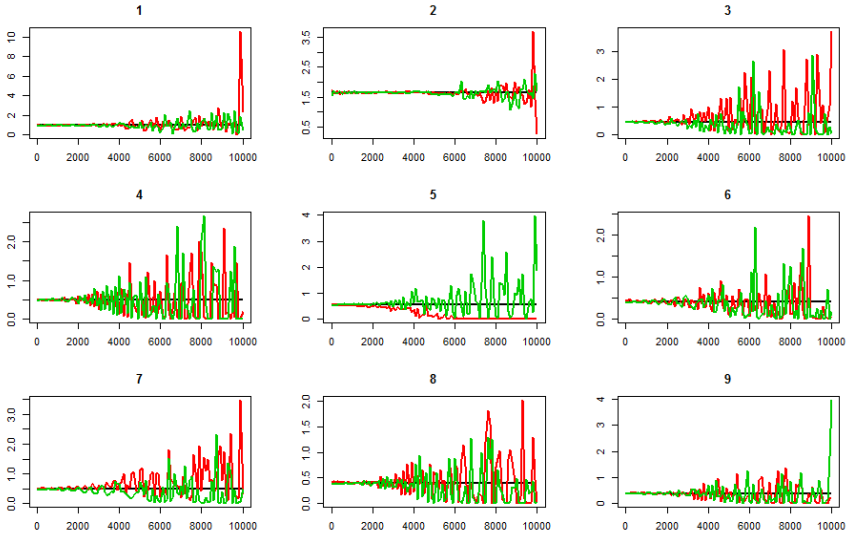

FIGURE 2. Output of the dynamic simulation for a niche food web.

were feasible and the flows could be quantified. Figure 2 represents the the dynamic simulation made for the niche food web. The output represents

a stable unperturbated foodweb (black line), and an increasingly instable foodweb when perturbed (the red and green lines).

4 Conclusion

The functions implemented in the open source framework R will allow to study the effect of human and environmental perturbations on artificially generated food webs.

Acknowledgments: The first author thanks to the CSIC grant program JAE-predoc, that made possible this study.

References

Cattin, M-F., Bersier, L-F., Banasek-Richter, C. et al. (2004). Phylogenetic constraints and adaptation explain food-web structure. *Nature*, **427**, 835-837.

Cohen, J.E. and Newman, C. M. (1985) A stochastic theory of community food webs I. Models and aggregated data. *Proceeding of the Royal Society of London, B*, **224**, 449461.

Soetaert, K., Van den Meersche, K., van Oevelen, D. (2009) limSolve: Solving Linear Inverse Models. R-package version 1.5.1.

Soetaert, K., Petzoldt, T. and Setzer R.W. (2010). Solving Differential Equations in R: Package deSolve. *Journal of Statistical Software*, **33** (9), 1-25.

van Oevelen, D., et al. (2010). Quantifying Food Web Flows Using Linear Inverse Models. *Ecosystems*, **13**, 32-45.

Williams, R.J. and Martinez, N.D. (2000). Simple rules yield complex food webs. *Nature*, **404**, 180183.

Estimation of the hazard function

Péter Raisz[1]

[1] Department of Applied Mathematics, University of Miskolc, H-3515 Miskolc-Egyetemváros, Hungary

E-mail for correspondence: matrp@uni-miskolc.hu

Abstract: The Cox's proportional hazards regression is widely used in the medical, biological, actuarial and engineering sciences, although the validity of the model is difficult to check. One possibility for this is simulation. Bender, Augustin and Blettner suggested a method for the simulations of data. This method is applied in MATLAB environment.

Keywords: Cox's proportional hazards; inverse baseline hazard; simulation.

1 Cox's proportional hazards model

Let random variable X be the lifetime of an item and let $f(x)$ and $F(x)$ denote the probability density function and the cumulative distribution function of random variable X. The survival function (reliability function) of the item is defined as

$$S(x) = P(X > x) = 1 - F(x).$$

The hazard rate function $\lambda(x)$ is defined as the probability per time unit that an individual (item) that has survived to the beginning of the time interval will die (fail) in this particular time interval

$$\lambda(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)} = -\frac{S'(x)}{S(x)} = -\frac{d[\ln S(x)]}{dx}.$$

In other words $\lambda(x)$ is the conditional probability that an individual of age x will die in the interval $(x, x + \Delta x)$.
There are several problems with survival data when traditional regression techniques are applied. They are typically non-normally distributed and the observations are generally heavily censorised. In Cox D. R. (1972) suggested a regression model and method for analyzing heavily censored survival data in order to know whether survival is influenced by certain factors (covariates) and if yes then how can one calculate the risk of a certain individual in a particular situation, with given covariates. The model is known as Cox's Proportional Hazards Model depends on 'proportionality'assumption. The

survival time is investigated as a function of a given set of independent variables (or covariates).
The continuous random variable T is investigated. Let $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ denote the covariates, then the hazard rate of random variable T is supposed to have the form

$$\lambda(t; x) = \lambda_0(t) e^{\beta' \mathbf{x}} = \lambda_0(t) e^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p},$$

where $\lambda_0(t)$ the so-called baseline hazard, while β is a $(p \times 1)$ vector of unknown parameters. There is no supposition on the specific form of $\lambda_0(t)$. The baseline hazard describes the hazard rate of an individual with zero covariates. The name 'proportional hazards model' describes the supposition on the hazard rate namely that the ratio of the hazard rates of two persons with the same covariates is constant over time. Otherwise there are no other assumptions on the distribution of random variable T, therefore this regression model can be considered as a nonparametric model. Cox's idea for the determination of the parameter β is the so-called partial likelihood. In the last two decades Cox-regression became extremly popular, it is widely used even in proving the efficiency of drogs . It is interesting to investigate the validity of the proportional hazards condition, which is the vital condition for the applicability of the Cox-regression. In these cases stochastic simulation is very important to test the validity of a model.

2 Generation of survival times

The generation of survival times for a CR-model is more complicated than in case of traditional regression methods, since the usual statistical softwares can generate just with a given probability distribution, not with a given hazard (rate) function.

The cumulative distribution function in terms of the hazard rate function

$$F(t) = 1 - S(t) = 1 + \exp\left(\int_0^t \lambda(u)\, du\right),$$

and in the special case of the proportional hazard model

$$\lambda(u) = \lambda_0(u) e^{\beta' \mathbf{x}}.$$

Introducing the baseline hazard function

$$\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$$

The distribution function can be expressed in terms of the baseline hazard function as

$$F(t; x) = 1 - \exp\left(-\Lambda_0(t) e^{\beta' \mathbf{x}}\right).$$

If $F(x)$ is the cumulative distribution function of random variable X,

$$P(X < x) = F(x),$$

then $F(X)$ is uniformly distributed in the interval $(0,1)$.
Therefore $F^{-1}(u)$ has cumulative distribution function $F(x)$ when u is uniformly distributed in the interval $(0,1)$. Furthermore if u is uniformly distributed in the interval $(0,1)$, then so is $1-u$.
Using these statements, if t is survival time in the CR-model, then

$$U = \exp\left(-\Lambda_0(t)\, e^{\beta' \mathbf{x}}\right)$$

is uniformly distributed in the interval $(0,1)$. If $\lambda_0(t) > 0$, then

$$t = \Lambda_0^{-1}\left(-\log(u) \cdot e^{-\beta' \mathbf{x}}\right),$$

where is uniformly distributed in the interval $(0,1)$.
Bender, Augustin and Blettner suggested to determine the inverse of the baseline hazard function numerically. This is solved in MATLAB.

Acknowledgments: The author was supported by TAMOP-4.2.1.B-10/2/KONV-2010-0001 project with support by the European Union, co-financed by the European Social Fund.

References

Bender R., Augustin T. and Blettner M. (2003). Generating Survival Times to Simulate Cox proportional Hazards Models. *LMU München*, http://epub.ub.uni-muenchen.de.

Cox D. R. (1972). Regression models and life tables (with discussions). *Journal of the Royal Statistical Society - B*, **34**, 187-220.

Optimization in regression models with LRD and restrictions

Elina Moldavskaya[1]

[1] Electrical Engineering, Technion, Haifa, Israel 32000

E-mail for correspondence: elinamoldavskaya@gmail.com

Abstract: We examine the solution of minimization problem of the least squares functional of the regression with long-memory and equality and inequality constraints on parameters. Approximate representation for least squares estimator is given. From these representation one can see the concrete structure of the estimators.

Keywords: Strong dependence; Regression; Least square estimators; Constraints, Approximate representation.

1 Introduction

The estimators of inequality-constrained regression models can be computed by iterative algorithms of mathematical programming, but they do not have analytical expressions in terms of given data. It brings obstacles for further analysis of the constrained regression.
Asymptotic behavior of such class of estimators in the models with strong dependence (in discrete and continuous cases) but without constraints on the parameter was investigated by many authors. In the papers by Taqqu, Dobrushin and Major a noncentral limit theorem describing the model mentioned above was formulated. Then asymptotic theory for least squares estimators in models with strong dependence (in cases without constraints) was developed, for example in papers by Yajima, Künsch, Dahlhaus, Ivanov and Leonenko etc.
On the other hand, regression models with independent or weakly dependent errors under restrictions on parameters were considered in papers by Korkhin, Knopov, Dupacova and Wets, Nagaraj and Fuller for discrete cases, Wang.
Asymptotic properties of least squares estimators in regression models with long memory and non-linear inequality-constraints on the parameter have been studied in papers by Moldavskaya. It was proved that least squares estimators converges in distribution to the optimal solution of the quadratic programming problem, but from this result it was not clear, what the specific asymptotic distribution of this estimators is.

We present approximate representation of the least squares estimators (LSE) in the regression models with long range dependence in the noise and inequality-constraints on the parameters. From this representation one can see the concrete structure of the estimators of these problems.

2 Basic model and main result

Consider the following inequality-constrained regression problem

$$\sum_{t=1}^{N}[y_t - f(x_t,\beta)]^2 \longrightarrow \min_{\beta\in\mathbf{R}^p} \qquad (1)$$
$$g_i(\beta) \le 0, \qquad i = 1,\dots,m \qquad h_j(\beta) = 0, \qquad j = 1,\dots,n$$

where $\beta = (\beta_1,\dots,\beta_p)'$ is the unknown vector-parameter to be estimated, $\{(x_t, y_t), t = 1,\dots,N\}$ is the sample data in the model

$$y_t = f(x_t,\beta_0) + \eta_t \qquad (2)$$

and β_0 is the true value of β. Denote the estimator of the problem (1) by $\widehat{\beta}_N$.

$$g_i(\beta), i = 1,\dots,m, h_j(\beta), j = 1,\dots,n, f(x_t,\beta)$$

are known functions and η_t, $t \in R$, is a random noise subordinated to process with long-range dependence.

$$V(N)(\hat{\beta}_N - \beta_0) = D\sum_{t=1}^{N}\nabla f(x_t,\beta_0)\eta_t + O_p(N^{-\alpha}L(N)\log\log N)^{\frac{1}{2}}), \quad (3)$$

where V(N) is a normalizing multiplier, D is a matrix, not related to the random noise. The first term on the right hand side of (3) is the main part and the second term is the remainder. From this expression we know what $V(N)(\hat{\beta}_N - \beta_0)$ looks like (approximately).

To prove main result, we follow Wang (2000) (who provided the analogous results for the case with independent errors), using results of Moldavskaya (2010) and law of the iterated logarithm for sums of non-linear functions of Gaussian variables exhibit a long range dependence provided by Taqqu (1977).

References

Moldavskaya, E.M. (2010). LSE of the parameters in the constrained regression models with LRD: approximate representation. In: *Proceedings of the International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management SCE-Shamoon College of Engineering.* 108-118, Beer Sheva, Israel.

Taqqu, M.S. (1977). Law of the iterated logarithm for sums of non-linear functions of Gaussian variables exhibit a long range dependence. *Z. Wahrshein. verw. Gebiete*, **Bd 40**, 203-238.

Wang, J. (2000). Approximate representation of estimators in constrained regression problem. *Scand. J. Statist.*, **27**, 21-33.

Application for back-testing day trading strategies

Juan Ricardo Rivera Peruyero[1], Pere Marti-Puig[1]

[1] Escola Politècnica Superior, Universitat de Vic, c/. Laura 13, 08500 Vic (Barcelona) Spain

E-mail for correspondence: juanricardo.rivera@uvic.cat

Abstract: An application for downloading any series of financial asset day prices from a free provider has been developed. This application also allows to back-test any day-trading strategy with the mined data easily.

Keywords: Day trading strategies; Back-testing; Algorithmic trading.

1 Introduction

An application for downloading any series of financial asset day prices from a free provider has been developed. This application also allows to back-test any day-trading strategy with the mined data easily. In a previous work we have proposed and evaluated different day trading methods in terms of the mean losses and benefits they produce under different market conditions, after 200 days of activity; see Marti and Rivera (2011). Those methods are formulated by algorithms that clearly specify all the actions to be taken and make investment decisions independent of the emotional aspects of trading because the strategy is governed by a set of defined rules. A trading strategy can be automated and performed by a computer that wraps trading formulas into automated order and execution systems. However, obtaining good trading algorithms is quite complicate and what we have observed is that a certain strategy may exhibit great results during a certain period of time, and instead, may work terribly wrong in another. Moreover, often it is not clear the reason that originates this disparity of results. The day trading strategies are mainly used by small investors. This contribution focuses on validating or invalidating different day trading strategies by performing a back-testing with data of different financial assets in a diary framework. The developed application is designed to study the methods mainly used by small investors to operate on the stock market.
This paper is organized as follows. In part 2 we present the kind of algorithms and strategies that we want to evaluate as well as the framework in which we observe the markets. In part 3 we present some simulation

experiments and the results obtained. Finally, section 4 contains some conclusions.

2 Approach, assumptions and application overview

The algorithms in which we are interested are those that do not assume any model to predict the financial asset behaviours as, for example, any future evolution of price, volatility or risk. We evaluate and explore techniques based on computing signals from the available series of prices and taking advantage of some patterns that those signals could exhibit in order to automatically generate orders of buying or selling. In some way we could say that we evaluate using statistics some empirical methods and we try to discover some operation rules that could work in practice. To develop the method of generating signals we use different signal processing tools, like linear, non-linear and adaptive filters, but we avoid the uses of models trying to model the behaviour of prices. The reason of avoiding complicated models of prediction is that the most common methods assume Gaussian statistics. However, the probability density functions of financial data differs from the Gaussian because exhibit bigger tails in both extremes. As a consequence, the common Gaussian assumption to model the variation of prices origins a lot of mistakes because the extreme events are underestimated; see Mandelbrot and Freeman (1982), Richard and Mandelbrot (2009), Mandelbrot and Taleb (2006). Long before the last financial crisis, Mandelbrot, who first reported this observation, strongly criticized the Black-Scholes model, widely used in banking to estimate the price of options and many other financial derivatives, because of its Gaussian behaviour assumption; see Mandelbrot and Freeman (1982).
In the day trading literature and on the Internet we can find many methods that fulfil our strategy requirements. Very often the authors state that their strategies perform quite well with poor statistic arguments justifying their statements. Some of these methods, e.g., Wilder (1978), Cava (2006), Ortiz de Zárate (2009), have been analysed with our application. The developed application mines the prices of financial assets from free web information providers like google/finaces and yahoo/finances. These two providers offer for free to their users the historic of day prices of any financial asset negotiated in any important market in the world. The huge among of available data will be very useful for the back-testing. That application is developed with Matlab because this software offers a lot of facilities for testing algorithms quickly and visualizing the results. With Matlab it is also easy to implement functions for mining web information.

3 Simulations and experiments

Almost all the trading algorithms, also those used in this work, depend on some parameters that are required to be tuned and that are commonly

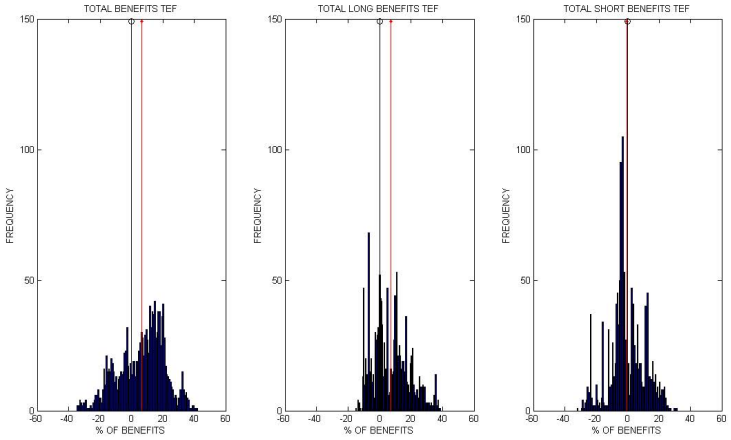

FIGURE 1. Histograms of the profits obtained with the ADX based method after 200 consecutive days in the market with $N = 6$ for signal $ADX_N(n)$ and $N = 24$ for signals $ADX_N^+(n)$ and $ADI_N^-(n)$. On the left, total profits; centre, profits from long operations; right, profits from short operations.

obtained from historical analysis using prices of the past. In some previous work Marti and Rivera (2011) we have explored the best parameter combinations that maximize the benefits over an historic of 10.000 prices of the Santander and Telefonica Spanish IBEX35 stocks. That operation was performed for both methods. Then, maintaining the best set of parameters for each method we have explored the same historic, but now computing the benefits/losses obtained from all periods of 200 consecutive days we can form from the historic. The results are presented by histograms. In Figures 1 and 2 there are represented tree histograms; the total of benefits obtained in 200 days (left), the benefits of the same period obtained from long operations (centre) and the benefits obtained from short operations (righ). These histograms are computed operating exclusively on the Telefonica stock. The histograms provide a fast interpretation of the dispersion of the results. The mean is represented using a red line. Figure 1 is obtained by applying the ADX method with $N = 6$ for signal $ADX_N(n)$ and N=24 for signals $ADX_N^+(n)$ and $ADI_N^-(n)$. In that case the mean of benefits is positive and we can conclude that most of benefits are given from long operations with a very poor contribution due to short operations. Figure 2 is obtained by applying the own method on the Telefonica stock prices data with parameters $P = 4$, $M_c = 5$ and $M_v = 1$. In that case the results are better than the ADX method with important expectations of mean benefits

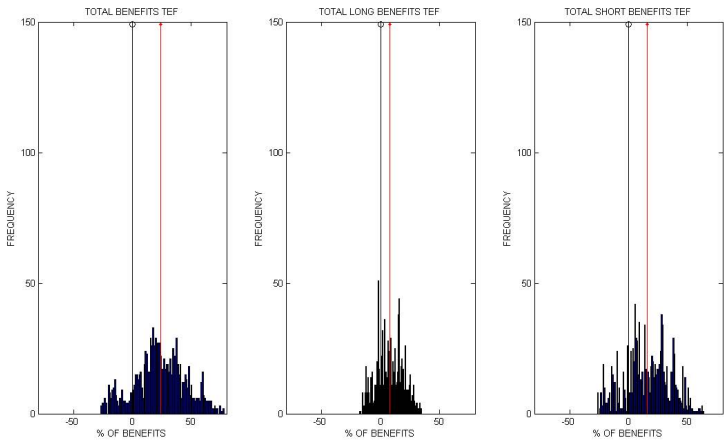

FIGURE 2. Histograms of the profits obtained using our own method after 200 consecutive days in the market with parameters $P = 4$, $M_c = 5$ and $M_v = 1$. On the left, total profits; centre, profits from long operations; right, profits from short operations.

even for short operations.
Some other experiments where done in order to provide an idea of how the systems work. An interesting result is that the benefits are strong dependent on the day in which the operations begin. Starting the operations one day or the day after can significantly modify the benefits. In Figure 3 we represent the total profits obtained as a function of time, so that the longitudinal axis begins with the first 200 day block analyzed and finishes with the block of 200 days that ends in the present. Note the strong dependency of benefits on the moment of acting in the market and its discrete nature.
In Figure 3 the time dependence of the benefits of operating 200 days consecutive days using the own method with P=19 and $M_c = 6$ and $M_v = 6$ against the Spanish Santader stock value is given.
A drawback of this experiment is that to identify the best set of parameters we use all the data available and then, with a selection of parameters, we use short periods (of 200 day) extracted from the same data used to tune the system.

A more realistic approach

The following set of experiments differs from the last ones because use the available data more realistically. Once an historic of prices of a particular financial asset is selected we always proceed in the same way. Consider

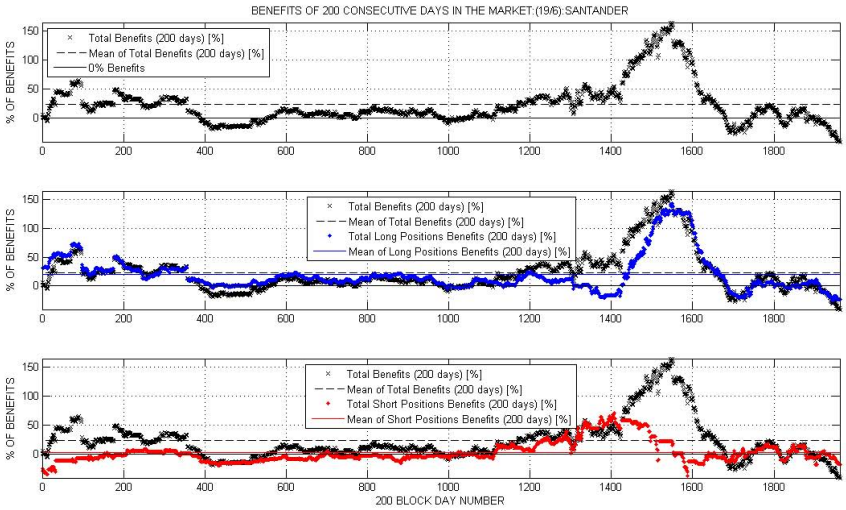

FIGURE 3. Evolution of 200-day profits over in time using our own method with parameters $P = 19$, $M_c = 6$ and $M_v = 6$. The discontinuous line represents the mean.

a test of benefits for the day n. Then we take N consecutive past days beginning with the day $n-1$ and going back to day $n-N$. We use those N past prices to find the best set of parameters that maximize the benefits in this period by a brute force exploration. Once the parameters are obtained, and maintaining these parameters, the trading method is evaluated with the next M consecutive days, indexed from $n+1$ to $n+M-1$, to show how the system performance on the next M future days. Then, after these M days in the market, the system closes positions and computes the total of benefits. So, for the simulation started at day n we obtain, at day $n+M$, one benefit (or loss) value. Next we do the same operation for the day $n+1$, for the day $n+2$, and so on until finishing all the historic available.
This kind of experiments is done to explore if updating each day the parameters with the recent past prices can improve or not the benefits.
In next pair of histograms we represent the results of those experiments. The histogram on the left represents the maximum of benefits obtained in the bloc of training days. The combination of parameters that maximize the benefits will be used to the operations of next bloc of 'future days.' On the right it is represented the histogram of benefits generated in the blocs of 'future days.' In the experiments below we have used the prices of the Santander stock.

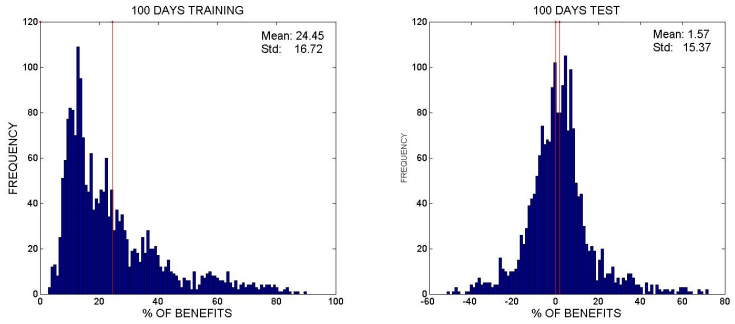

FIGURE 4. Performance of the ADX method when 100 days are used for the training and 100 days for the test. In the histogram on the left, the benefits reached in the training periods and, on the right, the performance of the method over 100 days. The red line indicates the mean.

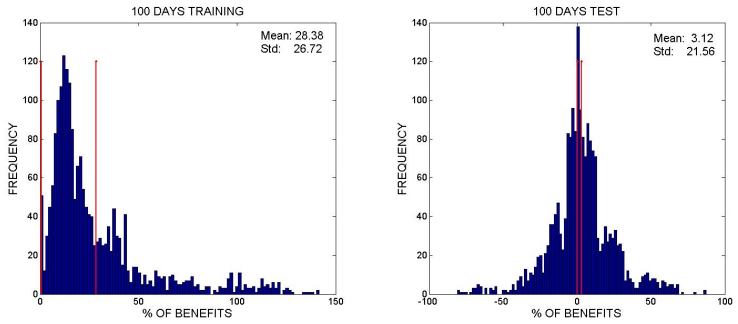

FIGURE 5. Performance of the own method when 100 days are used for the training and 100 days for the test. In the histogram on the left, the benefits reached in the training periods and on the right, the performance of the method over 100 days. The red line indicates the mean.

In Figure 4 we can see the performance of the ADX method when 100 days are used for the training and 100 days for the test. In that case we can observe, in the histogram on the left, that the benefits reached in the training periods, that are the maximum benefits possible for each period because he have found the set of parameters by a brute force exploration, has a mean of 24.45 while their standard deviation is only of 17.72. However, in the histogram on the right, we can observe that the parameters that maximize the benefits for the past 100 days do not work properly to the

next 100 days and the mean of benefits decreases to 1.57 while the standard deviation is practically the same.
A similar result is found when using the own method. The results of using 100 days for training are represented in Figure 5. In that case the mean of benefits reached in the training process is about 28.38% with a standard deviation of 26.72 as we can see in the histogram on the left. In that case, the benefits obtained in the next 100 days are showed in the histogram on the right, showing a mean of 3.12, which is better – the double – than the one of experiment of figure 4, but has decreased a lot.

4 Conclusions

We have developed a Matlab based application for downloading series of financial asset day prices. This application also allows to back-test any day-trading strategy with the mined data. The application visualize by graphic representations the generation of signals and orders produced by a selected strategy together with a price representation. So, observing the cases in which the system fails, it is possible to improve the strategy. Almost all the algorithms applied depend on a small number of parameters that can be adapted for different financial assets and different market conditions. One of the most interesting possibilities that the application offers is to perform exhaustive explorations of a set of parameters on all the historical to maximize some criteria in order to find the best combinations. These explorations can be time consuming intensive, however the Matlab platform is oriented to perform mathematical calculations and simulation. Other Matlab platform advantage is that we can use a lot of statistics and signal processing tools available in their toolboxes. Therefore, new trading strategies can be easily developed with some basic knowledge of programming on Maltlab. In this article we investigate the performance of two day trading strategies. Depending on the way to present the results obtained by these two systems it could seem that they work pretty well. In fact, after an exhaustive exploration in the space of tuning parameters on a past series of data we always find combinations that, on these data, produce spectacular benefits. The same also happens if we look for combinations that work terribly wrong. Some information about this kind of algorithms are often not presented whit rigor and small investors may be tempted to use strategies that offer few guarantees. Works like this can help to protect them. What we have observed is, for example, that the day the simulation begins on can strongly affect the resulting profits and that both methods can produce large benefits in only a few days but they can also lead to large losses. As the marked conditions change on time we have prepared some simulations to know what happen when we use the recent past data in order to determine the tuning strategy parameters and use them on a short future period of 100 day and we have observed that the histograms

of benefits obtained are centered near zero, with positive means close to zero and large standard deviations.

Acknowledgments: All authors have partially been supported by the Universitat de Vic under grant R0904.

References

Cava, J. L. (2006). *El Arte de especular. Las técnicas que mejor funcionan.* Madrid: Eds. M. A. Cava.

Mandelbrot, B. B., and Freeman, W. H. (1982). *The Fractal Geometry of Nature.* W. H. Freeman & Co.

Mandelbrot, B. B., and Taleb, N. (2006). A focus on the exceptions that prove the rule. *Financial Times*, March 23. Retrieved on 2010-10-17 from http://www.ft.com

Marti-Puig, P., and Rivera-Peruyero, J. R. (2011). Web-based system for evaluating day trading strategies. In: *Proceedings of 7th International Conference on the Next Generation Web Services Practices (NWeSP)*, 250-255, Salamanca, Spain. DOI: 10.1109.

Ortiz de Zárate, L. (2009). *Técnicas relevantes para la especulación en los mercados financieros. El método menos es más.* Madrid: Bolsa Relevante, S.L.

Richard, H. L., and Mandelbrot, B. B. (2009). *The (Mis)Behavior of Markets: A Fractal View of Risk, Ruin, and Reward.* New York: Basic Books.

Wilder, J. W. (1978). *New Concepts in Technical Trading Systems.* Greensboro, NC: Trend Research.

Asymptotics for SIMEX estimator in misclassification model

Iryna A. Sivak[1]

[1] National Taras Shevchenko University of Kyiv, Ukraine

E-mail for correspondence: sivirinaa@gmail.com

Keywords: SIMEX approach; misclassification.

We consider general regression problem with response Y and random discrete regressor X, which has possible outcomes $1, 2, ..., m$. Let X be true value which can be misclassified, and X^* be observed regressor. Misclassification error is described by the misclassification matrix Π, which is assumed to be known. The parameter of interest is regression vector parameter β. We investigate estimators for the MC-SIMEX approach suggested by Kuchenhoff et al. (2006).

We observe data $(Y_i, X_i^*)_{i=1}^n$ and construct naive estimator $\widehat{\beta}_{naive}[(Y_i, X_i^*)_{i=1}^n]$, neglecting the presence of misclassification. For a fixed grid of values $0 = \lambda_0 < \lambda_1 < ... < \lambda_M$, we simulate B new pseudo data sets with higher misclassification by

$$X_{b,i}^*(\lambda_k) := MC[\Pi^{\lambda_k}](X_i^*), \quad i = \overline{1,n},\ b = \overline{1,B},\ k = \overline{1,M},$$

where the misclassification operation $MC[M](X_i^*)$ denotes the simulation of a variable given X_i^* with misclassification matrix M. Then we replace regressors X_i^* by $X_{b,i}^*(\lambda_k)$ and for each pseudo samples we construct the naive estimator $\widehat{\beta}_{naive}[(Y_i, X_{b,i}^*)_{i=1}^n]$. Define

$$\widehat{\beta}_{\lambda_k} = B^{-1} \sum_{b=1}^{B} \widehat{\beta}_{naive}[(Y_i, X_{b,i}^*)_{i=1}^n], \quad k = \overline{1,M}.$$

Thus, $\widehat{\beta}_{\lambda_k}$ is the mean value of naive estimators that correspond to the matrix of misclassification Π^{λ_k+1}. Notice that $\widehat{\beta}_{\lambda_0} = \widehat{\beta}_{naive}[(Y_i, X_i^*)_{i=1}^n]$ for $\lambda_0 = 0$.

Then extrapolate back to the case of no misclassification. We select extrapolation function $G(1+\lambda, \Gamma) = \{G_i(1+\lambda, \Gamma_i)\}_{i=0}^{\dim \beta}$ of MC-SIMEX estimator in a form

$$G_i(1+\lambda, \Gamma_i) = \sum_{j=0}^{t} (1+\lambda)^j \gamma_{ij} = A_\lambda^T \Gamma_i,\ \lambda \geq 0,\ i = \overline{1, \dim \beta},$$

such that for each $\lambda \geq 0$:

$$\beta^*(\Pi^\lambda) \approx G(1+\lambda, \Gamma),$$

where β^* is the limit of naive estimators a.s. as $n \to \infty$, $A_\lambda = \{(1+\lambda)^j\}_{j=0}^t, \Gamma_i = \{\gamma_{ij}\}_{j=0}^t,\ i = \overline{1, \dim \beta}$.
The parameter Γ is estimated by the least squares method for $(1+\lambda_k, \widehat{\beta}_{\lambda_k})_{k=1}^M$ and the estimator is denoted by $\widehat{\Gamma}$. The MC-SIMEX estimator is then given by $\widehat{\beta}_{MC-SIMEX} = G(0, \widehat{\Gamma})$ which corresponds to $\lambda = -1$.
We assume conditions which ensure that Π^λ is again the misclassification matrix for any $\lambda \geq 0$.
Let $s = \dim \beta$. Suppose that the true value β_0 is an interior point of some compact K in $\mathbb{R}^s$ and U is an open set in $\mathbb{R}^s$, $K \subset U$. Assume that without misclassification, the estimation to the parameter β_0 can be obtained as a solution to the estimating equation for $\beta \in K$

$$\frac{1}{n}\sum_{i=1}^n \psi(Y_i, X_i, \beta) = 0,$$

where $\psi : \mathbb{R} \times \{1, 2, .., m\} \times U \to \mathbb{R}^s$.
The naive estimate with misclassification matrix Π is defined as a solution to estimating equation

$$\frac{1}{n}\sum_{i=1}^n \psi(Y_i, X_i^*, \beta) = 0,\ \beta \in K.$$

Therefore, the limit value $\beta^*(\Pi)$ is a solution to equation

$$E\psi(Y, X^*, \beta) = 0,\ \beta \in K.$$

Similarly $\beta^*(\Pi(\lambda_k))$ is a solution to equation

$$E\psi(Y, X^*(\lambda_k), \beta) = 0,\ \beta \in K.$$

We apply MC-SIMEX approach to the multiple-choice misclassification model ($m \geq 3$) and obtain expansions of naive estimator

$$\widehat{\beta}_{naive} = \beta^*(\Pi) + o(1) \text{ a.s. as } n \to \infty,$$

where

$$\beta^*(\Pi) = \beta_0 + \sum_{k=1}^l \frac{d^k \beta(I; \Pi - I)}{k!} + O(\|\Pi - I\|^{l+1}),\ \Pi \to I, \qquad (1)$$

and MC-SIMEX estimate satisfies

$$\widehat{\beta}_{MC-SIMEX} = \beta^*_{MC-SIMEX} + o(1) \text{ a.s. as } n \to \infty,$$

where

$$\beta^*_{MC-SIMEX} = \beta_0 + O(\|\Pi - I\|^{l+1}),\ \Pi \to I. \tag{2}$$

Here by the differential $d^k\beta(I; \Pi - I)$ we mean the value of multilinear form at point I with arguments $\Pi - I$,and I is identity matrix.
It follows from (1) and (2) that MC-SIMEX estimator is closer to the true value than naive estimator when $\Pi \to I$. This explains that MC-SIMEX approach gives better results than some consistent estimators in small and medium samples.
The results in this paper are joint with Prof. Alexander G. Kukush.

References

Kuchenhoff, H., Mwalili, S. M., and Lesaffre, E. (2006). A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX. *Biometrics*, **62**, 85-96.

Integral representations for cumulants of bilinear forms with applications in statistics

V. Zaiats[1]

[1] Universitat de Vic, c/. Sagrada Família, 7, 08500 Vic (Barcelona)

E-mail for correspondence: vladimir.zaiats@uvic.cat

Keywords: Bilinear forms; Cumulants; Cyclic products of kernels; Integrals.

This study deals with cumulant analysis of sample cross-correlograms and of other second-order statistics of stochastic processes and random fields. We will consider different types of second-order statistics and various classes of stochastic processes. In any case, a dominant role will be played by integral representations for cumulants of these statistics. Earlier studies of integral representations for the cumulants of different second-order statistics of stationary stochastic processes are due to R. Bentkus (1972, 1976). The cumulants of different polynomial statistics were considered in a later paper by the same author R. Bentkus (1977). Similar integral representations of the cumulants for periodograms of homogeneous random fields were considered in Guyon (1995) and Rosenblatt (1985).
Our main interest is with a particular integral representation for cumulants. This representation involves cyclic products of kernels, see Buldygin et el. (2002). We give some examples of integral representations for the cumulants of different bilinear forms of random vectors, stochastic processes, and random fields. All of these representations are finite sums of integrals involving cyclic products of kernels. We obtain a formula expressing the Gaussian component of cumulants of simple bilinear forms of random vectors. This representation follows from the Leonov-Shiryaev-Brillinger representation for cumulants. Since the cumulant of a simple bilinear form of a Gaussian random vector coincides with the Gaussian component of this cumulant, we obtain integral representations for the cumulants of different bilinear forms of Gaussian random vectors, stationary Gaussian stochastic processes, and homogeneous Gaussian random fields. We also establish some inequalities useful in applications. We consider the Rosenblatt distribution, see, e.g. Rosenblatt (1985). The explicit form of the logarithm of the characteristic function of the Rosenblatt distribution is an infinite sum of integrals involving cyclic products of kernels. We show that the Bentkus representation for the cumulants of spectral estimators of a stationary time series is reduced, after some algebra, to an integral involving cyclic products of kernels. We give a representative collection of statements and examples showing that

integrals involving cyclic products of kernels appear naturally in cumulants of second-order statistics of both Gaussian and non-Gaussian random vectors, time series, stochastic processes, and random fields. These examples make clear that integrals involving cyclic products of kernels merit a special attention.
This presentation is based on joint results obtained with V. Buldygin[†] and F. Utzet.

Acknowledgments: The author has been partially supported by the grant R0904 from the Universitat de Vic.

References

Bentkus, R. (1972). The asymptotic normality of an estimate of the spectral function. *Litovsk. Mat. Sb.*, **12**, 5-18.

Bentkus, R. (1976). Cumulants of estimates of the spectrum of a stationary sequence. *Lithuanian Math. J.*, **16**, 501-518.

Bentkus, R. (1977). Cumulants of multilinear forms of a stationary sequence. *Lithuanian Math. J.*, **17**, 27-46.

Buldygin, V., Utzet, F., and Zaiats., V (2002). A note on the application of integrals involving cyclic products of kernels. *Qüestiió*, **26**, 3-14.

Guyon, X. (1995). Random fields on a network. Modeling, statistics, and applications. New York, Springer.

Rosenblatt, M. (1985). Stationary sequences and random fields. Boston: Birkhäuser.

Estimates for the rate of strong approximation in the multidimensional invariance principle

Andrei Yu. Zaitsev[1]

[1] St. Petersburg Department of the Steklov Mathematical Institute, Russia

E-mail for correspondence: zaitsev@pdmi.ras.ru

Abstract: We formulate the new results about the rate of strong approximation in the multidimensional invariance principle which were published in the recent papers of Zaitsev (2006, 2007) and Götze and Zaitsev (2007). They can be considered as multidimensional generalizations and improvements of some results of Komlós, Major and Tusnády (1975–1976), Sakhanenko (1985) and Einmahl (1989).

Keywords: Multidimensional invariance principle; Strong approximation; Sums of independent random vectors

We consider the problem of constructing on a probability space a sequence of independent $\mathbf{R}^d$-valued random vectors $X_1, \ldots, X_n$ (with given distributions) and a corresponding sequence of independent Gaussian random vectors $Y_1, \ldots, Y_n$ so that the quantity $\Delta(X, Y) = \max_{1 \le k \le n} \big\| \sum_{j=1}^k X_j - \sum_{j=1}^k Y_j \big\|$ would be so small as possible with large probability. The estimation of the rate of strong approximation in the invariance principle may be reduced to this problem.

We formulate the results published in the papers of Zaitsev (2006, 2007) and Götze and Zaitsev (2007). They can be considered as multidimensional generalizations and improvements of some results of Komlós, Major and Tusnády (1975–1976), Sakhanenko (1985) and Einmahl (1989).

Let $\mathcal{H}$ be the class of non-negative non-decreasing continuous functions $H : [0, \infty) \to \mathbf{R}^1$ such that (for some $\delta > 0$ and $x_0 > 0$) the functions $H(x)/x^{2+\delta}$ and $x/\log H(x)$ are non-decreasing, for $x \ge x_0$. The distribution of a random vector ξ will be denoted below by $\mathcal{L}(\xi)$.

We consider the rate of strong approximation assuming that, for some function $H \in \mathcal{H}$, $\mathbf{E}\, H(\|X_j\|) < \infty$, $j = 1, 2, \ldots, n$. The $X_1, \ldots, X_n$ will be generally speaking non-i.i.d., but, for the sake of simplicity, we give below the results in the case of i.i.d. $X_1, \ldots, X_n$ only.

Theorem 1 *Let $H \in \mathcal{H}$ and ξ be a random vector with $\mathbf{E}\,\xi = 0$ and $\mathbf{E}\, H(\|\xi\|) < \infty$. Then, for any $z > 0$ and $n \ge 1$, there exists a construction*

such that

$$\mathcal{L}(X_j) = \mathcal{L}(\xi), \quad \mathbf{E}\, Y_j = 0, \quad \operatorname{cov} Y_j = \operatorname{cov} \xi, \tag{1}$$

for $j = 1, 2, \ldots, n$, *and*

$$\mathbf{P}\left(\Delta(X, Y) > c_1\, z\right) \le \frac{c_2\, n}{H(z)}, \tag{2}$$

where c_1 *and* c_2 *are positive quantities depending only on* $\mathcal{L}(\xi)$ *and on the function* $H(\,\cdot\,)$.

Theorem 2 *Let* H *and* ξ *satisfy the conditions of Theorem* 1. *Then there exists a construction such that* (1) *is satisfied for all* $j = 1, 2, \ldots$, *and*

$$\mathbf{P}\left(\limsup_{n\to\infty} \Big\| \sum_{j=1}^{n} X_j - \sum_{j=1}^{n} Y_j \Big\| / H^{-1}(n) < \infty\right) = 1.$$

Theorems 1 and 2 generalize to the multidimensional case the results of Komlós, Major and Tusnády (1975–1976). Einmahl (1989) proved the same statements for the functions H from the class $\widetilde{\mathcal{H}}$ of non-negative non-decreasing continuous functions H such that the functions $H(x)/x^{3+\delta}$ and $\sqrt{x}/\log H(x)$ are non-decreasing, for $x \ge x_0$. Clearly, there exists a lot of functions belonging to $\mathcal{H}$ and not belonging to $\widetilde{\mathcal{H}}$. For example, we may mention the functions $H(x) = \exp\bigl(\lambda x^{\beta}\bigr)$, $1/2 < \beta \le 1$, $\lambda > 0$.

Theorem 3 *Let* H *and* ξ *satisfy the conditions of Theorem* 1, *and the function* $x/\log(H(x)/L_H)$ *be non-decreasing for* $x > u$, *where* $L_H = n\,\mathbf{E}\, H\,(\|\xi\|)$ *and*

$$u = C_1\, H^{-1}(C_2\, L_H), \tag{3}$$

with some constants $C_1 \ge 1$ *and* $C_2 \ge 1$, *where* $H^{-1}(\,\cdot\,)$ *is the inverse function for* H. *Then, for any* $n \ge 1$, *there exists a construction such that* (1) *is satisfied for* $j = 1, 2, \ldots, n$, *and*

$$\mathbf{P}\left(\Delta(X, Y) > c_3\, z\right) \le \frac{c_4\, n}{H(z)}, \tag{4}$$

for any $z > 0$, *where* c_3 *and* c_4 *are positive quantities depending only on* C_1, C_2, $\mathcal{L}(X_1)$ *and on the function* $H(\,\cdot\,)$.

The conditions of Theorem 3 are satisfied, for example, for the function $H \in \mathcal{H}$ such that the function $H(x)/x^{\gamma}$ is non-increasing for some $\gamma > 2$. Then, in the proof of Corollary 2 of Zaitsev (2006), it was shown that one can take $u = H^{-1}(e^{\gamma} L_H)$ in (3).
Another example is given by $H(x) = \exp\bigl(\lambda x^{\beta}\bigr)$, $\lambda > 0$, $0 < \beta < 1$. In this case one can take $u = (1-\beta)^{-1/\beta}\, H^{-1}(L_H)$ in (3). It is clear that the list of examples may be prolonged.

The statement of Theorem 3 is much stronger than that of Theorem 1, since (4) is satisfied for all $z > 0$ simultaneously, on the same probability space, while, in Theorem 1, the probability space depends on z and (2) may be not valid for other z's. On the other hand, in Theorem 3, a condition with u from (3) is supposed to be satisfied. There exist functions which satisfy the conditions of Theorem 1, while the conditions of Theorem 3 are not satisfied for u from (3). For example, we may mention the functions which behave as $\exp\bigl(\lambda x\bigr)$, $\lambda > 0$, on some intervals of values of the argument x. Note, however, that the statement of Theorem 3 for $H(x) \equiv \exp\bigl(\lambda x\bigr)$ may be easily derived from the main result of Zaitsev (1998). It seems that Theorem 3 is new, even for $d = 1$.

Theorem 4 *Assume that $\gamma > 2$ and ξ is a random vector with $\mathbf{E}\,\xi = 0$, $\mathbf{E}\,\|\xi\|^{\gamma} < \infty$ and $\operatorname{cov}\xi = \mathbb{I}$, the identity operator. Then, for any $n \geq 1$, there exists a construction such that* (1) *is satisfied for $j = 1, 2, \ldots, n$, and*

$$\mathbf{E}\,\bigl(\Delta(X,Y)\bigr)^{\gamma} \leq c_5\, n\, \mathbf{E}\,\|\xi\|^{\gamma}\,, \tag{5}$$

where c_5 is a positive constant depending only on γ and on $\mathcal{L}(\xi)$.

Theorem 5 (Corollary) *Let ξ satisfy the conditions of Theorem* 4. *Then there exists a construction such that* (1) *is satisfied for all $j = 1, 2, \ldots$, and*

$$\mathbf{E}\,\bigl(\Delta(X,Y)\bigr)^{\gamma} \leq c_6\, n\, \mathbf{E}\,\|\xi\|^{\gamma}\,, \tag{6}$$

for all $n \geq 1$, where c_6 is a positive constant depending only on γ and on $\mathcal{L}(\xi)$.

Theorem 4 and Corollary 5 provide the statements which are stronger than that of Theorem 3 in the case, where $H\,(x) = x^{\gamma}$. Corollary 5 provides formally stronger assertion than Theorem 4, since (6) is valid in Corollary 5 for all $n \geq 1$ simultaneously, on the same probability space, while, in Theorem 4, the probability space depends on n. However, Corollary 5 follows from Theorem 4 by the application of an idea used by Lifshits (2007) for the corresponding generalization of a result of Sakhanenko (1985). One should construct independent blocks of 2^m summands, $m = 1, 2, \ldots$, according to Theorem 4, and then use the Rosenthal-type inequality for sums of non-negative random variables, see, e.g., Johnson, Schechtman and Zinn (1985). Theorems 1 and 3 were proved in Zaitsev (2007). The proof of Theorem 2 is based on the main result of Zaitsev (1998). It repeats the proof of Theorem 2 of Einmahl (1989). Theorem 4 is a i.i.d. case of Theorem 4 of Götze and Zaitsev (2007).

In this talk we also discuss the non-i.i.d. results from Zaitsev (2006) and Götze and Zaitsev (2007), which are multidimensional generalizations of weakened versions of the results of Sakhanenko (1985). In Götze and Zaitsev (2007), we considered the case $H\,(x) = x^{\gamma}$, $\gamma > 2$. The paper of Zaitsev (2006) is devoted to the general case, where $H \in \mathcal{H}$.

Acknowledgments: Research partially supported by the Grant of leading scientific scools NSh 638-2008.1.

References

Einmahl, U. (1989). Extensions of results of Komlós, Major and Tusnády to the multivariate case. *J. Multivar. Anal.*, **28**, 20-68.

Götze, F., and Zaitsev, A.Yu. (2007). Bounds for the rate of strong approximation in the multidimensional invariance principle. *Preprint SFB 701 no. 07-057*. Bielefeld University, Bielefeld, 1–27 (to appear in *Theor. Probab. Appl.*).

Johnson, W.B., Schechtman, G., and Zinn, J. (1985). Best constants in moment inequalities for linear combinations of independent and exchangeable random variables. *Ann. Probab.*, **13**, 234-253.

Komlós, J., Major, P., and Tusnády, G. (1975, 1976). An approximation of partial sums of independent RV'-s and the sample DF. *Z. Wahr. verw. Geb.*, I **32**, 111-131; II **34**, 34-58.

Lifshits, M.A. (2007). Lectures on the strong approximation. St. Petersburg University Press, St. Petersburg, 1–32 (in Russian).

Sakhanenko, A.I. (1985). Estimates in the invariance principles. In: *Trudy Inst. Mat. SO AN SSSR, 5*. 27-44, Nauka, Novosibirsk.

Zaitsev, A.Yu. (1998). Multidimensional version of the results of Komlós, Major and Tusnády for vectors with finite exponential moments. *ESAIM: Probability and Statistics*, **2**, 41-108.

Zaitsev, A.Yu. (2006). Estimates for the rate of strong approximation in the multidimensional invariance principle. *Zapiski Nauchnyh Seminarov POMI*, **339**, 37-53 (in Russian).

Zaitsev, A.Yu. (2006). Estimates for the rate of strong Gaussian approximation for the sums of i.i.d. multidimensional random vectors. *Zapiski Nauchnyh Seminarov POMI*, **351**, 141-157 (in Russian).

Estimates for the concentration functions in the Littlewood–Offord problem

Andrei Yu. Zaitsev[1]

[1] St. Petersburg Department of the Steklov Mathematical Institute, Russia

E-mail for correspondence: zaitsev@pdmi.ras.ru

Keywords: Concentration function; Sums of i.i.d. random variables.

Let $X, X_1, \dots, X_n$ be independent identically distributed random variables. This talk deals with the behavior of the concentration functions of the weighted sums $\sum_{k=1}^n a_k X_k$ with respect to the arithmetic structure of coefficients a_k. Such concentration results recently became important in connection with investigations about singular values of random matrices. We formulate some refinements of results of Rudelson and Vershynin (2009), Friedland and Sodin (2007), Vershynin (2011). which are proved in the recent preprints Eliseeva and Zaitsev (2012) and Eliseeva, Götze and Zaitsev (2012).

References

Eliseeva, Yu. S., Götze, F., and Zaitsev, A. Yu. (2012). Estimates for the concentration functions in the Littlewood–Offord problem, arXiv: 1203.6763.

Eliseeva, Yu. S., and Zaitsev, A. Yu. (2012). Estimates for the concentration functions of weighted sums of independent random variables, arXiv: 1203.5520. Submitted to *Theory Probab. Appl.*.

Friedland, O., and Sodin, S. (2007). Bounds on the concentration function in terms of Diophantine approximation. *C. R. Math. Acad. Sci. Paris*, **345**, 513-518.

Rudelson, M., and Vershynin, R. (2009). Smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math.*, **62**, 1707-1739.

Vershynin, R. (2011). Invertibility of symmetric random matrices, arXiv: 1102.0300. To appear in *Random Structures and Algorithms*.

Informational analysis of magnetotelluric data

Luciano Telesca[1], Michele Lovallo[2], Marianna Balasco[1], Vincnezo Lapenna[1], Gerardo Romano [1], Agata Siniscalchi[3]

[1] National Research Council, Institute of Methodologies for Environmental Analysis, C.da S.Loja, 85050 Tito (PZ) Italy
[2] ARPAB, via della Fisica 18 C/D, 85100 Potenza, Italy
[3] Dipartimento di Scienze della Terra e Geoambientali, Universit degli Studi di Bari, Italy

E-mail for correspondence: luciano.telesca@imaa.cnr.it

Abstract: The time dynamics of magnetotelluric parameters is investigated by studying their daily measurements of Earth's apparent resistivity and phase by using the Fisher-Shannon Information plane method. The obtained results suggest a relationship between the informational properties of the time series and the sounding depth.

Keywords: Magnetotellurics; Fisher Information Measure; Shannon entropy.

1 Introduction

The temporal fluctuations of a signal can be studied in the so-called Fisher-Shannon information plane (FS), which is a plane whose coordinate axes are two statistical measures, the Fisher Information Measure (FIM) and the Shannon entropy power (N_X). Both the statistical quantities are used to investigate complex and nonstationary signals; in particular the FIM quantifies the amount of organization or order of a system, while N_X quantifies the amount of uncertainty or disorder of a system. Fisher (1925) introduced in the framework of statistical estimation the FIM, which was later employed by Frieden (1929) to develop as a versatile method able to describe the evolution laws of physical systems. The accurate description of the behavior of dynamical systems and the characterization of complex signals generated by these systems is efficiently performed by the FIM (Vignat and Bercher, 2003). The characterization of the time dynamics of EEG records and the detection of significant variations in the evolution of nonlinear dynamical systems were performed by Martin et al. (1999) by using the FIM that was revealed to be a very useful tool in dealing with theoretical as well as observational characteristics of dynamical systems (Martin et al., 2001). FIM was used in studying several geophysical and

environmental phenomena, revealing its ability in describing the complexity of a system (Telesca et al., 2008; Telesca et al., 2009) and suggesting its use as to reveal reliable precursors of critical events (Telesca et al., 2009b, Telesca et al., 2005; Telesca et al., 2005b; Telesca et al., 2010). Shannon entropy represents the well known method used to capture the fundamental state of things (Angulo et al., 2008). It is generally applied to quantify the amount of uncertainty that is inherent in the prediction of the output of a probabilistic event (Shannon, 1948). For instance, in case of discrete distributions, if one is able to predict exactly the outcome of a probabilistic event before it happens, the probability assumes the maximum value, but, as a consequence, the Shannon entropy assumes the minimum value. Thus the Shannon entropy will be zero for deterministic events. In case of continuous distributions (probability densities), in which the variable ranges over the real line, the Shannon entropy can take any real value, positive or negative. To avoid the difficulty of dealing with negative information measures, we use the Shannon power entropy N_X that will be defined below.
In this work, we analyze the time series of magnetotelluric (MT) parameters measured in southern Italy by using the combination of the statistical measures of FIM and Shannon entropy (the Fisher-Shannon Information plane), in order to identify hidden dynamical patterns.

2 The magnetotelluric method

The magnetotelluric method (MT) is a geophysical technique used to image the subsurface electrical resistivity by using the Earth's natural varying electromagnetic field, characterized by a broad range of periods. On the base of the skin depth formula, $\delta = 503(\rho T)^{1/2}$(in meters), where ρ is the Earth's resistivity and T the period (Kaufmann and Keller, 1981), the investigation depth increases with period and can reach several tens of kilometers for longer periods. Simultaneously measuring the horizontal components of the electric (E) and magnetic (H) field, the frequency dependent impedance tensor (Z) (called transfer function) of the subsoil can be estimated:

$$\begin{vmatrix} E_x(\omega) \\ E_y(\omega) \end{vmatrix} = \begin{vmatrix} Z_{xx}(\omega) & Z_{xy}(\omega) \\ Z_{yx}(\omega) & Z_{yy}(\omega) \end{vmatrix} = \begin{vmatrix} H_x(\omega) \\ H_y(\omega) \end{vmatrix} \tag{1}$$

where ω is the angular frequency, (E_x, E_y) and (H_x, H_y) represent respectively the electric and magnetic components in an orthogonal reference and $Z(\omega)$ is the MT transfer function tensor. As a simple linear system, the transfer function $Z(\omega)$ acts as a filter, while the magnetic and electric fields represent the input and output respectively.
The apparent resistivity and phase is defined by the following equations:

$$\rho_{ij}(\omega) = \frac{1}{\mu_0 \omega} |Z_{ij}(\omega)|^2 \tag{2}$$

and

$$\phi_{ij}(\omega) = tan^{-1} \frac{Im[Z_{ij}(\omega)]}{Re[Z_{ij}(\omega)]} \tag{3}$$

where μ_0 is the permeability of the vacuum, Z_{ij} are the complex components of the tensor defined in Eq. 1, with i, $j = x$ or y.

3 The methods

Let $f(x)$ be the probability density of a signal x. Its FIM I is given by

$$I = \int_{-\infty}^{+\infty} \left(\frac{df(x)}{dx} \right)^2 \frac{dx}{f(x)}. \tag{4}$$

The Shannon entropy is given by

$$H_X = -\int_{-\infty}^{+\infty} f(x) log\left(f(x)\right) dx. \tag{5}$$

For convenience the alternative notion of Shannon entropy power will be used

$$N_X = -\frac{1}{2\pi e} e^{2H_X}. \tag{6}$$

The N_X satisfies the so-called 'isoperimetric inequality,' a lower bound to the Fisher-Shannon product, given by $IN_X 1$, in case of 1-dimensional space. Theisoperimetric inequality indicates that the FIM and the Shannon entropy power are intrinsically linked quantities; thus the analysis of the time dynamic of signals should be improved when analyzed in the so called Fisher-Shannon (FS) information plane, whose y- and x-axis are the FIM and the Shannon entropy power, respectively. Vignat and Bercher (2003) showed that examining simultaneously both the Shannon entropy power and the FIM by means of the FS plane the characterization of the non-stationary dynamics of complex signals could be improved. The product IN_X can also be employed as a statistical measure of complexity (Angulo et al., 2008). The line IN_X=1 separates the FS plane in two parts, of which one is allowed (IN_X >1) and the other not (IN_X <1).

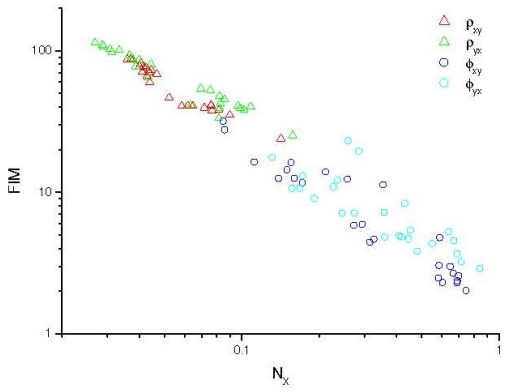

FIGURE 1. FS Information plane for the MT parameters.

4 Results

We analyzed the informational properties of the daily time series of normalized apparent Earth's resistivity and phase calculated from the electromagnetic field measured in Tramutola station, located in southern Italy, corresponding to 25 sounding periods (0.74473s, 0.93091s, 1.17029s, 1.46286s, 1.86182s, 2.40941s, 3.15077s, 4.096s, 5.28516s, 6.82667s, 8.62316s, 10.92267s, 13.65333s, 16.384s, 21.14064s, 27.30667s, 34.49263s, 43.69067s, 54.61333s, 65.536s, 84.56258s, 113.97565s, 154.20235s, 163.84s, 238.31273s). We investigated the two components of the magnetotelluric field (xy and yx). We calculated the Shannon entropy power N_X and the FIM I for all the analyzed series with the increase of the sounding period T (which is related to the sounding depth). In order to identify particular pattern and organization in the resistivity and phase, we analyzed their behavior in the Fisher-Shannon (FS) information planes. Figure 1 shows the FS plane for the analysed magnetotelluric parameters. The FS information plane shows that: 1) the phases are well discriminated from the resistivities; 2) the resistivities are characterized by lower organization and higher disorder than the phases. The complexity measures IN_X for the resistivity (Figure 2) and phases (Figure 3) were calculated. It is observed that in both cases the complexity measure is the highest for two periods 5.28s and 6.82s. The complexity pattern for the resistivity ρ_{xy} is approximately stable for all the sounding periods, while that of ρ_{yx} changes significantly with the sounding depth showing an abrupt increase starting from 34.49263s; both the components, however, are characterized by a minimum in the range [8.62316s,

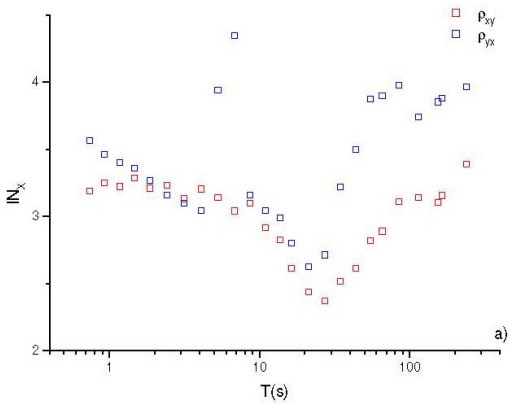

FIGURE 2. Complexity measure for resitivities (a) and phases (b).

34.49263s]. The complexity pattern of both phases is approximately similar for both directions, with a maximum at T=21.14s.

5 Conclusions

We studied the daily Earth's magnetotelluric apparent resistivity and phase in 25 sounding periods in a site in southern Italy by means of two information-theoretic measures: the FS information plane and the complexity measure. The series were calculated in directions xy and yx. It was observed that order or organization and uncertainty generally decrease and increase respectively as the sounding period (sounding depths) increases. Shannon entropy is characterized by a minimum in the period range [1.86182s, 8.62316s]. In this work, two different dynamics were found to drive resistivity fluctuations in two different period band with a transition zone corresponding more or less with the range[1.86182s, 8.62316s]. The phase presents a non trivial relationship between the order or the uncertainty with the sounding periods; the maximum (minimum) order degree (uncertainty degree) corresponds to sounding periods ranging around 10s. The higher regularity of the resistivity respect to that of the phases is also shown in the FS plane, in which the phases are less aligned than the resistivity. Our findings, although still preliminary, would suggest that investigating the informational properties of magnetotelluric parameters (apparent resistivity and phase) could contribute to better understand the complex processes occurring in the Earth's crust.

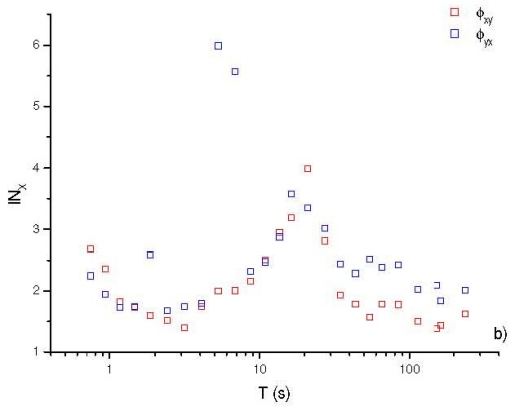

FIGURE 3. Complexity measure for resitivities (a) and phases (b).

Acknowledgments: The present study was supported by the project CNR-CSIC Development of of robust processing tools for the analysis of magnetotelluric data in the framework of the Bilateral Agreement for Scientific and Technological Cooperation between CNR and CSIC 2011-2012.

References

Angulo, J. C., Antolin, J., Sen, K. D. (2008) Fisher-Shannon plane and statistical complexity of atoms. *Phys. Lett. A*, **372**, 670-674.

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700-725.

Frieden, B. R. (1990). Fisher information, disorder, and the equilibrium distributions of physics. *Phys. Rev. A*, **41**, 4265-4276.

Kaufman, A. A., Keller, G. V. (1981). *The magnetotelluric sounding method, in: Methods in Geochemistry and Geophysics.* Elsevier Scientific Publ., Amsterdam, 583pp.

Martin, M. T., Pennini, F., Plastino, A. (1999) Fisher's information and the analysis of complex signals. *Phys. Lett. A*, **256**, 173-180.

Martin, M. T., Perez, J., Plastino, A. (2001) Fisher information and nonlinear dynamics. *Physica A*, **291**, 523-532.

Shannon, C. E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379-423, 623-656.

Telesca, L., et al. (2008) The Fisher information measure and Shannon entropy for particulate matter measurements. *Physica A*, **387**, 4387-4392.

Telesca, L., et al. (2009) Analysis of dynamics in Cd, Fe, and Pb in particulate matter by using the Fisher-Shannon method. *Water Air Soil Pollut.*, **201**, 33-41.

Telesca, L., Lovallo, M., Ramirez-Rojas, A., Angulo-Brown, F. (2009b) A nonlinear strategy to reveal seismic precursory signatures in earthquake-related selfpotential signals. *Physica A*, **388**, 2036-2040.

Telesca, L., Lapenna, V., Lovallo, M. (2005) Fisher information analysis of earthquake-related geoelectrical signals. *Nat. Hazards Earth Syst. Sci.*, **5**, 561-564.

Telesca, L., Lapenna, V., Lovallo, M. (2005b) Fisher information measure of geoelectrical signals. *Physica A*, **351**, 637-644.

Telesca, L., Lovallo, M., Carniel, R. (2010) Time-dependent Fisher information measure of volcanic tremor before 5 April 2003 paroxysm at Stromboli volcano, Italy. *J. Volcanol. Geoter. Res.*, **195**, 78-82.

Vignat, C., Bercher, J.-F. (2003) Analysis of signals in the Fisher-Shannon information plane. *Phys. Lett. A*, **312**, 27-33.