SORT

Statistics and Operations Research Transactions

Volume 28
Number 1, January - June 2004

ISSN: 1696-2281

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Aims

SORT (*Statistics and Operations Research Transactions*) – formerly *Qüestiió* - is an international journal launched in 2003, published twice-yearly by the Institut d'Estadística de Catalunya (Idescat), co-sponsored by the Universitat Politècnica de Catalunya (UPC), Universitat de Barcelona, Universitat Autònoma de Barcelona and Universitat de Girona and with the co-operation of the Spanish Region of the International Biometric Society. SORT promotes the publication of original articles of a methodological or applied nature on statistics, operations research, official statistics and biometrics.

The journal is described in the *Encyclopedia of Statistical Sciences*, and referenced in the *Current Index to Statistics*, the *Índice Español de Ciencia y Tecnología*, *Statistical Methods and Abstracts*, as well as MathSci of the American Mathematical Society (*Current Mathematical Publications and Mathematical Reviews*).

SORT represents the third series of the *Quaderns d'Estadística i Investigació Operativa* (*Qüestiió*), published by Idescat since 1992 until 2002, which in turn followed from the first series *Quaderns d'Estadística, Sistemes, Informàtica i Investigació Operativa* (1977-1992). The three series of *Qüestiió* have their origin in the *Cuadernos de Estadística Aplicada e Investigación Operativa*, published by the UPC till 1977.

ISSN: 1696-2281
SORT 28 (1) January - June 1-108 (2004)

SORT

Statistics and Operations Research Transactions

Sponsoring institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Institut d'Estadística de Catalunya

Supporting institution

Spanish Region of the International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Volume 28 (1), January-June 2004

Formerly Qüestiió

Contents

Articles

PAPER ECOLÒGIC

SORT 28 (1) January-June 2004, 1-8

On-line nonparametric estimation★

R. Khasminskii*

Wayne State University, USA

Abstract

A survey of some recent results on nonparametric on-line estimation is presented. The first result deals with an on-line estimation for a smooth signal *S(t)* in the classic 'signal plus Gaussian white noise' model. Then an analogous on-line estimator for the regression estimation problem with equidistant design is described and justified. Finally some preliminary results related to the on-line estimation for the diffusion observed process are described.

MSC: 62G05, 62G08, 62M05.

Keywords: White Gaussian Noise, on-line estimation, nonlinear filter.

1 On-line estimation of a signal in a Gaussian white noise model

We consider an observation process $X^\varepsilon(t)$ having the form

$$X^\varepsilon(t) = \int_0^t S(s)ds + \varepsilon W(t),\ t \in [0, 1]. \tag{1.1}$$

Here $W(t)$ is a standard Wiener process and $\varepsilon > 0$ is a small parameter. Denote by $\Sigma(\beta, L)$ a class of functions $S(t), t \in [0, T]$ having k derivatives on $(0, T)$ with k-th derivative $S^{(k)}(t)$ satisfying the Hölder condition with the exponent $\alpha \in (0, 1](\beta = k + \alpha)$:

$$|S^{(k)}(t + h) - S^{(k)}(t)| \le L|h|^\alpha.$$

* *Address for correspondence:* Department of Mathematics, Wayne State University, Detroit, MI 48202, USA. E-mail: rafail@math.wayne.edu

★ Partially supported by NSF Grant DMS 9971608.

Received: October 2003

Accepted: January 2004

The following problem was considered by Ibragimov and Khasminskii (1981): what is the rate of convergence to 0 for the best estimators of $S, S^{(1)}, \ldots, S^{(k)}$, as $\varepsilon \to 0$, and how can estimators with this rate be created? It was shown in Ibragimov and Khasminskii (1980a, 1981) that the kernel and projection estimators $\widehat{S}_\varepsilon(t)$ for a suitable choice of parameters have a property

$$\sup_{S \in \Sigma(\beta, L)} E\Big(\Big[\frac{\widehat{S}_\varepsilon(t) - S(t)}{\varepsilon^{2\beta/(2\beta+1)}}\Big]^2 + \sum_{j=1}^{k} \Big[\frac{\widehat{S}_\varepsilon^{(j)}(t) - S^{(j)}(t)}{\varepsilon^{2(\beta-j)/(2\beta+1)}}\Big]^2\Big) \le C, \tag{1.2}$$

and there are no estimators with uniformly in $\Sigma(\beta, L)$ better rate of convergence to 0 risks (here and below we denote by C, C_i generic positive constants, which may be different and do not depend on ε).

In some applications it is necessary to create a tracking (or on-line) type of estimator for S, that is estimators with the property: $\widehat{S}_\varepsilon(t+h)$ is based on $\widehat{S}_\varepsilon(t)$ and observation process on the time interval $[t, t+h]$ only. Unfortunately the well-known kernel and projection estimators do not have this property.

The tracking estimator for the model (1.1) was proposed by Chow *et al.* (1997). This estimator has the structure of a Kalman filter. Heuristically this estimator is based on the auxiliary filtering model

$$\begin{aligned} dS(t) &= S^{(1)}(t)dt \\ dS^{(j)}(t) &= S^{(j+1)}(t)dt, \ j = 1, \ldots, k-1 \\ dS^{(k)}(t) &= \sigma_\varepsilon dW'(t) \\ dX_t &= S(t)dt + \varepsilon dW(t). \end{aligned} \tag{1.3}$$

It is clear that the last equation in (1.3) is equivalent to (1.1). Assuming that the standard Wiener processes $W(t)$ and $W'(t)$ are independent and choosing a constant σ_ε in a suitable way, we arrive at the following estimator for $S(t) = S^{(0)}(t), \ldots, S^{(k)}(t)$ (see details in Chow *et al.* (1997))

$$\begin{aligned} d\widehat{S}_\varepsilon^{(j)}(t) &= \widehat{S}_\varepsilon^{(j+1)}(t)dt + \frac{q_j}{\varepsilon^{2(j+1)/(2\beta+1)}}(dX^\varepsilon(t) - \widehat{S}_\varepsilon(t)dt), \\ j &= 0, 1, \ldots, k-1 \\ d\widehat{S}_\varepsilon^{(k)}(t) &= \frac{q_k}{\varepsilon^{2(k+1)/(2\beta+1)}}(dX^\varepsilon(t) - \widehat{S}_\varepsilon(t)dt), \end{aligned} \tag{1.4}$$

subject to the initial conditions $\widehat{S}(0) = S_\circ, \widehat{S}^{(j)}(0) = S_\circ^j$, $j = 1, \ldots, k$, which reflect a priori information on $S(0), S^{(j)}(0)$, $j = 1, \ldots, k$.

Denote by $p_k(\lambda)$ the polynomial

$$p_k(\lambda) = \lambda^{k+1} + q_0\lambda^k + \ldots + q_{k-1}\lambda + q_k.$$

The following result was proven in Chow *et al.* (1997):

Theorem 1.1 *For any choice $q_0, \dots, q_k$ such that all roots of the polynomial $p_k(\lambda)$ have negative real parts, and for arbitrary bounded initial conditions $S_\circ, \dots, S_\circ^k$ the tracking filter (1.4) has the property: there exists an initial boundary layer $\Delta_\varepsilon = C_1 \varepsilon^{2/2\beta+1} \log(1/\varepsilon)$ such that for $t \geq \Delta_\varepsilon$ the inequality (1.2) is valid.*

Remark 1.2 *It is proven in Chow* et al. *(1997) also, that the initial boundary layer of the order $\varepsilon^{2/2\beta+1}$ is inevitable for any tracking type estimator.*

Remark 1.3 *An analogous result was proven in Chow* et al. *(2001) for the estimation of a time dependent spatial signal observed in a cylindrical Gaussian white noise model of the small intensity ε. It is proven in Chow* et al. *(2001) that outside of the inevitable boundary layer the symbiosis of a projection estimator in the space variables and tracking type estimator in the time variable also has an optimal rate of convergence of risks to 0, as $\varepsilon \to 0$, for a suitable choice parameters of a tracking filter and a projection estimator.*

2 On-line estimation of a smooth regression function

It is well known that the model (1.1) is a natural approximation for the regression estimation model with equidistant design. In more detail, consider the following statistical model. Let $f(t) \in \mathbb{R}^1, t \in [0, 1]$, be a function from $\Sigma(\beta, L)$, $t_{in} = \frac{i}{n}, i = 1, \dots, n$, and the observation model has the form

$$X_{in} = f(t_{in}) + \sigma(t_{in})\xi_{in}, \tag{2.1}$$

where $(\xi_{in})_{i \leq n}$ is a sequence of i.i.d. random variables with $E\xi_{in} = 0$, $E\xi_{in}^2 = 1$ and $\sigma^2(t_{in}) < C$. The natural analogy of an estimator (1.4) is the tracking estimator (hereafter we write for brevity t_i instead of t_{in} and X_i instead of X_{in})

$$\begin{aligned} \widehat{f_n^{(j)}}(t_i) &= \widehat{f_n^{(j)}}(t_{i-1}) + \frac{1}{n}\widehat{f_n^{(j+1)}}(t_{i-1}) + \frac{q_j}{n^{\frac{(2\beta-j)}{2\beta+1}}}(X_i - \widehat{f_n^{(0)}}(t_{i-1})) \\ j &= 0, 1, \dots, k-1 \\ \widehat{f_n^{(k)}}(t_i) &= \widehat{f_n^{(k)}}(t_{i-1}) + \frac{q_k}{n^{\frac{(2\beta-k)}{2\beta+1}}}(X_i - \widehat{f_n^{(0)}}(t_{i-1})) \end{aligned} \tag{2.2}$$

subject to some initial conditions $\widehat{f_n^{(0)}}(0), \widehat{f_n^{(1)}}(0), \dots, \widehat{f_n^{(k)}}(0)$.The following theorem, analogous to Theorem 1.1, was proven by Khasminskii and Liptser (2002):

Theorem 2.1 *Let $q_0, \dots, q_k$ are chosen so that all roots of the polynomial $p_k(\lambda)$ have negative real parts. Let an observation model has the form (2.1), $f \in \Sigma(\beta, L)$ and $\sigma^2(t) < C$. Then the estimator (2.2) with arbitrary bounded initial conditions*

$\widehat{f_n^{(0)}}(0), \widehat{f_n^{(1)}}(0), \ldots, \widehat{f_n^{(k)}}(0)$ *possesses the property: for* $t_l > C_1 n^{-\frac{1}{2\beta+1}} \log n := \delta_n$

$$\sup_{f\in\Sigma(\beta,L)} \sum_{j=0}^{k} E(f^{(j)}(t_\ell) - \widehat{f_n^{(j)}}(t_\ell))^2 n^{\frac{2(\beta-j)}{2\beta+1}} \le C_2. \tag{2.3}$$

Remark 2.2 *Similar the proof an analogous property of the estimator (1.4) it is easy to conclude from the results in Stone (1980) and Ibragimov and Khasminskii (1980b) that the rate of convergence of risks to zero for* $n \to \infty$ *in (2.3) is unimprovable. The boundary layer of order* $n^{-\frac{1}{2\beta+1}}$ *is also inevitable for any on-line estimator.*

Remark 2.3 *It is easy to apply the estimator (2.2) for the estimation* f *with the best rate of convergence of risk to* 0 *for* all $t \in [\delta_n, 1]$. *It is enough to set, for instance,* $\widehat{f_n^{(j)}}(t) = \widehat{f_n^{(j)}}(t_\ell)$ *for* $t_l \le t < t_{l+1}$.

Proof of Theorems 1.1 and 2.1 are similar. Making use of the choice parameters $q_0, \ldots, q_k$ and the recursive form of estimators (1.4), (2.2) one can find the suitable upper bounds for the bias and variance of these estimators. As an illustration, consider the simplest case of the estimation problem (2.1) with $f \in \Sigma(1, L)$ ($\beta = 1$). Then the estimator (2.2) takes the form

$$\widehat{f_n}(t_\ell) = \widehat{f_n}(t_{\ell-1}) + \frac{q_0}{n^{2/3}}(X_\ell - \widehat{f_n}(t_{\ell-1}));\ \widehat{f_n}(0) = f_0 \tag{2.4}$$

with arbitrary bounded f_0 and positive bounded q_0. Making use of (2.1) and notations $\Delta_n(\ell) = \widehat{f_n}(t_\ell) - f(t_\ell)$, $\Delta f(t_\ell) = f(t_{\ell+1}) - f(t_\ell)$, one can rewrite (2.4) as

$$\Delta_n(\ell) = (1 - \frac{q_0}{n^{2/3}})(\Delta_n(\ell-1) - (1 - \frac{q_0}{n^{2/3}})\Delta f(t_{\ell-1}) + \frac{q_0\sigma(t_\ell)\xi_\ell}{n^{2/3}}. \tag{2.5}$$

It follows from (2.5) that

$$\Delta_n(\ell) = (1 - \frac{q_0}{n^{2/3}})^\ell \Delta_n(0) - \sum_{i=0}^{\ell-1}(1 - \frac{q_0}{n^{2/3}})^{\ell-i}\Delta f(t_i) \tag{2.6}$$

$$+\frac{q_0}{n^{2/3}} \sum_{i=0}^{\ell-1}(1 - \frac{q_0}{n^{2/3}})^{\ell-i}\sigma(t_i)\xi_i.$$

It follows from the assumption $\beta = 1$ that $|\Delta f(t_i)| \le L/n$. Thus we have from (2.6) that

$$|E\Delta_n(\ell)| \le |\Delta_n(0)| \exp\{-\frac{q_0\ell}{n^{2/3}}\} + Cn^{-1/3}.$$

So $|E\Delta_n(\ell)| \le Cn^{-1/3}$, as $\ell \ge C_1 n^{2/3} \log n$, or, equivalently, as $t_\ell = l/n \ge C_1 n^{-1/3} \log n$. Analogously one can obtain

$$Var\Delta_n(\ell) \le \frac{C}{n^{4/3}} \sum_{i=0}^{\ell-1}(1 - \frac{q_0}{n^{2/3}})^{2(\ell-i)} \le Cn^{-2/3}.$$

These upper bounds for $|E\Delta_n(\ell)|$, $Var\Delta_n(\ell)$ imply the assertion of the Theorem 2.1 for the case $\beta = 1$. □

3 On-line estimation for the diffusion observed process

Recently we started (together with Y. Golubev) to study the problem of on-line estimation of an unknown signal $S(t)$ for the case of a diffusion observed process. A preliminary result concerns estimating a signal of the smoothness $\alpha, 0 < \alpha \leq 1$ only.

Assume that an observed process is a solution of the stochastic differential equation on $\mathbb{R}^1$

$$dX_\varepsilon(t) = F(t, X_\varepsilon(t), S(t))dt + \varepsilon\sigma(t, X_\varepsilon(t))dw(t); \; X_\varepsilon(0) = x_0. \tag{3.1}$$

(It is a natural generalization of an observation model (1.1)) Here $S(t) : \mathbb{R}^1 \mapsto \mathbb{R}^1$ is an unknown function, and the problem is to estimate this function on the interval $(0, T)$ making use of $X_\varepsilon(t), 0 \leq t \leq T$. Let the following conditions hold:

A1. The functions F, σ are Lipschitzian with respect to all variables, and σ is bounded.

A2. The function $S(t)$ satisfies the Hölder condition

$$|S(t+h) - S(t)| \leq L|h|^\alpha, \; 0 < \alpha \leq 1.$$

A3. For some positive C_i and all $0 \leq t \leq T, x \in \mathbb{R}^1, S \in \mathbb{R}^1$ the inequality $C_1 \leq |\frac{\partial F(t,x,S)}{\partial S}| \leq C_2$ holds.

We consider the following on-line estimator $S_\varepsilon(t)$

$$dS_\varepsilon(t) = \frac{dX_\varepsilon(t) - F(t, X_\varepsilon(t), S_\varepsilon(t))dt}{\gamma_\varepsilon F'_S(t, X_\varepsilon(t), S_\varepsilon(t))}; \; S_\varepsilon(0) = S^{(0)}. \tag{3.2}$$

Theorem 3.1 *Under conditions A1 – A3 the estimator (3.2) with* $\gamma_\varepsilon = k\varepsilon^{\frac{2}{2\alpha+1}}$ *(k is an arbitrary positive constant) has the property*

$$E|S_\varepsilon(t) - S(t)|^2 \leq C\varepsilon^{\frac{4\alpha}{2\alpha+1}} \tag{3.3}$$

as $t > C\varepsilon^{\frac{2}{2\alpha+1}} \log(1/\varepsilon)$ *(here C is large enough, but independent of* ε*).*

Proof. Introduce $S_\delta(t) = (2\delta)^{-1} \int_{\mathbb{R}^1} \exp\{-\frac{|t-u|}{\delta}\} S(u)du$. It is easy to see from A2 that

$$|S_\delta(t) - S(t)| \leq c_3\delta^\alpha, \; |S'_\delta(t)| \leq c_3\delta^{\alpha-1}. \tag{3.4}$$

Introduce a new process $x_\delta(t) = S_\varepsilon(t) - S_\delta(t)$. Then we have from (3.1) and (3.2)

$$dx_\delta(t) = \frac{1}{\gamma_\varepsilon}\Delta_\varepsilon(t)dt + \frac{\varepsilon\sigma(t, X_\varepsilon(t))}{\gamma_\varepsilon F'_S(t, X_\varepsilon(t), S_\varepsilon(t))}dw(t) - S'_\delta(t)dt. \tag{3.5}$$

Here we denote

$$\Delta_\varepsilon(t) = \frac{F(t, X_\varepsilon(t), S(t)) - F(t, X_\varepsilon(t), S_\delta(t) + x_\delta(t))}{F'_S(t, X_\varepsilon(t), S_\delta(t) + x_\delta(t))} \tag{3.6}$$

The equation (3.5) and Ito formula imply

$$d[x_\delta(t)]^2 = \frac{2}{\gamma_\varepsilon} x_\delta(t)\Delta_\varepsilon(t)dt + 2\frac{\varepsilon x_\delta(t)\sigma(t, X_\varepsilon(t))}{\gamma_\varepsilon F'_S(t, X_\varepsilon(t), S_\varepsilon(t))}dw(t) \tag{3.7}$$
$$+[\frac{\varepsilon\sigma(t, X_\varepsilon(t))}{\gamma_\varepsilon F'_S(t, X_\varepsilon(t), S_\varepsilon(t))}]^2 dt - 2x_\delta(t)S'_\delta(t)dt.$$

It follows from A3 and (3.4) that

$$x_\delta(t)\Delta_\varepsilon(t) = x_\delta(t)\frac{F(t, X_\varepsilon(t), S(t)) - F(t, X_\varepsilon(t), S_\delta(t))}{F'_S(t, X_\varepsilon(t), S_\delta(t) + x_\delta(t))}$$
$$+x_\delta(t)\frac{F(t, X_\varepsilon(t), S_\delta(t)) - F(t, X_\varepsilon(t), S_\delta(t) + x_\delta(t))}{F'_S(t, X_\varepsilon(t), S_\delta(t) + x_\delta(t))}$$
$$\leq \frac{C_2}{C_1}|x_\delta(t)||S(t) - S_\delta(t)| - \frac{C_1}{C_2}|x_\delta(t)|^2 \leq C_4|x_\delta(t)|\delta^\alpha - \frac{C_1}{C_2}|x_\delta(t)|^2. \tag{3.8}$$

Denote $V_\delta(t) = E[x_\delta(t)]^2$. Then it is clear from (3.7), (3.4) and (3.8) that

$$V'_\delta(t) \leq -\frac{k_1}{\gamma_\varepsilon}V_\delta(t) + k_2(\frac{\delta^{2\alpha}}{\gamma_\varepsilon} + \frac{\varepsilon^2}{\gamma_\varepsilon^2} + \gamma_\varepsilon\delta^{2\alpha-2}) \tag{3.9}$$

for small enough positive constant k_1 and large enough constant k_2 (both independent of $\varepsilon, \delta, \gamma_\varepsilon$). Now choose the parameters $\delta, \gamma_\varepsilon$ as $\delta \asymp \varepsilon^{\frac{2}{2\alpha+1}}$, $\gamma_\varepsilon \asymp \varepsilon^{\frac{2}{2\alpha+1}}$. Then we obtain from (3.9) for some positive constants k_3, k_4 independent of ε the inequality

$$V'_\delta(t) \leq -k_3\varepsilon^{-\frac{2}{2\alpha+1}}V_\delta(t) + k_4\varepsilon^{\frac{4\alpha-2}{2\alpha+1}} \tag{3.10}$$

It follows from (3.10) that

$$V_\delta(t) \leq V_\delta(0)\exp\{-k_3\varepsilon^{-\frac{2}{2\alpha+1}}t\} + \frac{k_4}{k_3}\varepsilon^{\frac{4\alpha}{2\alpha+1}}. \tag{3.11}$$

The initial value $V_\delta(0) = S^{(0)} - S_\delta(0)$ is bounded. Thus we can conclude from (3.11) that $V_\delta(t) \leq C\varepsilon^{\frac{4\alpha}{2\alpha+1}}$, as $t > C\varepsilon^{\frac{2}{2\alpha+1}}\log(1/\varepsilon)$. Note now that $E|S_\varepsilon(t) - S(t)|^2 \leq 2V_\delta(t) + 2|S_\delta(t) - S(t)|^2$. The theorem follows from these upper bounds and (3.4). □

Remark 3.1 *It follows from the s.1 that the rate of convergence in Theorem 3.1 is unimprovable: it is unimprovable even for the case* $F(t, x, S) = S, \sigma(t, x) = 1$.

5 Concluding remark

The estimators (1.4), (2.2), (3.2) can be used for extrapolation too. For instance, the expression

$$\overline{f_n(t_l + h)} = \sum_{j=0}^{k} \frac{h^j}{j!} \widehat{f_n}(t_l)$$

can be used for estimation of $f(t_l + h)$ on the basis of observations $X_{1n}, \dots, X_{ln}$. It is not hard to check with help of Theorem 2.1 that

$$E|\overline{f_n(t_l + h)} - f(t_l + h)|^2 \le C \max\{h, n^{-\frac{1}{2\beta+1}}\}^{2\beta},$$

and better rate of convergence of risk is unattainable.

References

Chow, P.-L., Khasminskii, R. & Liptser, R.Sh. (1997). Tracking of signal and its derivatives in Gaussian white noise, *Stochastic Processes Appl.*, 69, 2 259-273.

Chow, P.-L., Khasminskii, R. & Liptser, R.Sh. (2001). On estimation of time dependent spatial signal in Gaussian white noise, *Stochastic Processes Appl.*, 96, 161-175.

Ibragimov, I. & Khasminskii, R. (1980a). Estimation of a signal, its derivatives, and point of maximum for Gaussian observations, *Theory Probability Appl.*, 25, 703-720.

Ibragimov, I. & Khasminskii, R. (1980b). On nonparametric estimation of regression, *Soviet Math. Dokl.*, 21, 810-814.

Ibragimov, I. & Khasminskii, R. (1981). *Statistical estimation: Asymptotic Theory.* Springer Verlag (Russian ed. 1979).

Khasminskii, R. & Liptser, R. (2002). Tracking of a smooth regression function, *Theory Probability Appl.*, 47, 3, 567-575.

Stone, C. (1980). Optimal rates of convergence for nonparametric estimators, *Ann. Statist.*, 8, 1348-1360.

Resum

Es presenta un recull d'alguns resultats recents en estimació no-paramètrica en línia. El primer resultat tracta d'una estimació en línia per a un senyal suau S(t) en el model clàssic "senyal més soroll blanc Gaussià (GWN)". Aleshores es descriu i justifica un estimador en línia anàleg pel problema d'estimació de regressió amb disseny equidistant. Finalment, es descriuen alguns resultats preliminars en relació a l'estimació en línia pel procés de difusió observada.

MSC: 62G05, 62G08, 62M05.

Paraules clau: soroll blanc Gaussià, estimació en línia, filtres no linials

SORT 28 (1) January-June 2004, 9-26

Asymptotic normality of the integrated square error of a density estimator in the convolution model

C. Butucea*

Université de Paris

Abstract

In this paper we consider a kernel estimator of a density in a convolution model and give a central limit theorem for its integrated square error (ISE). The kernel estimator is rather classical in minimax theory when the underlying density is recovered from noisy observations. The kernel is fixed and depends heavily on the distribution of the noise, supposed entirely known. The bandwidth is not fixed, the results hold for any sequence of bandwidths decreasing to 0. In particular the central limit theorem holds for the bandwidth minimizing the mean integrated square error (MISE). Rates of convergence are sensibly different in the case of regular noise and of super-regular noise. The smoothness of the underlying unknown density is relevant for the evaluation of the MISE.

MSC: 62G05, 62G20

Keywords: convolution density estimation, nonparametric density estimation, central limit theorem, integrated squared error, noisy observations

1 Introduction

In this paper we consider the following convolution model:

$$Z_i = X_i + e_i,$$

where $X_i,\ i = 1, \dots, n$ are i.i.d. random variables of unknown density f which we need to recover from noisy observations $Y_i,\ i = 1, \dots, n$. The noise variables e_i are supposed i.i.d. of known fixed distribution, having a density function η in L_1 and L_2 and a characteristic function (c. f.) Φ^η.

* *Address for correspondence:* Modal'X, Université Paris X, Bât. G, F92001 Nanterre Cedex. LPMA, Université Paris VI, 175, rue du Chevaleret, F75013 Paris. butucea@ccr.jussieu.fr.

Received: October 2003
Accepted: March 2004

We suggest here an estimator f_n of f from noisy observations and study the asymptotic normality of its integrated square error (ISE)

$$ISE(f_n, f) = \int (f_n(x) - f(x))^2 dx. \tag{1}$$

Let us suppose for the beginning that f belongs to a Sobolev class $W(r, L)$ of densities, i.e.

$$W(r, L) = \left\{ f \text{ density} : f \in L_2, \quad \int |\Phi(u)|^2 |u|^{2r} du \leq 2\pi L \right\}$$

where $\Phi(u) = \int \exp(iux) f(x) dx$ denotes its Fourier transform, for some fixed $r > 1/2$ and a constant $L > 0$. This roughly means these densities are continuously derivable up to order r and their r-th derivative has bounded L_2 norm.

It is known from estimation theory in the convolution model, that the rates and behaviours of estimators are sensibly different if the characteristic function of the noise decreases polynomially or exponentially asymptotically. We suppose in a first part that the noise is "polynomial", i.e.

$$|\Phi^\eta(u)| \sim |u|^{-s}, \text{ as } |u| \to \infty,$$

where $\sim$ means that the functions behave similarly and $s > 0$ such that $r > s$.

Let us denote $g = f \star \eta$ the common density of $Y_i, i = 1, \dots, n$ and $\Phi^g = \Phi \cdot \Phi^\eta$ its Fourier transform.

In Section 3, we state our results for different setups. In Section 3.1 we consider classes of supersmooth densities in association with polynomial noise. We say that f is a supersmooth density if f belongs to the class

$$S(\alpha, r, L) = \left\{ f \text{ density} : f \in L_2, \quad \int |\Phi(u)|^2 \exp(2\alpha |u|^r) du \leq 2\pi L \right\},$$

for some $\alpha, r, L > 0$. In Section 3.2 we consider Sobolev densities in association with exponentially decreasing noise. Exponential noise means

$$|\Phi^\eta(u)| \sim \exp(-\gamma |u|^s), \text{ as } |u| \to \infty,$$

where $\gamma, s > 0$. We work here with a kernel estimator of the deconvolution density

$$f_n(x) = \frac{1}{n} \sum_{i=1}^n K_h^n (x - Y_i), \tag{2}$$

where $h > 0$ is small, K_h^n denotes $K^n(\cdot/h)/h$ and the kernel K^n is defined via its Fourier transform

$$\Phi^{K^n}(u) = \frac{\Phi^K(u)}{\Phi^\eta(u/h)}, \text{ where } \Phi^K(u) = I[|u| \leq 1]. \tag{3}$$

Since pioneering work by Carroll and Hall (1988), the deconvolution density was already estimated in many setups. We shall cite here only works very much related to our framework and problems. Such kernel estimates were used on classes similar to the Sobolev class by Fan (1991*a*), who computed the rates of convergence of the minimax L_2 risk. Recently wavelet estimators were proven to attain the same rates on Besov bodies and these rates are known to be optimal in the minimax approach, see Fan and Koo (2002).

In the setup of Sobolev densities, Goldenshluger (1999) generalized the minimax rate for estimating f with pointwise risk to adaptive (to the Sobolev smoothness) rates when the noise is either polynomial or exponential (without loss of rate in this last case). Efromovich (1997) computed exact asymptotic risks (pointwise and in L_2 norm) for estimating Sobolev densities in the presence of exponentially decreasing noise.

The kernel estimator in (2) (with adequate bandwidth) was proven to be minimax for estimating supersmooth densities with polynomial noise in Butucea (2004) and with exponential noise in Butucea and Tsybakov (2003). The same kernel estimator was proven asymptotically normal when the noise is either polynomial or exponential in Fan (1991*b*) and Fan and Liu (1997).

Here we study the asymptotic normality of the ISE in (1) and will discuss several important applications of results issued from these computations. Such computations can be found in Hall (1984) for a nonparametric density estimator with direct observations. His study is a direct application of a Central Limit Theorem of degenerate U-statistics of second order. He motivates this by the practical use in simulations of ISE as a measure of the performance of a density estimator. The main goal is to evaluate c_n and σ_n such that

$$\sigma_n^{-1}(ISE(f_n, f) - c_n) \to N(0, 1),$$

when $h \to 0$ and $n \to \infty$. This subject is strongly related to estimating the L_2 norm of the density f from noisy observations. Indeed, a natural estimator d_n^2 of $||f||_2^2$ can be decomposed such that one of the terms is the degenerate second order U-statistic S_2 defined later in (8). For not too smooth densities S_2 is the dominating term and this gives the rate of estimating $||f||_2^2$. Estimating the L_2 norm of a density is furthermore useful for nonparametric testing in the convolution model. These problems will be soon the subject of scientific communications.

Another related problem can be further investigated starting with these calculations, namely that of bandwidth selection for the kernel deconvolution density estimator f_n in (2), via cross-validation.

2 Results

As a first step it is natural to replace c_n by $E_f[ISE(f_n, f)]$ also denoted by $MISE(f_n, f)$ for mean integrated square error. From now on P_f, E_f, and V_f denote the probability, the expectation and the variance when the true underlying density of the model is f. We may use constants $c, C, C', \dots$ which are different throughout the whole proof.

Note that the density of our observations is $g = f \star \eta$. We note next that

$$\begin{aligned} ISE(f_n, f) &= \int (f_n(x) - E_f[f_n(x)] + E_f[f_n(x)] - f(x))^2 dx \\ &= \int (f_n(x) - E_f[f_n(x)])^2 dx + \int (E_f[f_n(x)] - f(x))^2 dx. \end{aligned}$$

Indeed, the cross product term is null, see Lemma 2. We replace from now on $E_f[f_n(x)]$ by its value $K_h \star f$. Then

$$MISE(f_n, f) = E_f[ISE(f_n, f))] = E_f\left[\int (f_n(x) - E_f[f_n(x)])^2 dx\right] + \int (E_f[f_n(x)] - f(x))^2 dx$$

and we write

$$ISE(f_n, f) - E_f[ISE(f_n, f))] = I_n - E_f[I_n],$$

where $I_n = \int (f_n(x) - E_f[f_n(x)])^2 dx$. Computation of $E_f[I_n]$ and of the bias $B^2(f_n) = \int (E_f[f_n(x)] - f(x))^2 dx$ is rather classical in minimax theory.

Lemma 1 *Let $f_n(\cdot, Y_1, \dots, Y_n)$ be the kernel density estimator defined in (2) based on the noisy observations in our convolution model with a bandwidth $h \to 0$ when $n \to \infty$. Then*

$$E_f[I_n] = \frac{1 + o(1)}{\pi(2s + 1)nh^{2s+1}}.$$

If the underlying density belongs to a Sobolev smoothness class $W(r, L)$ with $r > 1/2$, then

$$\sup_{f \in W(r,L)} B^2(f_n) = \sup_{f \in W(r,L)} \int \left(E_f[f_n(x)] - f(x)\right)^2 dx = Lh^{2r} = o(1).$$

In conclusion, $MISE(f_n, f)$ converges to 0, if and only if $nh^{2s+1} \to \infty$ when $n \to \infty$ and the bandwidth minimizing $\sup_{f \in W(r,L)} MISE(f_n, f)$ is

$$h_{MISE} = (L\pi(2s + 1)n)^{-\frac{1}{2(r+s)+1}}.$$

Proof. We present here only exact calculation of $E_f[I_n]$, since the remaining results are obvious or not entirely new. We have

$$\begin{aligned} E_f[I_n] &= \frac{1}{n} \int \left(\int \left(K_h^n(x - y) - K_h \star f\right)^2 (x) dx \right) g(y) dy \\ &= \frac{1}{n} \left(\int \left(\int \left(K_h^n(x - y)\right)^2 dx \right) g(y) dy - \|K_h \star f\|_2^2 \right). \end{aligned}$$

We know that $\|K_h \star f\|_2^2$ is equal to $\|f\|_2^2$ plus some estimation bias which tends to 0 when $h \to 0$ on a smoothness class like the Sobolev class, $W(r, L)$. So, the main term is $\int\left(\int\left(K_h^n(x-y)\right)^2 dx\right) g(y)dy$. Use Lemma 2:

$$\begin{aligned}
\int\left(\int\left(K_h^n(x-y)\right)^2 dx\right) g(y)dy &= \frac{1}{h}\int (K^n)_h^2 \star g(x)dx = \frac{1}{2\pi h}\Phi^{(K^n)_h^2\star g}(0)\\
&= \frac{1}{2\pi h}\Phi^g(0)\Phi^{(K^n)^2}(0) = \frac{1}{2\pi h}\int \Phi^{K^n}(-u)\Phi^{K^n}(u)du\\
&= \frac{1+o(1)}{\pi(2s+1)h^{2s+1}}.
\end{aligned}$$

□

Remark that in previous equations and in the following proofs, we compute integrals like $\int(\Phi^{K^n})^2$ by actually replacing the c. f. of the noise by $|u|^{-s}$, its asymptotic expression. We do this for simplicity, since calculation would actually need splitting integration domain into $|u| \leq M$ and $M < |u| < 1/h$, for some large enough, but fixed $M > 0$. If M is large enough, Φ^η is almost $|u|^{-s}$ and the second integral is always dominating over the first and gives the order of the whole expression. For a complete and explicit computation of $\|K_n\|_2^2$ see Butucea (2004).

Let us look closer at I_n:

$$\begin{aligned}
I_n &= \frac{1}{n^2}\int\left(\sum_{i=1}^n (K_h^n(x-Y_i) - K_h \star f(x))\right)^2 dx\\
&= \frac{1}{n^2}\sum_{i=1}^n \|K_h^n(\cdot - Y_i) - K_h \star f\|_2^2 + \frac{1}{n^2}\sum_{i\neq j=1}^n \langle K_h^n(\cdot - Y_i) - K_h \star f, K_h^n(\cdot - Y_j) - K_h \star f\rangle,
\end{aligned}$$

where $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the L_2 norm and the scalar product in L_2, respectively. If we denote by

$$U_i = U_i(x, h, Y_i) = K_h^n(x - Y_i) - K_h \star f(x), \tag{4}$$

these variables are centred and independent. We get

$$\begin{aligned}
I_n - E_f[I_n] &= \frac{1}{n^2}\sum_{i=1}^n \left(\|U_i\|_2^2 - E_f[\|U_i\|_2^2]\right) + \frac{1}{n^2}\sum_{i\neq j=1}^n \langle U_i, U_j\rangle\\
&= S_1 + S_2, \text{ say.}
\end{aligned}$$

It is easy to see that variables in S_1 and in S_2 are uncorrelated:

$$E_f[(\|U_k\|_2^2 - E_f[\|U_k\|_2^2])(\langle U_i, U_j\rangle)] = 0,$$

for all $k, i, j = 1, \ldots, n$ and $i \neq j$. It is necessary now to compute the variance of each sum and compare. What we prove in the following is that S_2 has a larger variance (in order) than S_1, for any $h \to 0$ and $n \to \infty$. Then we prove its asymptotic normality and deduce the asymptotic normality of $ISE(f_n, f) - E_f[ISE(f_n, f)]$. The main difficulty comes from the fact that S_2 is an U-statistic of order 2 and degenerate. Indeed,

$$\begin{aligned} E_f[\langle U_i, U_j \rangle / Y_j = y_j] &= E_f[\langle K_h^n(\cdot - Y_i) - K_h \star f, K_h^n(\cdot - y_j) - K_h \star f \rangle] \\ &= \langle E_f[K_h^n(\cdot - Y_i)] - K_h \star f, K_h^n(\cdot - y_j) - K_h \star f \rangle = 0. \end{aligned}$$

Nevertheless, each term of the sum depends on n and we apply a central limit theorem for degenerate U-statistics by Hall (1984), which he already applied in his paper for the ISE of a nonparametric estimator with direct observations. Here, we have noisy observations and a particular choice of the kernel (motivated by the minimax theory in this field) giving sensibly different asymptotic behaviours and rates.

Theorem 1 *Let $f_n(\cdot, Y_1, \dots, Y_n)$ be the kernel density estimator defined in (2) based on the noisy observations in our convolution model and a bandwidth $h \to 0$ such that $nh^{2s+1} \to \infty$, when $n \to \infty$. Then*

$$\sqrt{\frac{\pi(4s+1)n^2 h^{4s+1}}{2\|g\|_2^2}} \left(ISE(f_n, f) - E_f[ISE(f_n, f)] \right) \to N(0,1)$$

where the convergence is in law when $n \to \infty$.

Corollary 2 *Let $f_n(\cdot, Y_1, \dots, Y_n)$ be the kernel density estimator in (2) based on the noisy observations with noise having polynomially decreasing Fourier transform and a bandwidth $h \to 0$ such that $nh^{2s+1} \to \infty$, when $n \to \infty$. Then I_n is asymptotically normally distributed with*

$$E_f[I_n] = \frac{1 + o(1)}{\pi(2s+1)nh^{2s+1}} \text{ and } V_f[I_n] = \frac{2\|g\|_2^2(1 + o(1))}{\pi(4s+1)n^2 h^{4s+1}}; \tag{5}$$

if f belongs to the Sobolev class $W(r, L)$, the integrated square error $ISE(f_n, f)$ is asymptotically normally distributed with

$$MISE(f_n, f) \le Lh^{2r} + \frac{1 + o(1)}{\pi(2s+1)nh^{2s+1}} \text{ and } V_f[ISE(f_n, f)] = \frac{2\|g\|_2^2(1 + o(1))}{\pi(4s+1)n^2 h^{4s+1}}$$

and the $MISE$ becomes minimal (and of the order of the minimax L_2 risk) for $h_* = (L\pi(2s+1)n)^{1/(2(r+s)+1)}$

$$\inf_{h>0} \sup_{f \in W(r,L)} MISE(f_n, f) = L^{\frac{1}{2(r+s)+1}} (\pi(2s+1)n)^{-\frac{2r}{2(r+s)+1}}.$$

Notice that for constructing a confidence interval of $ISE(f_n, f)$ using its asymptotic normality, both $MISE(f_n, f)$ and $V_f[ISE(f_n, f)]$ still depend on unknown quantities. This was already noted by Hall (1984). The mean of $ISE(f_n, f)$ depends on unknown f

via the bias of f_n: $B^2(f_n) = \|E_f[f_n] - f\|_2^2$ that we can bound from above by Lh^{2r}. The variance of $ISE(f_n, f)$ depends on unknown $\|g\|_2^2$. Nevertheless, g is the density of our observations and can be directly evaluated at a faster rate than f (the same holds for the other frameworks). Indeed, not only we have direct observations, moreover, g is more regular than f due to the convolution (which adds smoothness). The estimation of the L_2 norm of a regular enough density, having a smoothness $> 1/4$, can be done efficiently at rate $1/\sqrt{n}$, see e.g. Laurent (1996).

Note also that if we use another bandwidth h satisfying $nh^{2r+2s+1} \to \infty$, when $n \to \infty$, the associated $MISE$ is $(1 + o(1))/(\pi(2s + 1)nh^{2s+1})$. Indeed, whatever the bias of the estimator f_n is, it is smaller than $Lh^{2r} = o(1/(nh^{2s+1}))$. In this case, the confidence interval $IC_{1-\delta}$ of risk $\delta > 0$, writes

$$IC_{1-\delta} = \left[\frac{1}{\pi(2s+1)nh^{2s+1}} \pm z_{1-\delta/2} \frac{\|g\|_2}{nh^{2s+1/2}} \sqrt{\frac{2}{\pi(4s+1)}} \right], \qquad (6)$$

where z_δ is the δ-quantile of $N(0, 1)$, a gaussian law.

Proof. **Convergence of S_1**

$$S_1 = \frac{1}{n^2} \sum_{i=1}^{n} \left(\|U_i\|_2^2 - E_f[\|U_i\|_2^2] \right).$$

Let us compute an upper bound of the variance of S_1. We have

$$\begin{aligned} V_f[S_1] &= \frac{1}{n^4} \sum_{i=1}^{n} E_f \left[\left(\|U_i\|_2^2 - E_f[\|U_i\|_2^2] \right)^2 \right] \\ &= \frac{1}{n^3} \left(E_f[\|U_1\|_2^4] - \left(E_f[\|U_1\|_2^2] \right)^2 \right) \leq \frac{E_f[\|U_1\|_2^4]}{n^3}. \end{aligned}$$

In order to evaluate an upper bound of this, we develop the square of sums in $E_f[\|U_1\|_2^4]$ and conclude by saying that the dominant term is given by one of positive terms (this expectation being a positive real number):

$$\begin{aligned} E_f[\|U_1\|_2^4] &= \int \left(\int \left(K_h^n(x-y) - K_h \star f(x) \right)^2 dx \right)^2 g(y)dy \\ &= \int \left(\int \left(K_h^n(x-y) \right)^2 dx \right)^2 g(y)dy \\ &\quad +2\|K_h \star f\|_2^2 \int \left(\int \left(K_h^n(x-y) \right)^2 dx \right) g(y)dy \\ &\quad +4 \int \left(\int K_h^n(x-y) K_h \star f(x) dx \right)^2 g(y)dy. \end{aligned}$$

Note that, by Cauchy-Schwarz and previous evaluations:

$$\begin{aligned}
&\int \left(\int K_h^n(x-y) K_h \star f(x) dx \right)^2 g(y) dy \\
\leq\ & \int \left(\int \left(K_h^n(x-y) \right)^2 dx \right)^{1/2} \|K_h \star f\|_2\, g(y) dy \leq \frac{O(1)}{h^{2s+1}}.
\end{aligned}$$

It remains to compute an asymptotic upper bound of $\int \left(\int \left(K_h^n(x-y) \right)^2 dx \right)^2 g(y)dy$. As previously,

$$\int \left(\int \left(K_h^n(x-y) \right)^2 dx \right)^2 g(y)dy \leq \frac{C}{h^2} \|K^n\|_2^4 \leq \frac{c}{h^{4s+2}}.$$

Then, for all $h > 0$ small such that $nh^{2s+1} \to \infty$,

$$V_f \left[\sqrt{\frac{\pi(4s+1)n^2 h^{4s+1}}{2\|g\|_2^2}} S_1 \right] \leq \frac{C}{nh} = o(1), \text{ when } n \to \infty \tag{7}$$

and then

$$\sqrt{\frac{\pi(4s+1)n^2 h^{4s+1}}{2\|g\|_2^2}} S_1 \to_P 0, \text{ when } n \to \infty.$$

Convergence of S_2:

$$S_2 = \frac{1}{n^2} \sum_{i \neq j = 1}^{n} \langle U_i, U_j \rangle. \tag{8}$$

The variables in S_2 are centred and, moreover, $E_f[\langle U_i, U_j \rangle \langle U_k, U_l \rangle] = 0$ as soon as $(i,j) \neq (k,l)$ and $(i,j) \neq (l,k)$. Then

$$V_f[S_2] = \frac{1}{n^4} E_f \left[\left(\sum_{i \neq j = 1}^{n} \langle U_i, U_j \rangle \right)^2 \right] = \frac{2}{n^4} n(n-1) E_f[\langle U_1, U_2 \rangle^2] = \frac{2 + o(1)}{n^2} E_f[\langle U_1, U_2 \rangle^2]$$

If we develop this, we get

$$E_f[\langle U_1, U_2 \rangle^2] = E_f[\langle K_h^n(x - Y_1), K_h^n(x - Y_2) \rangle^2] - \|K_h \star f\|_2^4.$$

We use again the fact that $\|K_h \star f\|_2^2$ is equal to $\|f\|_2^2$ plus some estimation bias which tends to 0 when $h \to 0$ on the class $W(r, L)$. So, the main term is the first one. Indeed:

$$\begin{aligned}
E_f[\langle K_h^n(\cdot - Y_1), K_h^n(\cdot - Y_2) \rangle^2] &= \int \int \left(\int K_h^n(x-u) K_h^n(x-v) dx \right)^2 g(u) g(v) du dv \\
&= \frac{1}{h} \int \int (M^n)_h^2 (v-u) g(u) g(v) du dv,
\end{aligned}$$

where we put $M^n(x) = \int K^n(z+x)K^n(z)dz$. Note that

$$\begin{aligned}\int (M^n(x))^2 dx &= \frac{1}{2\pi}\int \left|\Phi^{\langle K^n(x+\cdot),K^n(\cdot)\rangle}(u)\right|^2 du = \frac{1}{2\pi}\int \left|\Phi^{K^n}(u)\Phi^{K^n}(-u)\right|^2 du \\ &= \frac{1}{2\pi}\int_{|u|\le 1}\frac{du}{|\Phi^\eta(u/h)|^2\,|\Phi^\eta(-u/h)|^2} = \frac{1+o(1)}{\pi(4s+1)h^{4s}}.\end{aligned}$$

Since densities g are continuous functions, even $(r+s-1/2)$ - Lipschitz continuous, see Lemma 3, they are uniformly bounded over f in the Sobolev class $W(r,L)$ with any noise density η under our assumptions. Then for any small $\epsilon > 0$, such that $\epsilon/h \to \infty$, when $n \to \infty$:

$$\begin{aligned}&\left|\int\int (M^n)_h^2(v-u)g(u)g(v)dudv - \int (M^n)^2\|g\|_2^2\right| \\ =\ &\left|\int\int\left((M^n)_h^2(v-u)g(u) - g(v)\int (M^n)^2\right)dug(v)dv\right| \\ \le\ &\int\left|\int (M^n)^2(x)(g(v+hx)-g(v))dx\right|g(v)dv \\ \le\ &\int_{|hx|\le\epsilon}(M^n)^2(x)|hx|^{r+s-1/2}dx + 2\sup_{f,\eta}\|g\|_\infty\int_{|hx|>\epsilon}(M^n)^2(x)dx \le o\left(\int (M^n)^2\right).\end{aligned}$$

This means

$$E_f[\langle K_h^n(x-Y_1), K_h^n(x-Y_2)\rangle^2] = \frac{1+o(1)}{\pi(4s+1)h^{4s+1}}\|g\|_2^2$$

which implies that

$$V_f[S_2] = \frac{(2+o(1))\|g\|_2^2}{\pi(4s+1)n^2h^{(4s+1)}}. \tag{9}$$

Asymptotic normality of S_2. We apply here the following Proposition by Hall (1984):

Proposition 1 (see Theorem 1, Hall (1984)) *Assume $H_n(x,y)$ is a symmetric function such that $E[H_n(X_1,X_2)/X_1] = 0$ almost surely and $E[H_n^2(X_1,X_2)] < \infty$ for each n. Denote by*

$$G_n(x,y) = E[H_n(X_1,x)H_n(X_1,y)].$$

If

$$\left(E[G_n^2(X_1,X_2)] + n^{-1}E[H_n^4(X_1,X_2)]\right)/\left(E[H_n^2(X_1,X_2)]\right)^2 \to 0, \tag{10}$$

as $n \to \infty$, then

$$W_n \equiv \sum_{i<j=1}^{n} H_n(X_i,X_j)$$

is asymptotically normally distributed with zero mean and variance $n^2E[H_n^2(X_1,X_2)]/2$.

We apply this result to

$$n^2 S_2/2 = \sum_{i<j=1}^{n} \langle U_i, U_j \rangle.$$

We have seen already that this U-statistic is degenerate and that

$$E_f[\langle U_1, U_2 \rangle^2] = \frac{\|g\|_2^2 + o(1)}{\pi(4s+1)h^{4s+1}} < \infty.$$

In order to check (10) we evaluate and bound from above $E_f[G_n^2(Y_1, Y_2)]$ and $E_f[\langle U_1, U_2 \rangle^4]$. First, if we replace U_1 and U_2 and we keep the dominant term in the expectation:

$$\begin{aligned} E_f[\langle U_1, U_2 \rangle^4] &\leq \int \left(\int \frac{1}{h^2} K^n\left(\frac{u-y_1}{h}\right) K^n\left(\frac{u-y_2}{h}\right)\right)^4 g(y_1)g(y_2)dy_1dy_2 \\ &\leq \frac{1}{h^3} \int \frac{1}{h} \left(K^n\left(z + \frac{y_2 - y_1}{h}\right) K^n(z)dz \right)^4 g(y_1)g(y_2)dy_1dy_2 \\ &\leq \frac{1}{h^3} \int R_h^n(y_2 - y_1)g(y_1)g(y_2)dy_1dy_2, \end{aligned}$$

where $R^n(z) = (\int K^n(z+u)K^n(u)du)^4 = (M^n(z))^4$. As in the previous part of this proof, we need to evaluate

$$\int R^n(z)dz = \int (M^n)^4(z)dz = \frac{1}{2\pi} \int \left|\Phi^{M^n} \star \Phi^{M^n}(u)\right|^2 du \leq \left(\int \left|\Phi^{M^n}(u)\right|^2 du \right)^2 \leq \frac{c}{h^{8s}}.$$

Thus,

$$E_f[\langle U_1, U_2 \rangle^4] / \left(n(E_f[\langle U_1, U_2 \rangle^2])^2 \right) \leq \frac{c/h^{8s+3}}{n/h^{8s+2}} \leq \frac{C'}{nh} = o(1) \tag{11}$$

and this proves the first part of (10).

Now, recall (4) and write

$$G_n(y_1, y_2) = \int \langle U_1(\cdot, h, y_1), U_3(\cdot, h, y_3) \rangle \langle U_2(\cdot, h, y_2), U_3(\cdot, h, y_3) \rangle g(y_3)dy_3.$$

We have

$$\begin{aligned} \langle U_1(\cdot, h, y_1), U_3(\cdot, h, y_3) \rangle &= \int \frac{1}{h^2} K^n\left(\frac{u-y_1}{h}\right) K^n\left(\frac{u-y_3}{h}\right) du \\ &\quad -\frac{1}{h} \int K_h \star f(u) \left[K^n\left(\frac{u-y_1}{h}\right) + K^n\left(\frac{u-y_3}{h}\right) \right] du \\ &\quad + \|K_h \star f\|_2^2 \end{aligned}$$

By changing the variable, the first term on the right-hand side becomes

$$\frac{1}{h} \int K^n\left(u + \frac{y_3 - y_1}{h}\right) K(u)du = M_h^n(y_3 - y_1),$$

where again $M^n(z) = \int K^n(u+z)K^n(u)du$. Then, when we replace this into $E_f[G_n^2(Y_1,Y_2)]$, we keep only the dominant term:

$$\begin{aligned}
E_f[G_n^2(Y_1,Y_2)] &\leq \int\int\left(\int M_h^n(y_3-y_1)M_h^n(y_3-y_2)g(y_3)dy_3\right)^2 g(y_1)g(y_2)dy_1dy_2\\
&\leq \frac{1}{h}\int\int\frac{1}{h}\left(M^n\left(z+\frac{y_2-y_1}{h}\right)M^n(z)g(y_2+hz)dz\right)^2 g(y_1)g(y_2)dy_1dy_2\\
&\leq \frac{1}{h}\int\int\frac{1}{h}\int(M^n)^2\left(z+\frac{y_2-y_1}{h}\right)(M^n)^2(z)g(y_2+hz)dzg(y_1)g(y_2)dy_1dy_2\\
&\leq C\frac{1}{h}\int\int Q_h^n(y_2-y_1)g(y_1)g(y_2)dy_1dy_2,
\end{aligned}$$

where we used Jensen inequality, the fact that densities g are uniformly bounded by a constant C depending only on r, s, L. We denoted by

$$Q^n(z) = (\int(M^n)^2(z+x)(M^n)^2(x)dx)^2.$$

Similarly to previous calculation of $E_f[\langle U_1,U_2\rangle^2]$

$$\int Q^n(z)dz = \int\int(M^n)^2(z+x)(M^n)^2(x)dxdz = \left(\int(M^n)^2(x)dx\right)^2 \leq \frac{C''}{h^{8s}}.$$

Thus,

$$E_f[G_n^2(Y_1,Y_2)]/(E_f[\langle U_1,U_2\rangle^2])^2 \leq C'''h = o(1). \tag{12}$$

Inequalities (11) and (12) imply verification of (10) and the proof of asymptotic normality. Thus, together with (9), we get the theorem: $ISE(f_n,f) - MISE(f_n,f)$ is asymptotically normally distributed with mean 0 and variance $2\|g\|_2^2/(\pi(4s+1)n^2h^{4s+1})$. If we take in consideration Lemma 1, plus simple computations, we get the Corollary.

□

3 Other frameworks

We study here the same problem in the framework of supersmooth densities observed with polynomial noise (Section 4.1) and that of Sobolev densities with exponential noise (Section 4.2). As it is known from deconvolution density estimation, the bandwidth minimizing $MISE$ provides much slower rates for smoother noise distribution. Smoother is the noise, harder is the deconvolution problem and slower is the convergence rate to the asymptotic gaussian law.

3.1 Supersmooth densities and polynomial noise

In the previous context, condition $nh^{2s+1} \to \infty$ was necessary to ensure consistency of the $MISE$, but we only need the more classical, less restrictive condition $nh \to \infty$ in order to have S_1 converging in probability to 0 (see (7)) and for the asymptotic normality of the ISE, see (11) necessary to get (10). The fact that f was in the Sobolev class allowed us to evaluate the bias term in $MISE$ and to minimize over $h > 0$ the $MISE$. If we consider instead of Sobolev smoothness classes, a class $S(\alpha, r, L)$ of supersmooth densities f as defined in the Introduction. We know (see Butucea (2004)) that

$$B^2(f_n) = \int \left(E_f[f_n(x)] - f(x)\right)^2 du \le L\exp\left(-\frac{2\alpha}{h^r}\right).$$

Theorem 3 *Let $f_n(\cdot, Y_1, \dots, Y_n)$ be the kernel density estimator in (2) based on noisy observations with noise having polynomially decreasing Fourier transform and a bandwidth $h \to 0$ such that $nh^{2s+1} \to \infty$, when $n \to \infty$. Then Theorem 1 holds. Moreover, I_n is asymptotically normally distributed with mean and variance given by* (5) *in Corollary 2; if f belongs to the class $S(\alpha, r, L)$, the integrated square error $ISE(f_n, f)$ is asymptotically normally distributed with*

$$MISE(f_n, f) \le L\exp\left(-\frac{2\alpha}{h^r}\right) + \frac{1 + o(1)}{\pi(2s+1)nh^{2s+1}} \text{ and } V_f[ISE(f_n, f)] = \frac{2\|g\|_2^2(1 + o(1))}{\pi(4s+1)n^2h^{4s+1}}$$

and the $MISE$ becomes minimal for

$$h_* = \left(\frac{\log n}{2\alpha} - \frac{2s - r + 1}{2\alpha r}\log\log n\right)^{-1/r}$$

giving

$$\inf_{h>0} \sup_{f \in S(\alpha,r,L)} MISE(f_n, f) = \frac{1}{\pi(2s+1)n}\left(\frac{\log n}{2\alpha}\right)^{(2s+1)/r}.$$

The main density being here much smoother than the variance, we can at the same time choose a bandwidth h that minimizes the $MISE(f_n, f)$ and makes the bias term $\exp(-2\alpha/h^r)$ negligible. Indeed, consider,

$$h = \left(\frac{\log n}{2\alpha} - \sqrt{\log n}\right)^{-1/r}. \tag{13}$$

Then $h/h_* \to 1$, when $n \to \infty$,

$$\exp\left(-\frac{2\alpha}{h^r}\right) = \exp\left(-\frac{2\alpha}{h_*^r}\right)\exp\left(-2\alpha\sqrt{\log n} + \frac{2s+1}{r}\log\log n\right) = o\left(\exp\left(-\frac{2\alpha}{h_*^r}\right)\right)$$

and thus

$$MISE(f_n, f) = \frac{1 + o(1)}{\pi(2s+1)nh^{2s+1}} = \frac{1 + o(1)}{\pi(2s+1)nh_*^{2s+1}}$$

and the confidence interval can be written as in (6) for the bandwidth h in (13).

3.2 Sobolev densities and exponential noise

The situation changes completely if the noise is exponentially smooth. From Butucea and Tsybakov (2003) we know

$$E_f[I_n] = \frac{h^{s-1}(1+o(1))}{2\pi\gamma sn} \exp\left(\frac{2\gamma}{h^s}\right)$$

and this has to be $o(1)$ as a necessary condition for the $MISE$ to be consistent.

Theorem 4 *Let $f_n(\cdot, Y_1, \dots, Y_n)$ be the kernel density estimator defined in (2) based on noisy observations with noise having exponentially decreasing Fourier transform in our convolution model and a bandwidth $h \to 0$ such that $h^{s-1}\exp(2\gamma/h^s)/n \to 0$, when $n \to \infty$. Then*

$$\sqrt{\frac{2\pi\gamma sn^2}{h^{s-1}\exp(4\gamma/h^s)\|g\|_2^2}}\left(ISE(f_n, f) - E_f[ISE(f_n, f)]\right) \to N(0, 1)$$

where the convergence is in law when $n \to \infty$. Moreover, I_n is asymptotically normally distributed with

$$E_f[I_n] = \frac{h^{s-1}}{2\pi\gamma sn} e^{2\gamma/h^s}(1+o(1)) \text{ and } V_f[I_n] = \frac{h^{s-1}\|g\|_2^2}{2\pi\gamma sn^2} e^{4\gamma/h^s}(1+o(1));$$

if f belongs to the class $W(r, L)$, the integrated square error $ISE(f_n, f)$ is asymptotically normally distributed with

$$MISE(f_n, f) \le Lh^{2r} + \frac{h^{s-1}}{2\pi\gamma sn} e^{2\gamma/h^s}(1+o(1)) \text{ and } V_f[ISE(f_n, f)] = \frac{h^{s-1}\|g\|_2^2}{2\pi\gamma sn^2} e^{4\gamma/h^s}(1+o(1))$$

and the $MISE$ becomes minimal (and of the order of the minimax L_2 risk, see Efromovich (1997)) for h_* of order $(\log n/(2\gamma))^{-1/s}$

$$\inf_{h>0} \sup_{f \in W(r,L)} MISE(f_n, f) = L\left(\frac{\log n}{2\gamma}\right)^{-2r/s}.$$

In this case the bias term, the bias term Lh_*^{2r} is dominating in the expression of $MISE(f_n, f)$.

Proof. Indeed, we can see that

$$V_f[S_1] \le C\frac{\|K^n\|_2^4}{h^2 n^3} \le C\frac{h^{2s-2}}{(2\pi\gamma s)^2 n^3} \exp\left(\frac{4\gamma}{h^s}\right),$$

for some constant $C > 0$, and

$$V_f[S_2] = \frac{h^{s-1}\|g\|_2^2(1+o(1))}{2\pi\gamma sn^2} \exp\left(\frac{4\gamma}{h^s}\right).$$

We can see that $V_f[S_1]/V_f[S_2] \leq h^{s-1}/n = o(1)$ and thus S_2 is still the dominating term in the weak convergence to the normal law.

Moreover, $\int (M^n)^2 = (1 + o(1))h^s \exp(4\gamma/h^s)/(4\pi\gamma s)$ and finally

$$\left(E[G_n^2(X_1, X_2)] + n^{-1}E[H_n^4(X_1, X_2)]\right) / \left(E[H_n^2(X_1, X_2)]\right)^2 \leq O(h^3) + \frac{O(1)}{nh^3} = o(1).$$

By Proposition 1 we deduce the asymptotic normality. □

4 Auxiliary results

Lemma 2 *Let f_n be the kernel estimator defined in (2) with the particular choice of the kernel and for arbitrary $h > 0$ small. Then*

$$E_f[f_n(x)] = K \star f(x).$$

Moreover, due to the choice of the kernel the cross term in $ISE(f_n, f)$ is null

$$\int \left(f_n(x) - E_f[f_n(x)]\right)\left(E_f[f_n(x)] - f(x)\right) dx = 0.$$

Proof. For the first part, we use the Fourier inversion formula, the expression of the Fourier transform of the kernel and the fact that $\Phi^g = \Phi \cdot \Phi^\eta$:

$$\begin{aligned} E_f[f_n(x)] &= \int \frac{1}{h} K^n\left(\frac{x-y}{h}\right) g(y)dy = \frac{1}{2\pi} \int e^{-ixu} \Phi^{K^n}(hu)\Phi^g(u)du \\ &= \frac{1}{2\pi} \int e^{-ixu}\Phi^K(hu)\Phi(u)du = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy = K_h \star f(x). \end{aligned}$$

Next,

$$\begin{aligned} &\int \left(f_n(x) - E_f[f_n(x)]\right)\left(E_f[f_n(x)] - f(x)\right) dx \\ &= \int \left(f_n(x) - E_f[f_n(x)]\right) E_f[f_n(x)]dx - \int \left(f_n(x) - E_f[f_n(x)]\right) f(x)dx. \qquad (14) \end{aligned}$$

Now, the first term of the difference, we use again Plancherel formula (saying that $\int p \cdot q = \int \Phi^p \cdot \overline{\Phi^q}/2\pi$ for any functions p and q in L_1 and L_2):

$$\begin{aligned} &\int \left(f_n(x) - E_f[f_n(x)]\right) E_f[f_n(x)]dx \\ &= \frac{1}{n}\sum_{i=1}^n \int \left(K_h^n(x - Y_i) - K_h \star f\right) K_h \star f(x)dx \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2\pi n}\sum_{i=1}^{n}\int e^{-ixu}\left(\frac{\Phi^K(hu)e^{iuY_i}}{\Phi^\eta(u)} - \Phi^K(hu)\Phi(u)\right)\overline{\Phi^K(hu)\Phi(u)}du \\
&= \frac{1}{\pi n}\sum_{i=1}^{n}\int e^{-ixu}\left(\frac{\Phi^K(hu)e^{iuY_i}}{\Phi^\eta(u)} - \Phi^K(hu)\Phi(u)\right)\overline{\Phi(u)}du \\
&= \int \left(f_n(x) - E_f[f_n(x)]\right) f(x)dx,
\end{aligned}
$$

where $\overline{\Phi^K(u)}$ is the complex conjugate of $\Phi^K(u) = I[|u| \le 1]$ and we use the fact that $(\Phi^K)^2 = \Phi^K$. Then the difference in (14) is null. □

Lemma 3 *1) If f belongs to a Sobolev class $W(r, L)$ with $r > 1/2$, then $g = f \star \eta$, with η the density of a polynomial noise, is $(r + s - 1/2)$- Lipschitz continuous function. If f is a supersmooth density in $S(\alpha, r, L)$, then g is at least Lipschitz continuous.*
2) If f is either Sobolev or supersmooth density then f and $g = f \star \eta$ are uniformly bounded densities, whether the noise is polynomial or exponential. That means, there exists a constant $C > 0$, depending only on r, s, L, such that

$$
\sup_f \|f\|_\infty \le C \text{ and } \sup_f \|g\|_\infty \le C.
$$

3) If the noise is polynomial then the deconvolution kernel defined in (3) *has*

$$
\|K^n\|_2^2 = \frac{1 + o(1)}{\pi(2s + 1)h^{2s}},
$$

if the noise is exponential, then it has

$$
\|K^n\|_2^2 = \frac{h^s(1 + o(1))}{2\pi\gamma s}\exp\left(\frac{2\gamma}{h^s}\right).
$$

Proof. 1) If f is in the Sobolev class $W(r, L)$ and η is the density of a polynomial noise, we have:

$$
\begin{aligned}
|g(x + y) - g(x)| &= \frac{1}{2\pi}\left|\int (e^{-iu(x+y)} - e^{-iux})\Phi^g(u)du\right| \\
&\le \frac{1}{2\pi}\int \frac{|e^{-iuy} - 1|}{|u|^{r+s}}|\Phi(u)||u|^r|\Phi^\eta(u)||u|^s du \\
&\le \frac{1}{2\pi}\left(\int \frac{|e^{-iuy} - 1|^2}{|u|^{2(r+s)}}du \int |\Phi(u)|^2|u|^{2r}|\Phi^\eta(u)||u|^{2s}du\right)^{1/2} \\
&\le \frac{|y|^{r+s-1/2}}{2\pi}\left(\int \frac{|e^{-iv} - 1|^2}{|v|^{2(r+s)}}dv\right)^{1/2}\left(\int_{|u|\le M} |\Phi(u)|^2|\Phi^\eta(u)|^2|u|^{2(r+s)}du\right. \\
&\quad \left. + \int_{|u|>M} |\Phi(u)|^2|u|^{2r}du\right)^{1/2}
\end{aligned}
$$

and all the integrals are finite, for any $M > 0$ large enough but fixed. Then there exists a finite constant $C > 0$ that does not depend on x or y, such that

$$|g(x+y) - g(x)| \leq C|y|^{r+s-1/2}.$$

We omit the similar proofs in the cases where either the noise is exponential or the density f is supersmooth.

2) Probability density functions f in the Sobolev class are such that:

$$\begin{aligned} |f(x)| &= \frac{1}{2\pi}\left|\int e^{-ixu}\Phi(u)du\right| \\ &\leq \frac{1}{2\pi}\left(\int |\Phi(u)|^2(1+|u|^{2r})du \int (1+|u|^{2r})^{-1}du\right)^{1/2}, \end{aligned}$$

which is less than some constant C depending only on r and L. Similarly for g.

3) For this we refer to Butucea (2004) and Butucea and Tsybakov (2003). □

References

Butucea, C. (2004). Deconvolution of supersmooth densities with smooth noise, *Canadian J. Statist.*, 32, to appear.

Butucea, C. and Tsybakov, A.B. (2003). Fast asymptotics in density deconvolution, Manuscript.

Carroll, R.J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density, *J. Amer. Statist. Assoc.*, 83, 1184-1186.

Fan, J. (1991*a*). Global behavior of deconvolution kernel estimates, *Statist. Sinica*, 1, 541-551.

Fan, J. (1991*b*). Asymptotic normality for deconvolution kernel density estimators, *Sankhyā Ser. A*, 53, 97-110.

Fan, J. and Koo, J.-Y. (2002). Wavelet deconvolution, *IEEE Trans. Inform. Theory*, 48, 734-747.

Fan, Y. and Liu, Y. (1997). A note on asymtptotic normality for deconvolution kernel estimators, *Sankhyā Ser. A*, 59, 138-141.

Goldenshluger, A. (1999). On pointwise adaptive nonparametric deconvolution, *Bernoulli*, 5, 907-926.

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators, *Journal of Multivariate Analysis*, 14, 1-16.

Laurent, B. (1996). Efficient estimation of integral functionals of a density, *Ann. Statist.*, 24, 659-681.

Resum

En aquest article considerem un estimador nucli de la densitat en un model de convolució i donem un teorema central del límit pel seu error quadràtic integrat. L'estimador nucli és força usual en teoria mínimax quan la densitat subjacent es recupera a partir d'observacions amb soroll. El nucli està fixat i depèn fortament de la distribució de l'error, la qual se suposa totalment coneguda. L'amplada de banda no està fixada, els resultats es verifiquen per qualsevol seqüència d'amplades decreixents cap a 0. En particular, es pot aplicar el teorema central del límit per l'amplada de banda que minimitza l'error quadràtic integrat mitjà. Les velocitats de convergència són força diferents en el cas de sorolls regulars i de sorolls super-regulars. La suavitat de la densitat subjacent és rellevant en l'avaluació del l'error quadràtic integrat mitjà.

MSC: 62G05, 62G20

Paraules clau: error quadràtic integrat, estimació no paramètrica de la densitat, model de convolució, observacions amb soroll, teorema central del límit

SORT 28 (1) January-June 2004, 27-36

On best affine unbiased covariance-preserving prediction of factor scores

Heinz Neudecker*

University of Amsterdam, The Netherlands

Abstract

This paper gives a generalization of results presented by ten Berge, Krijnen, Wansbeek & Shapiro. They examined procedures and results as proposed by Anderson & Rubin, McDonald, Green and Krijnen, Wansbeek & ten Berge. We shall consider the same matter, under weaker rank assumptions. We allow some moments, namely the variance Ω of the observable scores vector and that of the unique factors, Ψ, to be singular. We require $T'\Psi T > 0$, where $T\Lambda T'$ is a Schur decomposition of Ω. As usual the variance of the common factors, Φ, and the loadings matrix A will have full column rank.

MSC: 62H25, 15A24

Keywords: Factor analysis, factor scores, covariance-preserving, Kristof-type theorem

1 Introduction

We consider the factor model $y = \mu_y + Af + \varepsilon$, where y is a $p \times 1$ vector of observable random variables called «scores», f is an $m \times 1$ vector of non-observable random variables called «common factors», A is a $p \times m$ matrix of full column rank whose elements are called «factor loadings» and ε is a $p \times 1$ vector of non-observable random variables called «unique factors». The usual moment definitions and assumptions are

$$E(\varepsilon) = 0, \quad E(f) = 0, \quad E(y) = \mu_y, \quad D(\varepsilon) = \Psi, \quad D(f) = \Phi, \quad C(f, \varepsilon) = 0.$$

* *Address for correspondence:* Oosterstraat, 13. 1741 GH Schagen. The Netherlands. Postal address: NL 1741 GH 13. E-mail: H.Neudecker@uva.nl
Received: December 2001
Accepted: September 2003

This yields the moment structure

$$\Omega = A\Phi A' + \Psi,$$

where $\Omega = D(y)$ and Ψ can be singular, Φ and A have full column rank. Notice that

$$\mathcal{M}(A) \subset \mathcal{M}(\Omega). \tag{1.1}$$

The following *additional* assumption is made:

$$T'\Psi T > 0.$$

It is inspired by the Schur decomposition $\Omega = T\Lambda T'$, with $T'T = I_r$ and diagonal $\Lambda > 0$. Obviously $p \geqslant r > m$.

In two recent publications Krijnen, Wansbeek & ten Berge (1996) and ten Berge, Krijnen, Wansbeek & Shapiro (1999) studied the problem of best linear prediction of f given y, subject to the constraint $E\hat{f}\hat{f}' = Eff'$, where $\hat{f} = B'y$ is their predictor function. Vectors f and y have a simultaneous distribution. The two expectations are taken with respect to this distribution.

The constraint $E\hat{f}\hat{f}' = Eff'$ is mistakenly referred to as «correlation-preserving». We shall call it «covariance-preserving», although at face value only the RHS expression is a variance matrix. We shall use an affine predictor function $\hat{f} = a + B'y$. It will be shown that $a + B'\mu_y = 0$. Hence the predictor function will become $\hat{f} = B'(y - \mu_y)$ which is linear and unbiased. Consequently the LHS expression will become a variance matrix.

In their article ten Berge *et al.* (1999) examine three prediction procedures, due to McDonald (1981) —who generalized a procedure proposed by Anderson & Rubin (1956)—, Green (1969) and Krijnen *et al.* (1996), respectively.

We shall consider the same three procedures. The second and third are based on the mean-squared-error matrix $M = E\left(\hat{f} - f\right)\left(\hat{f} - f\right)'$. Where Green minimizes its trace, $\operatorname{tr} M$, Krijnen *et al.* minimize its determinant, $|M|$. McDonald uses a different though related criterion $\operatorname{tr}\Psi^{-1}E\left(y - \mu_y - A\hat{f}\right)\left(y - \mu_y - A\hat{f}\right)'$ which he minimizes. Note that these authors assume $\Psi > 0$, hence $\Omega > 0$. ten Berge *et al.* conclude that McDonald's and Krijnen *et al.*'s solutions for B coincide.

In the present paper we shall again consider the above-mentioned procedures, under weaker rank assumptions. We shall show that the MSE matrix M is positive definite. Minimization of the trace and the determinant of M yields immediately $a + B'\mu_y = 0$. Minimization of McDonald's criterion function yields the same result. As mathematical methods we use a Kristof-type theorem and a matrix inequality developed by Zhang (1999). Finally we show that 1) $\hat{f}_G$, the Green predictor and $\hat{f}_K$, the Krijnen *et al.* predictor coincide when Φ and $A'\Omega^{+}A$ commute, 2) $\hat{f}_M$, the McDonald predictor and $\hat{f}_K$ coincide when Ψ and $A\Phi A'$ commute.

2 A Kristof-type theorem

Two of the three criterion functions can be seen to belong to the class $\operatorname{tr} P'X$, where P and X have dimension $p \times m$. The constant matrix P has rank q. The variable matrix X satisfies the constraint $X'X = I_m$. The aim is to maximize $\operatorname{tr} P'X$ subject to $X'X = I_m$. Define then the Lagrangean function

$$\varphi(X) = \operatorname{tr} P'X - \tfrac{1}{2} \operatorname{tr} L (X'X - I_m),$$

where L is a *symmetric* matrix of multipliers. Symmetry of L is vital. It is justified, of course, by the symmetry of the constraint.

The differential of the function, namely

$$d\varphi = \operatorname{tr} P' dX - \operatorname{tr} LX'dX = \operatorname{tr} (P - XL)' \, dX$$

has to be zero. This yields the equations

$$P = XL \tag{2.1}$$

$$X'X = I_m \tag{2.2}$$

From these we obtain

$$P'P = L^2 \tag{2.3}$$

$$P = X (P'P)^{\frac{1}{2}} \tag{2.4}$$

Which square root will be selected is still undecided. Consider equation (2.4). As

$$P (P'P)^{+\frac{1}{2}} (P'P)^{\frac{1}{2}} = P$$

it is consistent. The symbol «+» denotes the Moore-Penrose inverse. The symbols «+» and «$\frac{1}{2}$» are interchangeable in $(P'P)^{+\frac{1}{2}}$. The general solution of (2.4) is

$$X_\circ = P (P'P)^{+\frac{1}{2}} + Q - Q (P'P)^{\frac{1}{2}} (P'P)^{+\frac{1}{2}}, \quad Q \text{ arbitrary} \tag{2.5}$$

When we use the singular-value decomposition $P = F_1 \Gamma_1^{\frac{1}{2}} G_1'$, with $F_1'F_1 = G_1'G_1 = I_q$ and (diagonal) $\Gamma_1^{\frac{1}{2}} > 0$, we can write the solution as

$$X_\circ = F_1 G_1' + Q (I_m - G_1 G_1') \tag{2.6}$$

It follows from (2.5) that

$$\operatorname{tr} P'X_\circ = \operatorname{tr} (P'P)^{\frac{1}{2}}. \tag{2.7}$$

As we look for a *maximum*, we have to take the *positive* definite square root $(P'P)^{\frac{1}{2}}$. The solution $X_\circ$ is *not* unique, unless $q = m$. In that case it can be written as

$$X_\circ = P (P'P)^{-\frac{1}{2}} = F_1 G_1' \tag{2.8}$$

For the connaisseurs we shall examine the second differential

$$d^2\varphi = -\mathrm{tr}(dX)L(dX)' \tag{2.9}$$

When this expression is *negative* for all $dX \neq 0$ satisfying $(dX)'X_\circ + X_\circ' dX = 0$, a maximum has been found. The choice $L = (P'P)^{\frac{1}{2}} > 0$ guarantees this.

3 The Green procedure

As stated we use the MSE matrix $M = E\left(\hat{f} - f\right)\left(\hat{f} - f\right)' = \left(a + B'\mu_y\right)\left(a + B'\mu_y\right)' + B'\Omega B + \Phi - B'A\Phi - \Phi A'B$. Obviously $a + B'\mu_y = 0$, as we have to minimize tr$M$. As a consequence $E\hat{f}\hat{f}' = B'\Omega B$. Imposition of the constraint $E\hat{f}\hat{f}' = Eff'$ yields then $M = 2\Phi - B'A\Phi - \Phi A'B$. Green (1969) defines the problem:

$$\min_B \mathrm{tr}\,(2\Phi - B'A\Phi - \Phi A'B) \quad \text{subject to } B'\Omega B = \Phi.$$

We introduce $C' = \Phi^{-\frac{1}{2}}B'\Omega^{\frac{1}{2}}$. Clearly $C'C = I_m$. This yields the equivalent problem

$$\max_C \mathrm{tr}\,\Phi^{\frac{3}{2}}A'\Omega^{+\frac{1}{2}}C \quad \text{subject to } C'C = I_m$$

We used: $A'\Omega^{+\frac{1}{2}}C\,\Phi^{\frac{1}{2}} = R'\Omega^{\frac{1}{2}}\Omega^{+\frac{1}{2}}\Omega^{\frac{1}{2}}B = R'\Omega^{\frac{1}{2}}B = A'B$, with $A = \Omega^{\frac{1}{2}}R$ due to (1.1).

Application of the Kristof-type theorem gives the solution

$$C_G = \Omega^{+\frac{1}{2}}A\Phi^{\frac{3}{2}}\left(\Phi^{\frac{3}{2}}A'\Omega^{+}A\Phi^{\frac{3}{2}}\right)^{-\frac{1}{2}},$$

from which follows the solution

$$B_G = \Omega^{+}A\Phi^{\frac{3}{2}}\left(\Phi^{\frac{3}{2}}A'\Omega^{+}A\Phi^{\frac{3}{2}}\right)^{-\frac{1}{2}}\Phi^{\frac{1}{2}} + \left(I_p - \Omega^{+\frac{1}{2}}\Omega^{\frac{1}{2}}\right)Q, \quad Q \text{ arbitrary.}$$

The arbitrary component disappears in the predictor expression $B_G'(y - \mu_y)$, because $\left(I_p - \Omega^{\frac{1}{2}}\Omega^{+\frac{1}{2}}\right)\left(y - \mu_y\right) = 0$ with probability one (w. p. 1).

Hence we get as predictor

$$\hat{f}_G = \Phi^{\frac{1}{2}}\left(\Phi^{\frac{3}{2}}A'\Omega^{+}A\Phi^{\frac{3}{2}}\right)^{-\frac{1}{2}}\Phi^{\frac{3}{2}}A'\Omega^{+}\left(y - \mu_y\right).$$

The reader can verify that $A'\Omega^{+}A > 0$.

An alternative expression is

$$C_G = F_2G_2',$$

where we have used the singular-value decomposition

$$\Omega^{+\frac{1}{2}}A\Phi^{\frac{3}{2}} = F_2\Gamma_2^{\frac{1}{2}}G_2',$$

with $F_2'F_2 = G_2'G_2 = G_2G_2' = I_m$. Use was made of the fact that $\Omega^{+\frac{1}{2}}A$ has full column rank (m).

For nonsingular Ω the solution becomes that given by ten Berge *et al.* (1999) in their presentation, namely between (6) and (7).

4 The McDonald procedure

This approach is based on the weighted-least-squares function

$$\operatorname{tr}\Psi^{+}E\left(y-\mu_y-A\hat{f}\right)\left(y-\mu_y-A\hat{f}\right)'.$$

Clearly

$$E\left(y-\mu_y-A\hat{f}\right)\left(y-\mu_y-A\hat{f}\right)' = \left(I_p-AB'\right)\Omega\left(I_p-BA'\right)+$$
$$+A\left(a+B'\mu_y\right)\left(a+B'\mu_y\right)'A'.$$

Again we find that $a + B'\mu_y = 0$, now having to minimize

$$\operatorname{tr}\Psi^{+}E\left(y-\mu_y-A\hat{f}\right)\left(y-\mu_y-A\hat{f}\right)'.$$

Notice that $A'\Psi^{+}A > 0$.

Imposition of the constraint $E\hat{f}\hat{f}' = Eff'$ leads to the problem of minimizing

$$\operatorname{tr}\Psi^{+}\left(I_p-AB'\right)\Omega\left(I_p-BA'\right) \quad \text{subject to } B'\Omega B=\Phi.$$

Using $C' = \Phi^{-\frac{1}{2}}B'\Omega^{\frac{1}{2}}$ we define the problem:

$$\max_C \operatorname{tr}\Phi^{\frac{1}{2}}A'\Psi^{+}\Omega^{\frac{1}{2}}C \quad \text{subject to } C'C=I_m.$$

Application of the Kristof-type theorem yields the solution

$$C_M=\Omega^{\frac{1}{2}}\Psi^{+}A\Phi^{\frac{1}{2}}\left(\Phi^{\frac{1}{2}}A'\Psi^{+}\Omega\Psi^{+}A\Phi^{\frac{1}{2}}\right)^{-\frac{1}{2}},$$

from which follows the solution

$$B_M=\Omega^{+\frac{1}{2}}\Omega^{\frac{1}{2}}\Psi^{+}A\Phi^{\frac{1}{2}}\left(\Phi^{\frac{1}{2}}A'\Psi^{+}\Omega\Psi^{+}A\Phi^{\frac{1}{2}}\right)^{-\frac{1}{2}}\Phi^{\frac{1}{2}}+\left(I_p-\Omega^{+\frac{1}{2}}\Omega^{\frac{1}{2}}\right)Q, \quad Q \text{ arbitrary.}$$

Finally the predictor turns out to be

$$\hat{f}_M=\Phi^{\frac{1}{2}}\left(\Phi^{\frac{1}{2}}A'\Psi^{+}\Omega\Psi^{+}A\Phi^{\frac{1}{2}}\right)^{-\frac{1}{2}}\Phi^{\frac{1}{2}}A'\Psi^{+}\left(y-\mu_y\right).$$

Again we used

$$\left(I_p-\Omega^{\frac{1}{2}}\Omega^{+\frac{1}{2}}\right)\left(y-\mu_y\right)=0 \quad \text{w.p.1.}$$

The reader can verify that $A'\Psi^{+}\Omega\Psi^{+}A > 0$, using

$$A'\Psi^{+}\Omega\Psi^{+}A=A'\Psi^{+}\left(A\Phi A'+\Psi\right)\Psi^{+}A=A'\Psi^{+}A\Phi A'\Psi^{+}A+A'\Psi^{+}A.$$

An alternative expression is

$$C_M = F_3 G_3',$$

where

$$\Omega^{\frac{1}{2}}\Psi^{+}A\Phi^{\frac{1}{2}} = F_3\Gamma_3^{\frac{1}{2}}G_3',$$

with $F_3'F_3 = G_3'G_3 = G_3G_3' = I_m$.

For nonsingular Ω the solution becomes that given by ten Berge *et al.* (1999) in their presentation, namely between (4) and (5).

5 The Krijnen *et al.* procedure

Like Green's this approach uses the MSE matrix M of $\hat{f}$. Instead of tr $(2\Phi - B'A\Phi - \Phi A'B)$, Krijnen *et al.* use $|2\Phi - B'A\Phi - \Phi A'B|$ which has to be minimized. The first thing to do is to prove that $2\Phi - B'A\Phi - \Phi A'B > 0$.

We have

$$\begin{aligned} 2\Phi - B'A\Phi - \Phi A'B &= \Phi^{\frac{1}{2}}\left(2I_m - \Phi^{-\frac{1}{2}}B'A\Phi^{\frac{1}{2}} - \Phi^{\frac{1}{2}}A'B\Phi^{-\frac{1}{2}}\right)\Phi \\ &= \Phi^{\frac{1}{2}}\left(2I_m - \Phi^{-\frac{1}{2}}B'\Omega^{\frac{1}{2}}\Omega^{+\frac{1}{2}}A\Phi^{\frac{1}{2}} - \Phi^{\frac{1}{2}}A'\Omega^{+\frac{1}{2}}\Omega^{\frac{1}{2}}B\Phi^{-\frac{1}{2}}\right)\Phi^{\frac{1}{2}} \\ &= \Phi^{\frac{1}{2}}\left(2I_m - C'V - V'C\right)\Phi^{\frac{1}{2}} \\ &= \Phi^{\frac{1}{2}}\left[(C-V)'(C-V) + (I_m - V'V)\right]\Phi^{\frac{1}{2}} \end{aligned}$$

where

$$V = \Omega^{+\frac{1}{2}}A\Phi^{\frac{1}{2}} \quad \text{and hence} \quad V'V = \Phi^{\frac{1}{2}}A'\Omega^{+}A\Phi^{\frac{1}{2}}.$$

We shall show that all eigenvalues of $V'V$ are positive and less than unity. Pre-(post-) multiply the moment structure $\Omega = A\Phi A' + \Psi$ by $\Phi^{\frac{1}{2}}A'\Omega^{+}\left(\Omega^{+}A\Phi^{\frac{1}{2}}\right)$. This leads to $\Phi^{\frac{1}{2}}A'\Omega^{+}A\Phi^{\frac{1}{2}} = \left(\Phi^{\frac{1}{2}}A'\Omega^{+}A\Phi^{\frac{1}{2}}\right)^2 + \Phi^{\frac{1}{2}}A'\Omega^{+}\Psi\Omega^{+}A\Phi^{\frac{1}{2}}$, hence

$$\Phi^{\frac{1}{2}}A'\Omega^{+}A\Phi^{\frac{1}{2}} > \left(\Phi^{\frac{1}{2}}A'\Omega^{+}A\Phi^{\frac{1}{2}}\right)^2,$$

as $\Phi^{\frac{1}{2}}A'\Omega^{+}\Psi\Omega^{+}A\Phi^{\frac{1}{2}} > 0$, and $\lambda_i > \lambda_i^2$ where λ_i is any eigenvalue of $\Phi^{\frac{1}{2}}A'\Omega^{+}A\Phi^{\frac{1}{2}}$. This proves the property. Hence $I_m - V'V > 0$. As $(C-V)'(C-V) \geqslant 0$ we have shown that $2\Phi - B'A\Phi - \Phi A'B > 0$.

Hence $|2\Phi - B'A\Phi - \Phi A'B| > 0$. Consider then the positive definite matrix $2I_m - C'V - V'C$. We use (7.18) in Zhang (1999) which yields

$$C'V + V'C \leqslant 2U'\left(V'CC'V\right)^{\frac{1}{2}}U$$

where U is an orthogonal matrix.

As $C'C = I_m$ we have $CC' \leqslant I_p$. This in its turn leads to $V'CC'V \leqslant V'V$. The latter inequality gives $(V'CC'V)^{\frac{1}{2}} \leqslant (V'V)^{\frac{1}{2}}$. See Theorem 2.5.5 in Wang & Chow (1994).

Finally, we have

$$C'V + V'C \leqslant 2U'(V'V)^{\frac{1}{2}}U$$

or equivalently

$$2I_m - C'V - V'C \geqslant 2\left[I_m - U'(V'V)^{\frac{1}{2}}U\right].$$

From this we derive

$$\left|2I_m - C'V - V'C\right| \geqslant \left|2\left[I_m - U'(V'V)^{\frac{1}{2}}U\right]\right| = \left|2\left[I_m - (V'V)^{\frac{1}{2}}\right]\right|.$$

It is easy to see that $C_K = V(V'V)^{-\frac{1}{2}}$ leads to the equality

$$\left|2I_m - C_K'V - V'C_K\right| = \left|2\left[I_m - (V'V)^{\frac{1}{2}}\right]\right|.$$

Hence C_K solves the problem. It is not clear whether the solution is unique.

In fact, C_K also solves the related problem

$$\max_C \ \mathrm{tr}\, V'C \quad \text{subject to } C'C = I_m.$$

The (unique) solution is C_K by the Kristof-type theorem.

Application of Zhang's (7.18) yields

$$2\,\mathrm{tr}\, V'C = \mathrm{tr}\,(C'V + V'C) \leqslant 2\,\mathrm{tr}\, U'(V'V)^{\frac{1}{2}}U = 2\,\mathrm{tr}\,(V'V)^{\frac{1}{2}},$$

which again has solution C_K. We then get the solution

$$B_K = \Omega^+ A\Phi^{\frac{1}{2}}\left(\Phi^{\frac{1}{2}}A'\Omega^+A\Phi^{\frac{1}{2}}\right)^{-\frac{1}{2}}\Phi^{\frac{1}{2}} + \left(I_p - \Omega^{+\frac{1}{2}}\Omega^{\frac{1}{2}}\right)Q, \quad Q \text{ arbitrary.}$$

From this follows the *unique* predictor

$$\hat{f}_K = \Phi^{\frac{1}{2}}\left(\Phi^{\frac{1}{2}}A'\Omega^+A\Phi^{\frac{1}{2}}\right)^{-\frac{1}{2}}\Phi^{\frac{1}{2}}A'\Omega^+\left(y - \mu_y\right).$$

For nonsingular Ω the solution C_K coincides with that given by ten Berge *et al.* (1999), namely in (9).

6 Equality of $\hat{f}_G$ and $\hat{f}_K$ when Φ and $A'\Omega^+A$ commute

ten Berge *et al.* (1999) showed that $C_G = C_K$ under their assumptions when Φ and $A'\Omega^{-1}A$ commute. We shall prove that $\hat{f}_G = \hat{f}_K$ under our milder conditions.

When Φ and $A'\Omega^+A$ commute we have $\Phi = SMS'$ and $A'\Omega^+A = SNS'$, where M and N are positive definite diagonal matrices and S is orthogonal. Hence

$$\begin{aligned}\Phi^{\frac{3}{2}}\left(\Phi^{\frac{3}{2}}A'\Omega^+A\Phi^{\frac{3}{2}}\right)^{-\frac{1}{2}} &= SM^{\frac{3}{2}}S'\left(SM^{\frac{3}{2}}S'SNS'SM^{\frac{3}{2}}S'\right)^{-\frac{1}{2}}\\ &= SM^{\frac{3}{2}}S'\left(SM^3NS'\right)^{-\frac{1}{2}} = SM^{\frac{3}{2}}S'S\left(M^3N\right)^{-\frac{1}{2}}S'\\ &= SN^{-\frac{1}{2}}S' = (A'\Omega^+A)^{-\frac{1}{2}}.\end{aligned}$$

Further $$\Phi^{\frac{1}{2}}\left(\Phi^{\frac{1}{2}}A'\Omega^{+}A\Phi^{\frac{1}{2}}\right)^{-\frac{1}{2}} = (A'\Omega^{+}A)^{-\frac{1}{2}}.$$

This yields $$\hat{f}_G = \hat{f}_K = \Phi^{\frac{1}{2}}(A'\Omega^{+}A)^{-\frac{1}{2}}A'\Omega^{+}\left(y-\mu_y\right).$$

7 Equality of $\hat{f}_M$ and $\hat{f}_K$ when Ψ and $A\Phi A'$ commute

ten Berge *et al.* (1999) showed that $C_M = C_K$ under their assumptions when Ψ is nonsingular. Essential is the expression

$$\Omega^{-1} = \Psi^{-1} - \Psi^{-1}A\Phi^{\frac{1}{2}}\left(I_m + \Phi^{\frac{1}{2}}A'\Psi^{-1}A\Phi^{\frac{1}{2}}\right)^{-1}\Phi^{\frac{1}{2}}A'\Psi^{-1}.$$

Under our assumptions $I_m + \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}}$ is nonsingular because $A'\Psi^{+}A > 0$ which follows from $T'\Psi T > 0$ and (1.1). When we additionally assume that Ψ and $A\Phi A'$ commute we can establish the equality

$$\Omega^{+} = \Psi^{+} - \Psi^{+}A\Phi^{\frac{1}{2}}\left(I_m + \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}}\right)^{-1}\Phi^{\frac{1}{2}}A'\Psi^{+}.$$

Proof. When Ψ and $A\Phi A'$ commute we have $\Psi = SMS'$ and $A\Phi A' = SNS'$ where M and N are positive definite diagonal matrices and $S'S = I_m$. Further $A\Phi^{\frac{1}{2}} = SN^{\frac{1}{2}}T'$, with orthogonal T, a singular-value decomposition. Hence

$$\begin{aligned}
&\Psi^{+} - \Psi^{+}A\Phi^{\frac{1}{2}}\left(I_m + \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}}\right)^{-1}\Phi^{\frac{1}{2}}A'\Psi^{+}\\
&= SM^{-1}S' - SM^{-1}S'SN^{\frac{1}{2}}T'\left(I_m + TN^{\frac{1}{2}}S'SM^{-1}S'SN^{\frac{1}{2}}T'\right)^{-1}TN^{\frac{1}{2}}S'SM^{-1}S'\\
&= SM^{-1}S' - SM^{-1}N^{\frac{1}{2}}T'\left(I_m + TM^{-1}NT'\right)^{-1}TM^{-1}N^{\frac{1}{2}}S'\\
&= SM^{-1}S' - SM^{-1}N^{\frac{1}{2}}T'T\left(I_m + M^{-1}N\right)^{-1}T'TM^{-1}N^{\frac{1}{2}}S'\\
&= SM^{-1}S' - SM^{-1}N^{\frac{1}{2}}\left(I_m + M^{-1}N\right)^{-1}M^{-1}N^{\frac{1}{2}}S'.
\end{aligned}$$

Further $\Omega = A\Phi A' + \Psi = S(M+N)S'$, and $\Omega^{+} = S(M+N)^{-1}S'$. It is easy to see that

$$(M+N)^{-1} = M^{-1} - M^{-1}N^{\frac{1}{2}}\left(I_m + M^{-1}N\right)^{-1}M^{-1}N^{\frac{1}{2}}.$$

This yields the result. □

Recall that

$$\hat{f}_M = \Phi^{\frac{1}{2}}\left(\Phi^{\frac{1}{2}}A'\Psi^{+}\Omega\Psi^{+}A\Phi^{\frac{1}{2}}\right)^{-\frac{1}{2}}\Phi^{\frac{1}{2}}A'\Psi^{+}(y-\mu_y)$$

and

$$\hat{f}_K = \Phi^{\frac{1}{2}}\left(\Phi^{\frac{1}{2}}A'\Omega^{+}A\Phi^{\frac{1}{2}}\right)^{-\frac{1}{2}}\Phi^{\frac{1}{2}}A'\Omega^{+}(y-\mu_y).$$

Consider

$$\begin{aligned}
\Phi^{\frac{1}{2}}A'\Omega^{+}A\Phi^{\frac{1}{2}} &= \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}} - \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}}\left(I_m + \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}}\right)^{-1} \\
&\quad \times \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}} \\
&= \left(I_m + \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}}\right)^{-1}\Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}} \\
&= (I_m + E)^{-1}E, \\
\Phi^{\frac{1}{2}}A'\Psi^{+}\Omega\Psi^{+}A\Phi^{\frac{1}{2}} &= \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi A'\Psi^{+}A\Phi^{\frac{1}{2}} + \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}} \\
&= \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}} + \left(\Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}}\right)^2 \\
&= E + E^2, \\
\Phi^{\frac{1}{2}}A'\Omega^{+} &= \Phi^{\frac{1}{2}}A'\Psi^{+} - \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}}\left(I_m + \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}}\right)^{-1}\Phi^{\frac{1}{2}}A'\Psi^{+} \\
&= \left(I_m + \Phi^{\frac{1}{2}}A'\Psi^{+}A\Phi^{\frac{1}{2}}\right)^{-1}\Phi^{\frac{1}{2}}A'\Psi^{+} \\
&= (I_m + E)^{-1}\Phi^{\frac{1}{2}}A'\Psi^{+}, \\
\hat{f}_K &= \Phi^{\frac{1}{2}}\left[(I_m + E)^{-1}E\right]^{-\frac{1}{2}}(I_m + E)^{-1}\Phi^{\frac{1}{2}}A'\Psi^{+}\left(y - \mu_y\right), \\
\hat{f}_M &= \Phi^{\frac{1}{2}}\left(E + E^2\right)^{-\frac{1}{2}}\Phi^{\frac{1}{2}}A'\Psi^{+}\left(y - \mu_y\right).
\end{aligned}$$

Clearly

$$\left[(I_m + E)^{-1}E\right]^{-\frac{1}{2}}(I_m + E)^{-1} = \left(E + E^2\right)^{-\frac{1}{2}} \quad \text{as } E > 0.$$

This establishes the equality of $\hat{f}_K$ and $\hat{f}_M$.

8 Comments

1. ten Berge *et al.* (1999) claim that the McDonald method is undefined when Ψ is singular. This is unjustified. What matters is the nonsingularity of $T'\Psi T$. We make that assumption. It implies that $A'\Psi^{+}A > 0$ which we use several times.
2. Application of Zhang's result shows immediately that C_G and C_M yield the maximum. The Kristof-type theorem shows the *unicity* of the solutions.

Acknowledgement

The author is grateful to Götz Trenkler for drawing his attention to Zhang's result (7.18) which was fruitfully applied in Section 5. The reasoning why $B'\mu_y + a$ should be equal

to zero in all three procedures is due to Albert Satorra. This yields $\hat{f} = B'\left(y - \mu_y\right)$, an unbiased predictor.

9 References

Anderson, T. W. & Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.* Vol. V, 111-50. University of California Press, Berkeley and Los Angeles.

ten Berge, J. M. F. (1993). *Least Squares Optimization in Multivariate Analysis. DSWO Press*, Leiden, Netherlands.

ten Berge, J. M. F., Krijnen, W. P., Wansbeek, T. & Shapiro, A. (1999). Some new results on correlation-preserving factor scores prediction methods. *Linear Algebra Appl.*, 289, 311-8.

Green, B. F. (1969). Best linear composites with a specified structure. *Psychometrika*, 34, 301-18.

Krijnen, W. P., Wansbeek, T. J. & ten Berge, J. M. F. (1996). Best linear predictors for factor scores. *Communications in Statistics: Theory and Methods*, 25, 3013-25.

Kristof, W. (1970). A theorem on the trace of certain matrix products and some applications. *Journal of Mathematical Psychology*, 7, 515-30.

Löwner, K. (1934). Über monotone Matrixfunktionen. *Mathematische Zeitschrift*, 38, 177-216.

Magnus, J. R. & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics. Revised Edition.* Wiley, Chichester, England.

McDonald, R. P. (1981). Constrained least squares estimators of oblique common factors. *Psychometrika*, 46, 337-41.

Wang, S. G. & Chow, S. C. (1994). *Advanced Linear Models: Theory and Applications.* Marcel Dekker, New York.

Zhang, F. (1999). *Matrix Theory. Basic Results and Techniques.* Springer, New York.

Resum

Es dóna una generalització dels resultats presentats per ten Berge, Krijnen, Wansbeek and Shapiro . Aquests autors examinen mètodes i resultats basats en Anderson i Rubin. Mc Donald, Green i Krijnen, Wansbeek i ten Berge. Considerarem el mateix plantejament però sota condicions de rang més dèbils. Així suposarem que alguns moments, com les matrius de covariàncies Ω del vector de mesures observades dels factors comuns i ψ dels factors únics, siguin singulars. Imposem la condició $T'\psi T > 0$, essent $T\Lambda T'$ la descomposició de Schur de Ω. Com és usual, suposem que tenen rang màxim per columnes les matrius de covariàncies Φ dels factors comuns i la matriu A del model factorial.

MSC: 62H25, 15A24

Paraules clau: anàlisi factorial, mesures de factors, preservació de la covariància, teorema tipus Kristof

SORT 28 (1) January-June 2004, 37-54

Local superefficiency of data-driven projection density estimators in continuous time

Denis Bosq and Delphine Blanke*

Université Pierre et Marie Curie, Paris

Abstract

We construct a data-driven projection density estimator for continuous time processes. This estimator reaches superoptimal rates over a class $\mathcal{F}_0$ of densities that is dense in the family of all possible densities, and a «reasonable» rate elsewhere. The class $\mathcal{F}_0$ may be chosen previously by the analyst. Results apply to $\mathbb{R}^d$-valued processes and to $\mathbb{N}$-valued processes. In the particular case where square-integrable local time does exist, it is shown that our estimator is strictly better than the local time estimator over $\mathcal{F}_0$.

MSC: 62G07, 62M

Keywords: Density estimation, data-driven, superefficiency, continuous time processes

1 Introduction

We study a data-driven projection density estimator $\hat{f}_T$ in a general framework where data are in continuous time. The purpose is to reach a superoptimal rate on a class $\mathcal{F}_0$ of densities that is dense in $\mathcal{F}$, the family of all possible densities, and a «reasonable» rate elsewhere. The class $\mathcal{F}_0$ can be previously chosen by the analyst.

The results are, in some sense, extensions of those which where obtained in the i.i.d. case (cf Bosq 2002*a*, 2002*b*), but in this new context the methods are often different.

Section 2 contains notation and assumptions. In Section 3 we study the estimator over $\mathcal{F}_0$. We obtain a $\frac{1}{T}$-rate with respect to the mean integrated square error, a $(\frac{\ln\ln T}{T})^{1/2}$-

* *Address for correspondence:* (Université Pierre et Marie Curie, Paris-France)
Received: October 2003
Accepted: January 2004

rate with respect to uniform error, and a Gaussian limit in distribution with coefficient of normalization $\sqrt{T}$. Results concerning the asymptotic behaviour of $\hat{f}_T$ over $\mathcal{F} - \mathcal{F}_0$ appear in Section 4. Finally, Section 5 is devoted to comparison of $\hat{f}_T$ with the local time estimator $f_{T,0}$ when this estimator exists. It is shown that, in a special case, $\hat{f}_T$ is strictly better than $f_{T,0}$. The proofs are postponed until Section 6.

2 Notation and assumptions

Let $(E, \mathcal{B}, \mu)$ be a measure space, with μ σ-finite, and such that $L^2(\mu)$ is infinite dimensional. The norm of $L^2(\mu)$ will be denoted $\|.\|$. Let $(e_j,\ j \geq 0)$ be an orthonormal system in $L^2(\mu)$.

We consider a stochastic process $X = (X_t,\ t \in \mathbb{R})$ defined on a probability space $(\Omega, \mathcal{A}, P)$ and with values in $(E, \mathcal{B})$. X is supposed to be measurable and such that the X_t's are identically distributed with density f with respect to μ.

Denote $\mathcal{F}$ the family of densities f such that

$$f = \sum_{j=0}^{\infty} a_j e_j, \quad \sum_{j=0}^{\infty} a_j^2 < \infty. \tag{2.1}$$

The class of the observable processes will be denoted $\mathcal{X}$. Note that two different processes may have the same f. In order to estimate f from the data $(X_t,\ 0 \leq t \leq T)$ $(T > 0)$ we use a data-driven projection estimator :

$$\hat{f}_T = \sum_{j=0}^{\hat{k}_T} \widehat{a}_{j_T} e_j \text{ with } \widehat{a}_{j_T} = \frac{1}{T} \int_0^T e_j(X_t)\, dt, \quad j \geq 0$$

and

$$\hat{k}_T = \max\left\{ j\ :\ 0 \leq j \leq k_T,\ \left|\widehat{a}_{j_T}\right| \geq \gamma_T \right\}$$

where γ_T and the integer k_T are chosen by the analyst. If $\{\ldots\} = \emptyset$ one sets $\hat{k}_T = k_T$.

We always suppose that (unless otherwise stated)

$$k_T \to \infty, \quad \frac{k_T}{T} \to 0, \quad \gamma_T \to 0, \quad \text{as } T \to \infty.$$

If $\gamma_T = 0$ one obtains the projection density estimator

$$f_T = \sum_{j=0}^{k_T} \widehat{a}_{j_T} e_j \tag{2.2}$$

Now $\mathcal{F}_0(K)$ will denote the class of $f \in \mathcal{F}$ such that

$$f = \sum_{j=0}^{K} a_j e_j, \quad a_K \neq 0, \text{ and } \mathcal{F}_0 = \bigcup_{K=0}^{\infty} \mathcal{F}_0(K),$$

and finally we put

$$\mathcal{F}_1 = \mathcal{F} - \mathcal{F}_0.$$

In order to study the rates of convergence of $\hat{f}_T$ over $\mathcal{F}_0$ and $\mathcal{F}_1$ we shall use strong mixing coefficients of the form

$$\alpha(C,\mathcal{D}) = \sup_{C\in C, D\in\mathcal{D}} |P(C\cap D) - P(C)P(D)| \tag{2.3}$$

where C and $\mathcal{D}$ are sub-σ-algebra of $\mathcal{A}$.

For a given process $Y = (Y_t,\ t \in I)$, where $I \subseteq \mathbb{R}$, one defines its strong mixing functions as

$$\alpha_Y^{(2)}(u) = \sup_{h\in I, h+u\in I} \alpha(\sigma(Y_h), \sigma(Y_{h+u})),\ \ u \geq 0 \ \text{ and}$$

$$\alpha_Y(u) = \sup_{h\in\mathbb{R}} \alpha(\sigma(Y_t, t\leq h, t\in I),\ \ \sigma(Y_t, t \geq h+u, t\in I)),\ \ u\geq 0$$

with the convention $\alpha(.,.) = 0$ if one of the two sub-σ-algebras is not defined. These two classical coefficients will be used in the sequel.

Now the main assumptions and conditions are H_1 and H_2 :

$$H_1 \begin{cases} A_1 & : \ P_{(X_{s+h},X_{t+h})} = P_{(X_s,X_t)};\ \ s,t,h\in\mathbb{R}\ \textit{(2-stationarity)},\\ B_1(r) & : \ M_r = \sup_{j\geq 0}\left\|e_j(X_0)\right\|_r < \infty,\ \textit{where } 2<r\leq\infty,\\ C_1(r) & : \ \displaystyle\int_0^\infty \left[\alpha_X^{(2)}(u)\right]^{(r-2)/r} du < \infty,\\ c_1 & : \ \gamma_T \simeq T^{-\gamma}\ (\gamma>0)\ \textit{and}\ k_T \simeq T^\beta\ \ (0<\beta<1).\end{cases}$$

$$H_2 \begin{cases} A_2 & : \ X \textit{ is strictly stationary},\\ B_2 = B_1(\infty) & : \ M = \sup_{j\geq 0}\left\|e_j(X_0)\right\|_\infty < \infty,\\ C_2 & : \ \alpha_X(u) \leq a\,e^{-bu}\ \ (a>0, b>0)\\ & \ \ \ (X \textit{ is geometrically strongly mixing, (GSM)}).\\ c_2 & : \ \gamma_T = \left(\dfrac{\ln T \ln\ln T}{T}\right)^{1/2}.\end{cases}$$

Note that A_2 and C_2 are satisfied as soon as X is an enough regular stationary diffusion process (cf Doukhan, 1994). Note also in some situations, one may choose $\gamma_T = c(\frac{\ln T}{T})^{1/2}$ with constant c large enough.

Concerning B_2, it is satisfied in many classical cases, for example if (e_j) is a trigonometric system on a compact interval or the Hermite functions over $\mathbb{R}$. In the particular case where $E = \mathbb{N}$ and μ is the counting measure, the natural system $(1_{\{j\}},\ j\geq 0)$ is, of course, uniformly bounded.

Finally some special assumptions concerning local time will appear in Section 5.

3 Rates of $\hat{f}_T$ over $\mathcal{F}_0$

If $f \in \mathcal{F}_0$ we shall denote $K(f)$ the only integer K such that $f \in \mathcal{F}_0(K)$. The following proposition shows that $\hat{k}_T$ is actually a consistent estimator of $K(f)$.

Proposition 3.1 *If $f \in \mathcal{F}_0$, then*

1) if H_1 holds,

$$P(\hat{k}_T \neq K(f)) = O(T^{\beta+2\gamma-1}) \tag{3.1}$$

thus, if $\beta + 2\gamma < 1$, $\hat{k}_T \to K(f)$ in probability.

2) If H_2 holds,

$$P(\hat{k}_T \neq K(f)) = o(T^{-\delta}), \tag{3.2}$$

for each $\delta > 0$, in particular, if $T = T_n \uparrow \infty$ with $\sum_n T_n^{-\delta} < \infty$, for some $\delta > 0$, then

$$\hat{k}_{T_n} = K(f) \text{ almost surely for } n \text{ large enough.} \tag{3.3}$$

These results show that the adaptive estimator $\hat{f}_T$ has asymptotically the same behaviour as the pseudo-estimator

$$g_T = \sum_{j=0}^{K(f)} \widehat{a}_{j_T} e_j. \tag{3.4}$$

The following lemma makes this fact explicit :

Lemma 3.1 *If $M = \sup_{j\geq 0} \|e_j(X_0)\|_\infty < \infty$, one has*

$$\mathrm{E}\, \|\hat{f}_T - g_T\|^2 \leq M^2 k_T P(\hat{k}_T \neq K(f)). \tag{3.5}$$

We now indicate the rates of $\hat{f}_T$ on $\mathcal{F}_0$, we begin with the mean integrated square error (MISE).

Proposition 3.2 *If $f \in \mathcal{F}_0$, then*
1) If H_1 holds, we have

$$\mathrm{E}\, \|\hat{f}_T - f\|^2 = O(\frac{1}{T^{1-\beta}}) \tag{3.6}$$

2) If H_2 holds,

$$T.\mathrm{E}\, \|\hat{f}_T - f\|^2 \xrightarrow[T\to\infty]{} 2 \sum_{j=0}^{K(f)} \int_0^\infty \mathrm{Cov}\,(e_j(X_0), e_j(X_u))\, \mathrm{d}u. \tag{3.7}$$

The next statement gives a uniform result.

Corollary 3.1

$$\limsup_{T\to\infty} \sup_{X\in\mathcal{X}_0(a_0,b_0,K_0)} T.\mathrm{E}\,\|\hat{f}_T - f\|^2 \le \frac{8a_0M^2K_0}{b_0}. \tag{3.8}$$

Here $\mathcal{X}_0(a_0, b_0, K_0)$ denotes the family of processes that satisfy H_2 with $f \in \mathcal{F}_0(K)$, $K \le K_0$ and $\alpha_X(u) \le ae^{-bu}$ where $a \le a_0$ and $b \ge b_0$.

We now turn to the $\|.\|_\infty$-error :

Proposition 3.3 *If $f \in \mathcal{F}_0$ and H_2 holds, then*

$$(\forall\, \varepsilon > 0), (\forall\, \delta > 0), \quad P(\|\hat{f}_T - f\|_\infty \ge \varepsilon) = O(T^{-\delta}), \tag{3.9}$$

and if $T = T_n = nh$ $(h > 0)$, $n \to \infty$,

$$\|\hat{f}_T - f\|_\infty = O\Big(\big(\frac{\ln\ln T}{T}\big)^{1/2}\Big), \text{ almost surely.} \tag{3.10}$$

Finally the limit in distribution appears in the following statement:

Proposition 3.4 *If $f \in \mathcal{F}_0$, H_2 holds and $T = nh$ $(h > 0)$ then*

$$\sqrt{T}(\hat{f}_T - f) \Rightarrow N \tag{3.11}$$

where «$\Rightarrow$» means weak convergence in $L^2(\mu)$ and N is a zero-mean Gaussian $L^2(\mu)$-valued random variable with $K(f) + 1$-dimensional support.

Proposition 3.2(2), 3.3 and 3.4 exhibit superoptimal rates if $f \in \mathcal{F}_0$. In general these rates appear if the Castellana-Leadbetter condition holds (see Castellana and Leadbetter (1986), Bosq (1998)). Here this condition is *not* needed; this means that local irregularity of the sample paths is not necessary for obtaining these parametric rates over $\mathcal{F}_0$.

4 Asymptotic behaviour of $\hat{f}_T$ over $\mathcal{F}_1$

In order to study consistency of $\hat{f}_T$ when $f \in \mathcal{F}_1$ we need results concerning the behaviour of the truncation index $\hat{k}_T$ as T tends to infinity.

Below the first statement expresses the fact that $\hat{k}_T \to \infty$ in some sense when the second one shows that $\hat{k}_T$ is not far from an «optimal k_T».

Proposition 4.1 *If $f \in \mathcal{F}_1$ then*

1) If H_1 holds

$$P(\hat{k}_T < A) = O(T^{-1}), \quad A > 0, \tag{4.1}$$

2) If H_2 holds

$$P(\hat{k}_T < A) = O(\exp(-c_A\sqrt{T})), \quad (c_A > 0), \ A > 0. \tag{4.2}$$

Now we specify the asymptotic behaviour of $\hat{k}_T$. For this purpose we set

$$q(\eta) = \min\left\{q \in \mathbb{N},\ |a_j| \leq \eta \text{ for all } j > q\right\},\ \eta > 0. \tag{4.3}$$

Note that $q(\eta)$ does exist since $a_j \to 0$, and that, if $q(\eta) > 0$, then $|a_{q(\eta)}| > \eta$. On the other hand $\eta < \eta'$ implies $q(\eta') \leq q(\eta)$.

We put $q_T(\varepsilon) = q((1+\varepsilon)\gamma_T)$, $\varepsilon > 0$; $q'_T(\varepsilon') = q((1-\varepsilon')\gamma_T)$, $0 < \varepsilon' < 1$ and we consider the event

$$E_T := \left\{q_T(\varepsilon) \leq \hat{k}_T \leq q'_T(\varepsilon')\right\}.$$

Then:

Proposition 4.2 *If $f \in \mathcal{F}_1$ and $q_T(\varepsilon) \leq k_T$, we have*

1) Under H_1,

$$P(E_T^c) = O(T^{\beta+2\gamma-1}), \tag{4.4}$$

2) Under H_2,

$$P(E_T^c) = o(T^{-\delta}) \textit{ for all } \delta > 0. \tag{4.5}$$

We indicate two applications of these results:

Example 4.1 Under H_1, if $|a_j| \simeq j^{-\eta}$ $(\eta > \frac{1}{2})$ one has $q_T(\varepsilon) \simeq T^{\gamma/\eta}$, then $2\gamma \leq \beta$ ensures $q_T(\varepsilon) \leq k_T$ for T large enough and $\beta < \frac{1}{2}$ yields $P(E_T^c) \to 0$.

Example 4.2 Under H_2, if $|a_j| \simeq \alpha\rho^j$ $(\alpha > 0,\ 0 < \rho < 1)$ and $k_T > [1+(2\ln 1/\rho)^{-1}]\ln T$, one has $q_T(\varepsilon) \simeq \frac{\ln T}{2\ln(1/\rho)}$,

$$P\left(\left|\frac{\hat{k}_T}{\ln T} - (2\ln 1/\rho)^{-1}\right| > \xi\right) = o(T^{-\delta}),\ \xi > 0,\ \delta > 0. \tag{4.6}$$

In particular, if $T = T_n$ with $\sum_n T_n^{-\delta} < \infty$ for some $\delta > 0$, then

$$\frac{\hat{k}_{T_n}}{\ln T_n} \to (2\ln 1/\rho)^{-1} \text{ almost surely.} \tag{4.7}$$

Note that, from (4.7), one may deduce an estimator of ρ, namely $\widehat{\rho}_T = T^{-\frac{1}{2\hat{k}_T+1}}$ which converges almost surely.

We now may state results concerning the MISE of $\hat{f}_T$.

Proposition 4.3 *If $f \in \mathcal{F}_1$ and $q_T(\varepsilon) \leq k_T$ then*

1) Under H_1,

$$\mathrm{E}\,\|\hat{f}_T - f\|^2 = O(T^{-(1-\beta-2\gamma)}) + \sum_{j>q_T(\varepsilon)} a_j^2. \tag{4.8}$$

2) Under H_2,

$$\mathrm{E}\,\|\hat{f}_T - f\|^2 = O\left(\frac{q'_T(\varepsilon')}{T}\right) + \sum_{j>q_T(\varepsilon)} a_j^2. \tag{4.9}$$

Thus if H_1 and conditions in Example 4.1 hold then, taking $\beta = \frac{1}{2\eta}$, yields

$$\mathrm{E}\,\|f_T - f\|^2 = O(T^{-\frac{2\eta-1}{2\eta}}), \tag{4.10}$$

when $\mathrm{E}\,\left\|\hat{f}_T - f\right\|^2 = O(T^{-\frac{2\eta-1}{2\eta}+2\gamma})$.

Suppose now that conditions in Example 4.2 and H_2 hold. Then, if $\ln T = O(k_T)$, we have

$$\mathrm{E}\,\left\|\hat{f}_T - f\right\|^2 = O\Big(\frac{\ln T \ln\ln T}{T}\Big), \tag{4.11}$$

when, if $k_T \simeq a \ln T$ with $a \geq (2\ln 1/\rho)^{-1}$,

$$\mathrm{E}\,\|f_T - f\|^2 = O\Big(\frac{\ln T}{T}\Big). \tag{4.12}$$

In some special cases one may construct a process for which the rates (4.10) and (4.12) are the true rates for f_T. For example, if (e_j) is the trigonometric basis over $L^2[0,1]$, one may consider the process

$$X_t = Y_{[t]}, \quad t \in \mathbb{R}$$

where $(Y_n,\ n \in \mathbb{Z})$ is a sequence of independent $[0,1]$-valued random variables with common density f. For this process the rates are $T^{-(2\eta-1)/2\eta}$ and $\frac{\ln T}{T}$ respectively. This trick has been used previously in Blanke and Bosq (2000) and Bosq (1998) for the kernel density estimator.

Finally, at least in this special case, the loss of rate for $\hat{f}_T$ is a logarithm. Thus $\hat{f}_T$ has a $1/T$-rate on $\mathcal{F}_0$ and a «good» rate on $\mathcal{F}_1$.

We now turn to uniform rate. We have the following proposition:

Proposition 4.4 *Under H_2, if $|a_j| \simeq \alpha\rho^j$ ($\alpha > 0$, $0 < \rho < 1$), $j \geq 0$ and $k_T \gg \ln T$, if $T = T_n$ where $\sum \frac{\ln T_n}{T_n^\delta} < \infty$ for some $\delta > 0$ then for $f \in \mathcal{F}_1$:*

$$\limsup_{T_n\to\infty} \frac{\sqrt{T_n}}{(\ln T_n)^{3/2}} \left\|\hat{f}_{T_n} - f\right\|_\infty \leq 2\sqrt{\frac{2a\delta}{b}\frac{M^2}{\ln(1/\rho)}} \quad \textit{(almost surely)}. \tag{4.13}$$

Note that the rate in (4.13) is almost optimal since the law of the iterated logarithm shows that the rate cannot be better than $(\frac{\ln\ln T}{T})^{1/2}$.

5 Comparison with the local time estimator

We now suppose that X admits an *occupation density* (or *local time*) with respect to μ. More precisely we make the following assumption:

$H_3\ : \forall\, T \geq 0,\ \exists\, \ell_T \in L^2(\mu \otimes P)$:

$$\int_0^T \varphi(X_t)\,\mathrm{d}t = \int_E \varphi(x)\ell_T(x)\,\mathrm{d}\mu(x), \quad \varphi \in \mathcal{M}(E, \mathbb{R}^+), \tag{5.1}$$

where $\mathcal{M}(E, \mathbb{R}^+)$ is the family of $\mathcal{B}$-$\mathcal{B}_{\mathbb{R}}$ measurable positive real functions defined on E ($\mathcal{B}_{\mathbb{R}}$ is the Borel σ-algebra on $\mathbb{R}$).

In such a situation one defines the local time density estimator as

$$f_{T,0} = \frac{\ell_T}{T}, \quad T > 0 \tag{5.2}$$

$f_{T,0}$ is then the density of the empirical measure μ_T defined by

$$\mu_T(B) = \frac{1}{T}\int_0^T \mathbf{I}_B(X_t)\,\mathrm{d}t, \quad B \in \mathcal{B}.$$

Example 5.1 *If $E = \mathbb{N}$ and μ is the counting measure then H_3 is satisfied and*

$$f_{T,0}(x) = \frac{1}{T}\int_0^T \mathbf{I}_{\{x\}}(X_t)\,\mathrm{d}t, \quad x \in \mathbb{N} \tag{5.3}$$

Example 5.2 *If $E = \mathbb{R}$, and μ is Lebesgue measure, H_3 is equivalent to*

$$\liminf_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}\int_{[0,T]^2} P\left(|X_t - X_s| \le \varepsilon\right)\,\mathrm{d}s\,\mathrm{d}t < \infty, \; T > 0 \tag{5.4}$$

(cf Geman and Horowitz, 1980).

Example 5.3 *If $(E, \mathcal{B}, \mu) \subseteq (E_0, B_0, \mu_0)$ with $\mu = g.\mu_0$ and $0 < m \le g \le m' < \infty$ then if H_3 holds for μ_0 with local time $\ell_T^{(0)}$, it holds for μ with local time $\ell_T = \ell_T^{(0)}/g$.*

Note that, if $E = \mathbb{R}$, the Castellana-Leadbetter condition, 1986 (cf also Bosq, 1998) implies H_3 under mild regularity conditions, if X is strictly stationary.

Results and references concerning the local time estimator appear in Bosq and Davydov (1999) and Bosq (1998). Note that, in particular, $f_{T,0}$ is an unbiased estimator of f : $\mathrm{E}\, f_{T,0} = f$ (a.e.).

Now we need a result concerning the MISE of $f_{T,0}$. For this purpose we denote $\ell_{(k)}$ the local time of X on $]k-1, k]$, $k \in \mathbb{Z}$ and make the following assumption :

H_4 : *X is strictly stationary and the series $L = \sum_{k \ge 1} \int_E \mathrm{Cov}\left(\ell_{(1)}(x), \ell_{(k)}(x)\right) d\mu(x)$ converges.*

Note that the Davydov's inequality shows that a sufficient condition for H_4 is

H_4' : *X is strictly stationary and there exists $r > 2$ such that*
$\int_E \left[E\ell_{(1)}^r(x)\right]^{2/r} \mathrm{d}\mu(x) < \infty$ *and* $\sum_{k \ge 1} [\alpha_X(k)]^{(r-2)/r} < \infty.$

Now the following statement exhibits superefficiency of $f_{T,0}$:

Proposition 5.1 *If H_3 and H_4 hold, then*

$$T.\mathrm{E}\left\|f_{T,0} - f\right\|^2 \to L, \quad f \in \mathcal{F}. \tag{5.5}$$

Concerning $\hat{f}_T$ we have

Proposition 5.2 *If H_3 and H_4 hold, then*

$$\mathrm{E}\,\left\|\hat{f}_T - f\right\|^2 = O(\frac{1}{T}) + \mathrm{E}(\sum_{j>\hat{k}_T} a_j^2). \tag{5.6}$$

Note that the key of the proof of Proposition 5.2 is the fact that $\hat{f}_T = \Pi^{\hat{k}_T} f_{T,0}$ where $\Pi^{\hat{k}_T}$ is the orthogonal projector of $\mathrm{sp}(e_j,\ 0 \le j \le \hat{k}_T)$. A similar property for f_T has been noticed in Frenay (2001). Thus

$$\left\|\hat{f}_T - \Pi^{\hat{k}_T} f\right\| \le \left\|f_{T,0} - f\right\| \quad \text{and} \tag{5.7}$$

$$\mathrm{E}\,\left\|\hat{f}_T - \Pi^{\hat{k}_T} f\right\|^2 \le \mathrm{E}\,\left\|f_{T,0} - f\right\|^2 = O(\frac{1}{T}). \tag{5.8}$$

Consequently the efficiency of $\hat{f}_T$ depends on the «pseudo-bias» $\mathrm{E}\,\sum_{j>\hat{k}_T} a_j^2$. Under conditions in Proposition 4.3 this pseudo-bias may be replaced by $\sum_{j>q_T(\varepsilon)} a_j^2$ and the rates (4.11) and (4.12) do not change. However, $\hat{f}_T$ is better than $f_{T,0}$ over $\mathcal{F}_0$ because $f_{T,0} = \sum_{j=0}^{\infty} \widehat{a}_{j,T} e_j$, when $\hat{f}_T$ has the same asymptotic behaviour as $g_T = \sum_{j=0}^{K(f)} \widehat{a}_{j_T} e_j$ and more precisely :

Proposition 5.3 *If $f \in \mathcal{F}_0$ and H_2, H_3, H_4 hold then*

$$\liminf_{T\to\infty} T.\mathrm{E}\,\left\|f_{T,0} - f\right\|^2 \ge 2\sum_{j=0}^{\infty}\int_0^{\infty} \mathrm{Cov}\,(e_j(X_0), e_j(X_u))\,\mathrm{d}u \tag{5.9}$$

when

$$T.\mathrm{E}\,\left\|\hat{f}_T - f\right\|^2 \to 2\sum_{j=0}^{K(f)}\int_0^{\infty} \mathrm{Cov}\,(e_j(X_0), e_j(X_u))\,\mathrm{d}u. \tag{5.10}$$

It is easy to construct examples where $\int_0^{\infty} \mathrm{Cov}\,(e_j(X_0), e_j(X_u))\,\mathrm{d}u > 0$ for some $j > K(f)$; in that case $\hat{f}_T$ is strictly better than $f_{T,0}$ on $\mathcal{F}_0$.

6 Proofs

6.1 Proof of Proposition 3.1

Set $B_T = \left\{\exists\, j \,:\, 0 \le j \le k_T,\ \left|\widehat{a}_{j_T}\right| \ge \gamma_T\right\}$, then, we have for T large enough and $K = K(f)$, $B_T^c \Rightarrow \left|\widehat{a}_{KT}\right| < \gamma_T \le \frac{|a_K|}{2} \Rightarrow \left|a_{KT} - \widehat{a}_{KT}\right| \ge \frac{|a_K|}{2}$ thus $P(B_T^c) \le \dfrac{4\mathrm{Var}\,\widehat{a}_{KT}}{|a_K|^2}$. Now, 2-stationarity yields

$$\operatorname{Var}\widehat{a}_{KT} = \frac{2}{T}\int_0^T (1-\frac{u}{T})\operatorname{Cov}(e_K(X_0), e_K(X_u))\,du, \tag{6.1}$$

using Davydov's inequality, see Bosq (1998, p. 21), one obtains

$$\operatorname{Var}\widehat{a}_{KT} \leq \frac{2}{T}\int_0^T (1-\frac{u}{T})\frac{2r}{r-2}2^{\frac{r-2}{r}}[\alpha_X^{(2)}(u)]^{\frac{r-2}{r}}\,\|e_K(X_0)\|_r^2\,du$$

and H_1 implies

$$\operatorname{Var}\widehat{a}_{KT} \leq \frac{c_r}{T}, \tag{6.2}$$

where $c_r = \frac{4rM_r^2}{r-2}2^{\frac{r-2}{r}}\int_0^\infty [\alpha_X^{(2)}(u)]^{\frac{r-2}{r}}\,du$ thus

$$P(B_T^c) \leq \frac{4c_r}{a_K^2}\frac{1}{T}. \tag{6.3}$$

Now, as soon as $k_T > K$ and $\gamma_T \leq \frac{|a_K|}{2}$,

$$\left\{\hat{k}_T > K, B_T\right\} \Rightarrow \bigcup_{j=K+1}^{k_T}\left\{|\widehat{a}_{jT}| \geq \gamma_T\right\} \tag{6.4}$$

and

$$\left\{\hat{k}_T < K, B_T\right\} \Rightarrow |\widehat{a}_{KT} - a_K| > \frac{|a_K|}{2} \Rightarrow |\widehat{a}_{KT} - a_K| > \gamma_T \tag{6.5}$$

thus

$$P(\hat{k}_T \neq K, B_T) \leq \frac{1}{\gamma_T^2}\sum_{j=K}^{k_T}\operatorname{Var}\widehat{a}_{KT}, \tag{6.6}$$

again using Davydov's inequality one obtains

$$P(\hat{k}_T \neq K, B_T) = O(\frac{k_T+1}{\gamma_T^2 T}) = O(T^{\beta+2\gamma-1}). \tag{6.7}$$

Now, since (6.3) implies

$$P(\hat{k}_T \neq K, B_T^c) = O(\frac{1}{T}), \tag{6.8}$$

(3.1) follows. □

The proof of (3.2) is similar. It uses the following exponential inequality:

Lemma 6.1 *Let $Y = (Y_t,\ 0 \le t \le T)$ be a real measurable stationary strong mixing process such that $\int_0^\infty \alpha_Y(u)\,\mathrm{d}u < \infty$ and $M_Y = \sup_{0\le t\le T} \|Y_t\|_\infty < \infty$. Then for all $r \in [1, \frac{T}{2}]$ and all positive constants η, κ one has*

$$P(|\frac{1}{T}\int_0^T Y_t - \mathrm{E}\,Y_t\,\mathrm{d}t| \ge \eta) \le 4\exp\Big(-\frac{T\eta^2/M_Y^2}{c_1 + c_2\frac{r}{T} + c_3 M_Y^{-1}\eta r}\Big) + \frac{c_4}{\eta}M_Y\alpha_Y(r) \quad (6.9)$$

with $c_1 = 32(1+\kappa)^2\int_0^\infty \alpha_Y(u)\,\mathrm{d}u$, $c_2 = 4c_1$, $c_3 = \frac{16}{3}(1+\kappa)$, $c_4 = 16\frac{(1+\kappa)}{\kappa}$.

Proof of Lemma 6.1. For q, r such that $2qr = T$, we consider blocks of variables $V_T(j)$, $j = 1, \dots, 2[q]-1$, defined by

$$V_T(j) = \int_{(j-1)r}^{jr} (Y_t - \mathrm{E}\,Y_t)\,\mathrm{d}t \text{ and } V_T(2[q]) = \int_{(2[q]-1)r}^{2qr} (Y_t - \mathrm{E}\,Y_t)\,\mathrm{d}t.$$

So, for any $\eta > 0$,

$$P(|\frac{1}{T}\int_0^T Y_t - \mathrm{E}\,Y_t\,\mathrm{d}t| \ge \eta) \le P(|\sum_{j=1}^{[q]} V_T(2j)| > \frac{T\eta}{2}) + P(|\sum_{j=1}^{[q]} V_T(2j-1)| > \frac{T\eta}{2}).$$

The two terms may be handled similarly. Consider the first one, for example: we use Rio's (2000) coupling result recursively to approximate $V_T(2), \dots, V_T(2[q])$ by independent variables. For any $j \ge 1$, there exists a random variable $V_T^*(2j)$, measurable function of $V_T(2), \dots, V_T(2j)$ such that $V_T^*(2j)$ is independent of $V_T(2), \dots, V_T(2j-2)$ and with same law as $V_T(2j)$. Moreover :

$$\mathrm{E}\,\big|V_T^*(2j) - V_T(2j)\big| \le 2\|V_T(2j)\|_\infty\big(\sup|P(AB) - P(A)P(B)|\big)$$

where the supremum is taken over all sets A and B belonging to σ-algebras of events generated by respectively $\{V_T(2), \dots, V_T(2j-2)\}$ and $V_T(2j)$.

For any positive κ, one may write

$$P(|\sum_{j=1}^{[q]} V_T(2j)| > \frac{T\eta}{2}) \le P(|\sum_{j=1}^{[q]} V_T^*(2j)| > \frac{T\eta}{2(1+\kappa)})$$
$$+ P(|\sum_{j=1}^{[q]} V_T(2j) - V_T^*(2j)| > \frac{T\eta\kappa}{2(1+\kappa)})$$

Since the $V_T^*(2j)$ are independent, Bernstein's inequality (written as in Pollard (1984)) implies

$$P(|\sum_{j=1}^{[q]} V_T^*(2j)| > \frac{T\eta}{2(1+\kappa)}) \le 2\exp\Big(-\frac{T\eta^2/M_Y^2}{c_1 + c_2\frac{r}{T} + c_3 M_Y^{-1}\eta r}\Big)$$

with the help of Billingsley's inequality (1979), and constants c_i as stated as in Lemma 6.1. Moreover, Markov's inequality yields

$$P(|\sum_{j=1}^{[q]} V_T(2j) - V_T^*(2j)| > \frac{T\eta\kappa}{2(1+\kappa)}) \leq \frac{2(1+\kappa)}{T\eta\kappa} \sum_{j=1}^{[q]} \mathrm{E}\,|V_T(2j) - V_T^*(2j)|$$

and the result follows from Rio's coupling result. □

Now the proof of (3.2) consists in applying (6.9) to the processes $(e_j(X_t)-a_j,\ 0 \leq t \leq T)$ for $j = K, \ldots, k_T$. This allows to bound the quantities $P(|\widehat{a}_{j_T} - a_j| \geq \eta)$ for suitable η. In particular, one obtains

$$P(B_T^c) = O(\exp(-A\sqrt{T})), \quad (A > 0) \tag{6.10}$$

Technical details are omitted.

Finally (3.3) comes from Borel-Cantelli lemma. □

6.2 Proof of Lemma 3.1

It suffices to write $\|\hat{f}_T - g_T\|^2 = \|\hat{f}_T - g_T\|^2 \mathbf{1}_{\{\hat{k}_T \neq K\}} \leq (\sum_{j=1}^{\hat{k}_T} \widehat{a}_{j_T}^2)\mathbf{1}_{\{\hat{k}_T \neq K\}}$ $\leq M^2 k_T \mathbf{1}_{\{\hat{k}_T \neq K\}}$, hence (3.5) by taking expectations. □

6.3 Proof of Proposition 3.2

First we have,

$$\mathrm{E}\,\|\hat{f}_T - f\|^2 = \mathrm{E}(\sum_{j=0}^{\hat{k}_T} (\widehat{a}_{j_T} - a_j)^2) + \mathrm{E}(\sum_{j>\hat{k}_T} a_j^2) \tag{6.11}$$

then, by Davydov's inequality: $\mathrm{E}(\sum_{j=0}^{\hat{k}_T} (\widehat{a}_{j_T} - a_j)^2) \leq \sum_{j=0}^{k_T} \mathrm{Var}\,\widehat{a}_{j_T} \leq c_r \frac{k_T}{T}$. On the other hand, if $f \in \mathcal{F}_0(K)$, $\sum_{j>\hat{k}_T} a_j^2 = \sum_{j>\hat{k}_T} a_j^2 \mathbf{1}_{\{\hat{k}_T < K\}}$, hence $\mathrm{E}(\sum_{j>\hat{k}_T} a_j^2) \leq \|f\|^2 P(\hat{k}_T < K)$. Now from (6.5) and (6.8) it follows that $P(\hat{k}_T < K) \leq P(|\widehat{a}_{K_T} - a_K| > \frac{|a_K|}{2}) + O(\frac{1}{T})$. Using Davydov's inequality one obtains the bound

$$\mathrm{E}(\sum_{j>\hat{k}_T} a_j^2) = O(\frac{1}{T}), \tag{6.12}$$

and (6.11) gives (3.6). Concerning (3.7) first note that, if $f \in \mathcal{F}_0(K)$, $P(\hat{k}_T \neq K) = o(T^{-\delta})$ for each $\delta > 0$ (cf (3.2)), thus Lemma 3.1 entails $\mathrm{E}\,\|\hat{f}_T - g_T\| = o(T^{-1})$. Thus it is only necessary to study

$$\mathrm{E}\,\|g_T - f\|^2 = \sum_{j=0}^{K} \mathrm{Var}\,\widehat{a}_{j_T}, \tag{6.13}$$

but using Billingsley's inequality one obtains

$$\int_0^\infty \left|\mathrm{Cov}\,(e_j(X_0), e_j(X_u))\right| \mathrm{d}u \le 4M^2 \int_0^\infty ae^{-bu} \le \frac{4aM^2}{b} < \infty. \tag{6.14}$$

Now since

$$T.\mathrm{Var}\,\widehat{a}_{j_T} = 2\int_0^T (1 - \frac{u}{T})\mathrm{Cov}\,(e_j(X_0), e_j(X_u))\,\mathrm{d}u, \tag{6.15}$$

the dominated convergence theorem gives

$$T.\mathrm{Var}\,\widehat{a}_{j_T} \to 2\int_0^\infty \mathrm{Cov}\,(e_j(X_0), e_j(X_u))\,\mathrm{d}u \tag{6.16}$$

and (6.13) yields (3.7). □

6.4 Proof of Corollary 3.1

It suffices to apply Billingsley's inequality in (6.15) and to verify that the other bounds are uniform over $\mathcal{X}_0(a_0, b_0, K_0)$; details are omitted. □

6.5 Proof of Proposition 3.3

First, putting $K(f) = K$ one has

$|\hat{f}_T - g_T| = |(\hat{f}_T - g_T)\mathbf{I}_{\{\hat{k}_T \neq K\}}| \le \sum_{j=1}^{k_T} |\widehat{a}_{j_T}|\,|e_j|\mathbf{I}_{\{\hat{k}_T \neq K\}} \le M^2 k_T \mathbf{I}_{\{\hat{k}_T \neq K\}}$, one obtains, for all $\varepsilon > 0$ and all $\delta > 0$,

$$P(\left\|\hat{f}_T - g_T\right\|_\infty \ge \varepsilon) \le P(\hat{k}_T \neq K) = o(T^{-\delta}). \tag{6.17}$$

Now, $P(\|g_T - f\|_\infty \ge \varepsilon) \le \sum_{j=0}^{K} P(\left|\widehat{a}_{j_T} - a_j\right| \ge \frac{\varepsilon}{KM})$, then, using (6.9) for $Y_t = e_j(X_t)$, $0 \le t \le T$; $0 \le j \le K$, with $r = B\ln T$ one arrives at the bound

$$P(\left|\widehat{a}_{j_T} - a_j\right| \ge \frac{\varepsilon}{KM}) \le 4\exp(-\frac{T}{\ln T}\frac{3\varepsilon/KM^2B}{16(1+\kappa)(1+o(1))})$$
$$+ 64\frac{1+\kappa}{\kappa}\frac{KM^2}{\varepsilon}a\exp(-bB\ln T)$$

For a given $\delta > 0$ and choosing $B = \delta b^{-1}$ one obtains (3.9).

Concerning (3.10), note that $(e_j(X_t),\ t \in \mathbb{R})$ satisfies the law of the iterated logarithm (LIL) : actually using the LIL for strongly mixing discrete time processes (cf Rio, 2000) one obtains the LIL for the processes $(Z_{ij}^{(h)} = \frac{1}{h}\int_{(i-1)h}^{ih}(e_j(X_t) - a_j)\,\mathrm{d}t,\ i \ge 0)$ since these processes are bounded and geometrically strongly mixing. It follows that $\|g_T - f\|_\infty = O\big((\frac{\ln\ln T}{\ln T})^{1/2}\big)$ almost surely, hence (3.10) by using (6.17) for $T = nh$. □

6.6 Proof of Proposition 3.4

Since $\sqrt{T}(\hat{f}_T - f) = \sqrt{T}(\hat{f}_T - g_T) + \sqrt{T}(g_T - f)$ and $\sqrt{T}\left\|g_T - \hat{f}_T\right\|_\infty \to 0$ in probability (see (6.17)), Theorem 4.4 in Billingsley (1979) shows that it suffices to study asymptotic normality of

$$\sqrt{T}(g_T - f) = \sum_{j=0}^{K}(\widehat{a}_{j_T} - a_j)e_j.$$

This is equivalent to asymptotic normality of the finite dimensional random vector $\sqrt{T}(\widehat{a}_{0_T} - a_0, \ldots, \widehat{a}_{K_T} - a_K)$ which in turn is equivalent to this of the real random variables $\sqrt{T}\sum_{j=0}^{K}\lambda_j(\widehat{a}_{j_T} - a_j)\,;\lambda_1, \ldots, \lambda_K \in \mathbb{R}$. Finally using the processes $(Z_{ij}^{(h)},\ i \geq 0), 0 \leq j \leq K$ and Rio (2000), the desired result follows. □

6.7 Proof of Proposition 4.1

1) Let j_0 such that $a_{j_0} \neq 0$, similarly as in the proof of Proposition 3.1 one obtains

$$\left\{\hat{k}_T < j_0\right\} \Rightarrow \left|\widehat{a_{j_0}}_T - a_{j_0}\right| > \frac{|a_{j_0}|}{2} \tag{6.18}$$

as soon as $k_T \geq j_0$, hence $P(\hat{k}_T < j_0) = O(T^{-1})$. Since $f \in \mathcal{F}_1$, j_0 may be taken arbitrarily large, hence (4.1).

2) (6.18) and the exponential inequality (6.9) lead to (4.2). Details are omitted. □

6.8 Proof of Proposition 4.2

For T large enough we have $\left|a_{q_T(\varepsilon)}\right| > (1 + \varepsilon)\gamma_T$.

1) From Davydov's inequality we get $P(\hat{k}_T < q_T(\varepsilon), B_T) \leq P(\left|\widehat{a}_{q_T(\varepsilon),T} - a_{q_T(\varepsilon)}\right| > \varepsilon\gamma_T) \leq \frac{c_T}{\varepsilon^2 T\gamma_T^2}$.
Now, if $q'_T(\varepsilon') \geq k_T$ one has $P(\hat{k}_T > q'_T(\varepsilon')) = 0$, if not, since $\left|a_j\right| \leq (1 - \varepsilon')\gamma_T$ for $j > q'_T(\varepsilon')$, we have $\left\{\hat{k}_T > q'_T(\varepsilon'), B_T\right\} \Rightarrow \bigcup_{k_T \geq j > q'_T(\varepsilon')}\left|\widehat{a}_{j_T} - a_j\right| > \varepsilon'\gamma_T$
thus $P(\hat{k}_T > q'_T(\varepsilon'), B_T) \leq \frac{c_T(k_T+1)}{\varepsilon'^2 T\gamma_T^2}$ and (4.4) follows.

2) For proving (4.5) we may and do suppose that $q'_T(\varepsilon') < k_T$. Then

$$P(E_T^c \cap B_T) \leq P(|\widehat{a}_{q_T(\varepsilon),T} - a_{q_T(\varepsilon)}| \geq \varepsilon\gamma_T) + \sum_{q'_T(\varepsilon') < j \leq k_T} P(|\widehat{a}_{j_T} - a_j| \geq \varepsilon'\gamma_T),$$

Choosing $r = c\ln T$ in (6.9) one arrives at

$$P(E_T^c \cap B_T) = O(k_T T^{-(c'\ln\ln T)}) + O(k_T\gamma_T^{-1}T^{-cb})$$

for some constant c' and the choice $c > (\frac{3}{2} + \delta)b^{-1}$ leads to (4.5) since $P(B_T^c) = o(T^{-\delta})$ for all $\delta > 0$. □

6.9 Proof of Proposition 4.3

We start from (6.11) and write

$\mathrm{E}(\sum_{j>\hat{k}_T} a_j^2 \mathbf{I}_{E_T^c \cap B_T}) \leq \|f\|^2 P(E_T^c)$, $\mathrm{E}(\sum_{j>\hat{k}_T} a_j^2 \mathbf{I}_{E_T \cap B_T}) \leq \sum_{j>q_T(\varepsilon)} a_j^2$, $\mathrm{E}(\sum_{j=0}^{\hat{k}_T} (\widehat{a}_{j_T} - a_j)^2 \mathbf{I}_{E_T \cap B_T}) \leq \sum_{j \leq q'_T(\varepsilon')} \mathrm{Var}\, \widehat{a}_{j_T}$. Finally, under H_1 we write

$\mathrm{E}(\sum_{j=0}^{\hat{k}_T} (\widehat{a}_{j_T} - a_j)^2 \mathbf{I}_{E_T^c \cap B_T}) \leq \sum_{j=0}^{k_T} \mathrm{Var}\, \widehat{a}_{j_T}$, when under H_2,

$\mathrm{E}(\sum_{j=0}^{\hat{k}_T} (\widehat{a}_{j_T} - a_j)^2 \mathbf{I}_{E_T^c \cap B_T}) \leq 4M^2 k_T P(E_T^c)$, using the above bounds, (6.10) and (6.11) one obtains (4.8) and (4.9). □

6.10 Proof of Proposition 4.4

Let ξ be a positive constant, for any positive κ_i, $i = 1, 2$ one obtains

$$P(\|\hat{f}_T - f\|_\infty \geq \xi) \leq P(\|\hat{f}_T - f\|_\infty \mathbf{I}_{E_T} \geq \frac{\xi}{1+\kappa_1}) + P(\|\hat{f}_T - f\|_\infty \mathbf{I}_{E_T^c} \geq \frac{\xi\kappa_1}{1+\kappa_1})$$
$$\leq P_1 + P_2 + P_3$$

with $P_1 = \sum_{j=1}^{q'_T(\varepsilon')} P(Mq'_T(\varepsilon') \left|\widehat{a}_{j_T} - \mathrm{E}\,\widehat{a}_{j_T}\right| \geq \frac{\xi}{(1+\kappa_1)(1+\kappa_2)})$, $P_3 = P(E_T^c)$ and $P_2 = P(M \sum_{j=q_T(\varepsilon)+1}^{\infty} |a_j| \geq \frac{\xi\kappa_2}{(1+\kappa_1)(1+\kappa_2)})$.

Concerning P_1, the assumptions imply in particular that $q'_T(\varepsilon')$ is of the same order as $\ln T/(2\ln(1/\rho))$. Now (6.9) and the choices $Y_t = e_j(X_t)$, $M_Y = M$, $r = R\ln T$, $\eta = \frac{2\ln(1/\rho)\xi}{M(1+\kappa_1)(1+\kappa_2)\ln T}$ with $\xi^2 = c\frac{(\ln T)^3}{T}$ and $T = T_n$ yield $\sum_n P_1 = O(\frac{\ln T_n}{T_n^\delta})$ as soon as $R = (\frac{1}{2} + \delta)b^{-1}$ and $c = \frac{8M^4(1+\kappa_1)^2(1+\kappa_2)^2(1+\kappa)^2 a\delta}{b\ln^2(1/\rho)}$.

Now noting that $\sum_{j=q_T(\varepsilon)+1}^{\infty} |a_j| \leq C(\alpha, \rho)\gamma_T$, it is easy to see that for T_n large enough, $P_2 = 0$ with previous choices of γ_T and ξ. Moreover, Proposition 4.2 implies also $P_3 = o(T_n^{-\delta})$. Finally, collecting these results, one obtains Proposition 4.4 with the help of Borel-Cantelli's lemma since $\sum_n \frac{\ln T_n}{T_n^\delta} < \infty$. □

6.11 Proof of Proposition 5.1

Using additivity of local time one may write $\frac{\ell_T}{T} = \frac{\ell_{\{0\}}}{T} + \frac{1}{T}\sum_{j=1}^{[T]} \ell_{(j)} + \frac{\ell_{][T],T]}}{T}$. Since $\mathrm{E}\left\|\frac{\ell_{\{0\}}}{T}\right\|^2 = o(\frac{1}{T})$ and $\mathrm{E}\left\|\frac{\ell_{][T],T]}}{T}\right\|^2 \le \frac{\mathrm{E}\|\ell_{(1)}\|^2}{T} = o(\frac{1}{T})$ it suffices to study

$$n\mathrm{E}\|\frac{1}{n}\sum_{j=1}^{n} \ell_{(j)} - f\|^2 =$$

$$\mathrm{E}\|\ell_{(1)} - f\|^2 + 2\sum_{k=1}^{n-1}(1-\frac{k}{n})\int_E \mathrm{Cov}\,(\ell_{(1)}(x), \ell_{(k+1)}(x))\,\mathrm{d}\mu(x) \quad (6.19)$$

where $n = [T]$. A classical trick allows to prove that the second member of (6.19) tends to L, hence (5.5). □

6.12 Proof of Proposition 5.2

Let $\Pi^{\hat{k}_T}$ be the orthogonal projector of $\mathrm{sp}(e_j,\ 0 \le j \le \hat{k}_T)$, we have

$\left\|\Pi^{\hat{k}_T}(f_{T,0} - f)\right\| \le \left\|f_{T,0} - f\right\|$ thus $\mathrm{E}\left\|\hat{f}_T - \Pi^{\hat{k}_T} f\right\|^2 \le \mathrm{E}\left\|f_{T,0} - f\right\|^2$ and (5.5) implies $\limsup_{T\to\infty} T.\mathrm{E}\left\|\hat{f}_T - \Pi^{\hat{k}_T} f\right\|^2 \le L$ hence (5.6) from (6.11) and the fact that $P(E_T^c \cup B_T^c) = o(\frac{1}{T})$. □

6.13 Proof of Proposition 5.2

This is clear from (6.11), (5.7) and (5.8). □

6.14 Proof of Proposition 5.3

(5.10) has been proved in Proposition 3.2. Concerning (5.9) first note that (5.1) implies $\frac{1}{T}\int_0^T e_j(X_t)\,\mathrm{d}t = \frac{1}{T}\int_E e_j(x)\ell_T(x)\,\mathrm{d}x$ thus $\widehat{a}_{j_T} = \int_E f_{T,0}(x)e_j(x)\,\mathrm{d}\mu(x)$, $j \ge 0$, hence $f_{T,0} = \sum_{j=0}^{\infty} \widehat{a}_{j_T} e_j$ and $\sum \widehat{a}_{j_T}^2 < \infty$ (almost surely); then we have $T\mathrm{E}\left\|f_{T,0} - f\right\|^2 = \sum_{j=0}^{\infty} T\mathrm{Var}\,\widehat{a}_{j_T}$ but H_2 yields $\int_0^\infty \left|\mathrm{Cov}\,(e_j(X_0), e_j(X_u))\right|\,\mathrm{d}u < \infty$ and (6.16) holds. This implies (5.9) by using Fatou lemma for the counting measure. □

References

Billingsley, P. (1979). *Probability and Measure*, Wiley, New-York.

Blanke, D. and Bosq, D. (2000). A family of minimax rates for density estimators in continuous time, *Stochastic Anal. Appl.*, 18, 871-900.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes. Estimation and Prediction*, Lecture Notes in Statistics, 110, Second edn, Springer-Verlag, New-York.

Bosq, D. (2002*a*). Estimation localement suroptimale et adaptative de la densité, *C.R. Acad. Sci. Paris Ser. I*, 334, 591-595.

Bosq, D. (2002*b*). Estimation suroptimale de la densité par projection. Submitted for publication.

Bosq, D. and Davydov, Yu. (1999). Local time and density estimation in continuous time, *Math. Methods Statist.*, 8, 22-45.

Castellana, J.V. and Leadbetter, M.R. (1986). On smoothed probability density estimation for stationary processes, *Stochastic Process. Appl.*, 21, 179-193.

Doukhan, P. (1994). *Mixing : Properties and Examples*, Lecture Notes in Statistics, 85, Springer-Verlag.

Frenay, A. (2001). Sur l'estimation de la densité marginale d'un processus à temps continu par projection orthogonale, *Ann. I.S.U.P.*, 45, 55-92.

Geman, D. and Horowitz, J. (1980). Occupation densities, *Ann. Probab.*, 8, 1-67.

Pollard, D. (1984). *Convergence of Stochastic Processes*, Springer-Verlag, New-York.

Rio, E. (2000). *Théorie Asymptotique des Processus Aléatoires Faiblements Dépendants*, Mathématiques & Applications, (31), Springer-Verlag.

Resum

Construim un estimador de projecció conduïda per les dades per a processos en temps continu. Aquest estimador assoleix taxes super-òptimes sobre una classe $\mathcal{F}_0$ de densitats que és densa en la família de totes les densitats, i assoleix, a la vegada, taxes "raonables". La classe $\mathcal{F}_0$ pot ésser escollida prèviament per l'Estadístic.
Els resultats s'apliquen a processos a valors $\mathbb{R}^d$ i a valors N. En el cas particular on existeix un temps local de quadrat integrable, es demostra que el nostre estimador és estrictament millor que l'estimador temps local sobre $\mathcal{F}_0$.

MSC: 62G07, 62M

Paraules clau: estimació de densitats, conduït per les dades, processos a temps continu

SORT 28 (1) January-June 2004, 55-68

Modelling Stock Returns with AR-GARCH Processes*

Elżbieta Ferenstein[1,2] and Mirosław Gąsowski[3]

Warsaw, Poland

Abstract

Financial returns are often modelled as autoregressive time series with random disturbances having conditional heteroscedastic variances, especially with GARCH type processes. GARCH processes have been intensely studying in financial and econometric literature as risk models of many financial time series. Analyzing two data sets of stock prices we try to fit AR(1) processes with GARCH or EGARCH errors to the log returns. Moreover, hyperbolic or generalized error distributions occur to be good models of white noise distributions.

MSC: Primary 62M10, 91B84; secondary 62M20

Keywords: autoregressive process, GARCH and EGARCH models, conditional heteroscedastic variance, financial log returns

1 Introduction

Let S_t, $t = 0, 1, \ldots, T$, denote share prices observed at discrete moments. In the considered examples they are daily close prices of Elektrim and Okocim enterprise shares from the Warsaw Stock Exchange over a period 1994–2002. Graphs of the analyzed prices are given in Figures 1 and 3. Let R_t denote the log return at time t, so

* This work was supported by the grant PBZ-KBN-016/P03/99.

[1] *Address for correspondence:* Faculty of Mathematics and Information Science. Warsaw University of Technology. Pl. Politechniki 1, 00-661 Warsaw, Poland

[2] *Address for correspondence:* Polish-Japanese Institute of Information Technologies. Koszykowa 86, 02-008 Warsaw, Poland

[3] Bank Gospodarki Żywnościowej S.A. Kasprzaka 10/16, 01-211 Warsaw, Poland

Received: October 2003
Accepted: January 2004

$$R_t = \ln\left(\frac{S_t}{S_{t-1}}\right), \qquad t = 1,2,\ldots T. \tag{1}$$

Let $X_t = R_t - \overline{R}$ be the mean-centred process, where $\overline{R}$ denotes the sample mean over the observation period. Within a class of autoregressive processes with white noises having conditional heteroscedastic variances we try to find reasonable models of $\{X_t\}$. Some well known processes from a broad class of GARCH processes are listed below. $\{X_t\}$ is called an autoregressive process of order k with an GARCH noise of order p, q, in short AR(k)-GARCH(p,q) process, if for $t = 0, \pm 1, \pm, 2, \ldots$:

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_k X_{t-k} + \varepsilon_t, \tag{2}$$

$$\varepsilon_t = \sigma_t v_t, \tag{3}$$

where $\{v_t\}$ is a strong white noise (iid (0,1)), and $\{\sigma_t\}$ satisfies the recurrence equation

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_q \sigma_{t-q}^2. \tag{4}$$

Thus $\mathrm{Var}(\varepsilon_t|\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) = \sigma_t^2$, $E(\varepsilon_t) = 0$, $\mathrm{Cov}(\varepsilon_t, \varepsilon_s) = 0$, $t \neq s$. The residuals $\{\varepsilon_t\}$ satisfy (3) and (4) is an GARCH(p,q) process (Bolereslev 1986)). If in (4) all $\beta_i = 0$, then $\{\varepsilon_t\}$ is an ARCH(p) process introduced by Engle (1982). Above processes with heteroscedastic variances model such features of financial time series as risk variability, clustering data, leptokurtic property. Popular models of $\{v_t\}$ distributions are: normal, t-Student, GED (generalized error distribution), and hyperbolic. The latter three distributions have heavier tails than normal distributions (leptokurtic property). For instance, Eberlein and Keller (1995) have found that returns of some financial time series from German Stock may be considered as strong white noise from a hyperbolic distribution with $\mu = 0$, where the hyperbolic density is written as

$$f(x; \alpha, \beta, \delta, \mu) = \frac{\sqrt{\alpha^2 - \beta^2}}{2\alpha\delta K(\delta\sqrt{\alpha^2 - \beta^2})} \exp\left(-\alpha\sqrt{\delta^2 + (x-\mu)^2} + \beta(x-\mu)\right),$$

$\alpha > 0$, $|\beta| < \alpha$, K is a type I Bessel function, i.e.

$$K(t) = \frac{1}{2}\int_0^\infty \exp\left(-\frac{1}{2}t\left(x + \frac{1}{2}\right)\right) dx, \qquad t > 0.$$

Some disadvantage of GARCH processes in modelling financial returns is a symmetry of conditional variance of ε_t with respect to positive and negative values of $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$. In practice, one observes an leverage effect, i.e. asymmetric consequences of positive and negative innovations (the conditional variance tends to decrease if noise is positive-implying bigger returns). An EGARCH process (exponential GARCH) introduced by Nelson (1991) does not have this disadvantage. An EGARCH(p,q) process $\{\varepsilon_t\}$ satisfies (3) and below relations:

$$\ln(\sigma_t^2) = \alpha_0 + \alpha_1 g(v_{t-1}) + \cdots + \alpha_p g(v_{t-p}) + \beta_1 \ln(\sigma_{t-1}^2) + \cdots + \beta_q \ln(\sigma_{t-q}^2),$$

where $g(v_t) = \theta v_t + \delta(|v_t| - E|v_t|)$.

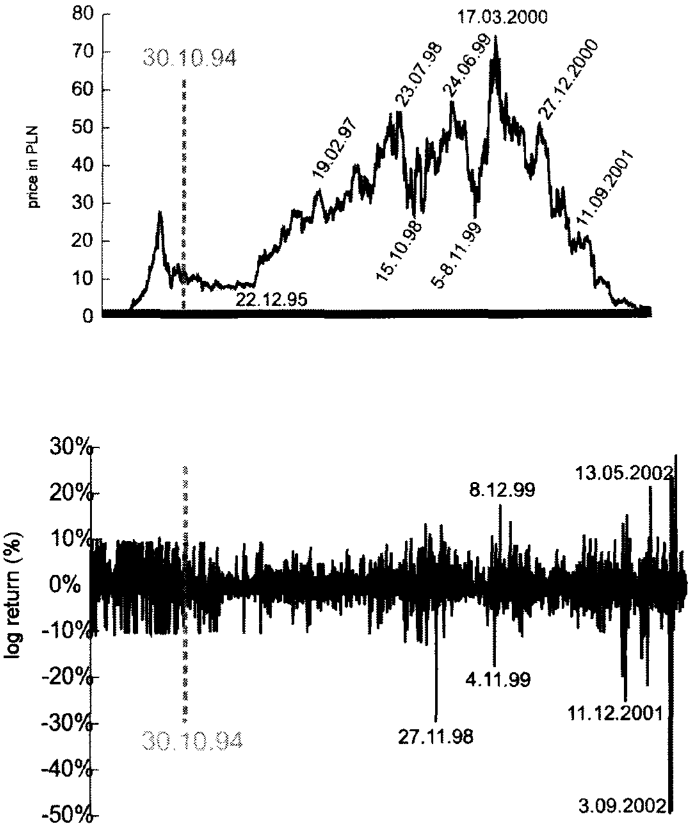

***Figure 1**: Elektrim 26.03.92–09.12.02 (2398 observations).*

Another process modelling the leverage effect was introduced by Glosten *et al.* (1993). Its conditional variance is as follows

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i(|\varepsilon_{t-i}| + \gamma_i \varepsilon_{t-i})^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2.$$

Sometimes, the simplest models of returns may be useful as proposed for instance, in the RiskMetrics process: $X_t = \sigma_t v_t$, $\{v_t\} \sim \text{iid}\, N(0, 1)$, where σ_t^2 is a historical variance estimated as

$$\hat{\sigma}_t^2 = (1 - \lambda) \cdot \sum_{j=0}^{n} \lambda^j X_{t-1-j}^2 = (1 - \lambda)X_{t-1}^2 + \lambda \hat{\sigma}_{t-1}^2,$$

where the smoothing constant $\lambda = 0.94$ (0.96) for daily (monthly) returns, $n = \frac{\ln \gamma}{\ln \lambda}$, with γ such that $1 - \lambda^n = 1 - \gamma$.

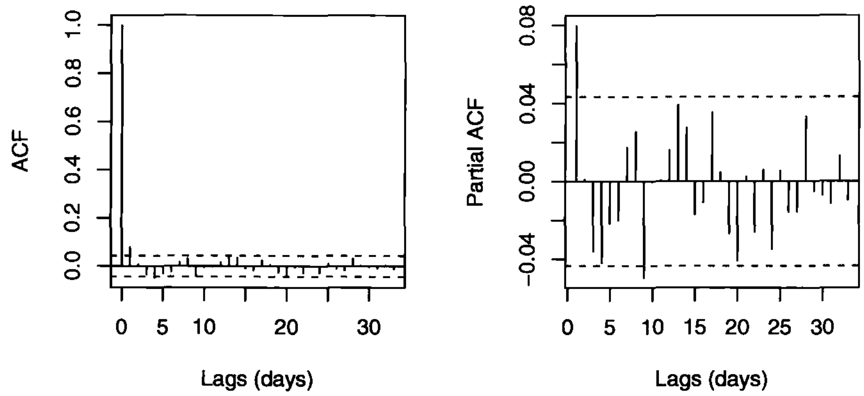

Figure 2: *Elektrim – sample ACF and PACF of log returns.*

2 Modelling empirical returns

Figures 1 and 3 present log returns of daily close prices of two enterprise shares–Elektrim and Okocim, from Warsaw Stock. If they were realizations of strong white noises their empirical autocorrelations

$$\widehat{\rho}(j) = \widehat{\gamma}(j)/\widehat{\gamma}(0), \quad j = 1, 2, \dots, h, \quad h < T,$$

$\widehat{\gamma}(j) = T^{-1}\sum_{t=1}^{T-j}(R_{t+j} - \overline{R})(R_t - \overline{R})$, would have behaved for large T approximately as random samples from normal distribution with the mean 0 and variance T^{-1}. Hence, the null hypothesis that log returns are realizations of random samples should be rejected if more than 5% of sample autocorrelations fall out of $[-1.96/\sqrt{T}, 1.96/\sqrt{T}]$ or at least one is significantly far from this interval. Analysing graphs of sample autocorrelations (ACF) at Figures 2 and 4 we tend to reject the hypothesis. Small p-values of autocorrelation and portmanteau tests in Tables 1 and 2 justify the latter.

Table 1: *Elektrim.*

Test	Value of test statistic	p-value
Autokorrelation, lag 1	$\widehat{\rho}(1)\sqrt{T} = 3.610296$	0.0003
Ljung-Box	$Q_{LB}(1) = 13.0528$	0.0003
	$Q_{LB}(6) = 22.9662$	0.0008
	$Q_{LB}(12) = 29.5553$	0.0032
	$Q_{LB}(24) = 46.3575$	0.0040

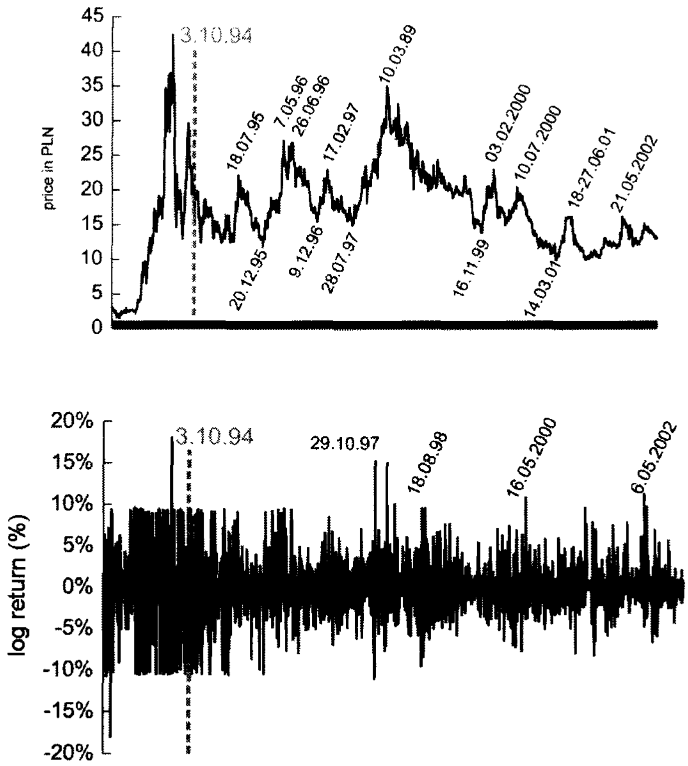

***Figure** 3: Okocim 13.02.92–03.01.03 (2423 observations).*

***Table** 2: Okocim.*

Test	Value of test statistic	p-value
Autocorrelation, lag 1	$\widehat{\rho}(1)\sqrt{T} = 2.474692$	0.0133
Ljung-Box	$Q_{LB}(1) = 6.1221$ $Q_{LB}(6) = 11.9245$ $Q_{LB}(12) = 20.9687$ $Q_{LB}(24) = 32.0897$	0.0133 0.0637 0.0508 0.1248

Ljung-Box statistics $Q_{LB}(h) = T(T+2)\sum_{j=1}^{h} \frac{\widehat{\rho}^2(j)}{T-j}$, based on strong white noise, have for long observation period T approximately the chi-square distribution wih h degrees of freedom. Graphs of partial autocorrelation functions (PACF) at Figures 2 and 4 have specific pikes for lags 1, and for lags bigger than 1 partial autocorrelations are close to 0. One may suspect then that AR(1) model can describe well the observed returns:

$$R_t - \mu = \Phi(R_{t-1} - \mu) + \varepsilon_t. \tag{5}$$

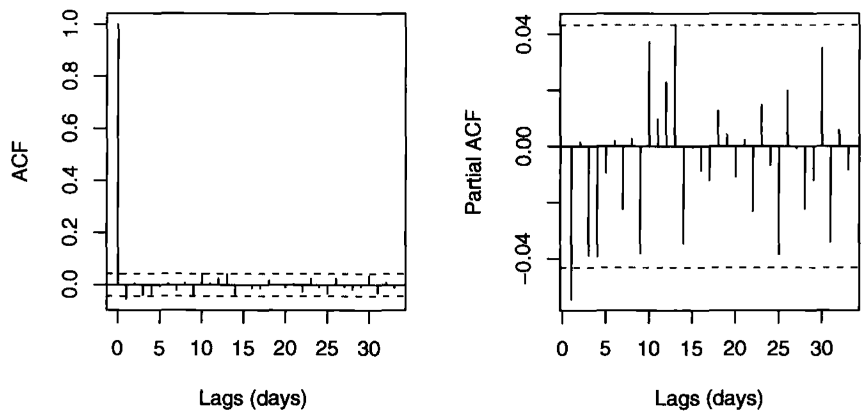

Figure 4: *Okocim – sample ACF and PACF of log returns.*

Assuming (5) and using quasi maximum likelihood method for the Elektrim data set we get the estimated model:

$$R_t + 0.0007429 = 0.07988(R_{t-1} + 0.0007429) + \widehat{\varepsilon}, \tag{6}$$

where estimates of model parameters in (6) maximize maximum likelihood function value obtained for $\{\varepsilon_t\} \sim \text{iid}\, N(0, \sigma)$.

Based on p-values of statistics in Table 3, obtained for sample residuals $\{\widehat{\varepsilon}_t\}$, one cannot reject the hypothesis that residuals $\{\widehat{\varepsilon}_t\}$ form a strong white noise, and thus our log returns follow (5).

Table 3: *Elektrim.*

Test	Value of test statistic	p-value
Autocorrelation, lag 1	$\widehat{\rho}(1)\sqrt{T} = 0.008595$	0,993142
Ljung-Box	$Q_{LB}(1) = 0.0001$	0.9931
	$Q_{LB}(6) = 7.9749$	0.2399
	$Q_{LB}(12) = 15.0624$	0.2380
	$Q_{LB}(24) = 30.9048$	0.1565

Analyzing the graph (Figure 5) of sample residuals $\{\widehat{\varepsilon}_t\}$ one can try to model residuals as a GARCH process since there is some clustering of close in time values. The Lagrange Multiplier test rejects the homoskedasticity hypothesis.

Processes listed in Table 4 have been chosen for farther analysis of residuals $\{\widehat{\varepsilon}_t\}$ behaviour. Particular models' orders minimize values of Akaike AIC$= -2\ln L + 2k$ or Schwarz SBC$= -2\ln L + k \ln n$ criteria, where k is a number of unknown parameters, n is a sample size, L denotes the maximum likelihood function. If more than one model have

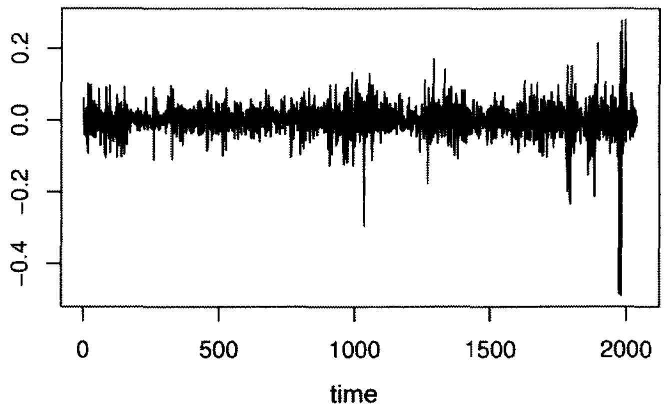

Figure 5: *Elektrim.*

Table 4:

Model	MSE	MAE	MPE	MMEO	MMEU
Homoscedastic	7.80208E-05	0.0023135	12396.3013	0.0122335	0.0310000
GARCH(2,2)	7.46337E-05	0.0021997	9243.13979	0.0126269	0.0265368
GARCH(2,2) with intercept	7.46492E-05	0.0022033	9147.99753	0.0126242	0.0265767
Stationary GARCH(1,3)	7.4127E-05	0.0021899	8985.63329	0.0126725	0.0264366
Stationary GARCH(1,3) with intercept	7.41663E-05	0.0021928	8902.94863	0.0126701	0.0264665
EGARCH(2,1)	7.496E-05	0.0021222	8339.03033	0.0124950	0.0261317
EGARCH(2,1) with intercept	7.49005E-05	0.0021256	8379.59961	0.0124678	0.0262241
Nonsymmetric GARCH(1,1)(1)	7.59514E-05	0.0022561	7785.23459	0.0125085	0.0269828
Nonsymmetric GARCH(2,2)(1)	7.5816E-05	0.0022114	7740.82165	0.0126797	0.0263438
Nonsymmetric GARCH(2,2)(2)	0.000606997	0.0206617	127872.13	0.0213179	0.1377665
GARCH(1,1) NTD	7.43914E-05	0.0022401	8989.68920	0.0126005	0.0266435
GARCH(2,2) NTD	7.50511E-05	0.0022210	9281.23602	0.0126727	0.0266436

comparable values of both criteria, then the model with lower number of parameters or lower p-values of significance parameters' tests is chosen.

Efficiency of model fitting may be evaluated via various measures of errors of squared residuals (estimators of conditional variance of log returns) forecasts, given in Table 4, where

$$MSE = \frac{1}{T}\sum_{t=1}^{T}(\widehat{\varepsilon}_t^2 - \overline{\varepsilon}_t^2)^2 \quad \text{(Mean Squared Error)}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|\widehat{\varepsilon}_t^2 - \overline{\varepsilon}_t^2| \quad \text{(Mean Absolute Error)}$$

$$MPE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{\widehat{\varepsilon}_t^2 - \overline{\varepsilon}_t^2}{\widehat{\varepsilon}_t^2}\right| \quad \text{(Mean Absolute Percentage Error)}$$

$$MMEO = \frac{1}{T}\left(\sum_{t=1}^{T}(\widehat{\varepsilon}_t^2 - \overline{\varepsilon}_t^2)^+ + \sum_{t=1}^{T}\sqrt{(\widehat{\varepsilon}_t^2 - \overline{\varepsilon}_t^2)^-}\right)$$ (Mean Mixed Error of Over-predictions)

$$MMEU = \frac{1}{T}\left(\sum_{t=1}^{T}(\widehat{\varepsilon}_t^2 - \overline{\varepsilon}_t^2)^- + \sum_{t=1}^{T}\sqrt{(\widehat{\varepsilon}_t^2 - \overline{\varepsilon}_t^2)^+}\right),$$ (Mean Mixed Error of Under-predictions),

where $\overline{\varepsilon}_t^2 = E(\hat{\varepsilon}_t^2|\hat{\varepsilon}_{t-1}, \hat{\varepsilon}_{t-2}, \dots) = E(\sigma_t^2 v_t^2|\hat{\varepsilon}_{t-1}, \hat{\varepsilon}_{t-2}, \dots) = \sigma_t^2$,

$$a^+ = \max\{a, 0\}, \quad a^- = \max\{-a, 0\}.$$

Stationary GARCH(1,3) process and stationary GARCH(1,3) process with intercept are best models with respect to MSE measure:
$\hat{\varepsilon}_t = \sigma_t v_t$, where $\sigma_t = 0.000105 + 0.2413\hat{\varepsilon}_{t-1}^2 + 0.0766\sigma_{t-1} + 0.4302\sigma_{t-2} + 0.2026\sigma_{t-3}$,
and
$\hat{\varepsilon}_t - 0.000878 = \sigma_t v_t$, where

$$\sigma_t = 0.000103 + 0.2414(\hat{\varepsilon}_{t-1} - 0.000878)^2 + 0.082\sigma_{t-1} + 0.4343\sigma_{t-2} + 0.1982\sigma_{t-3}.$$

Now, we will try to identify a distribution of the noise $\{v_t\}$ in the latter model. Table 5 presents results of testing the hypothesis that $\{v_t\}$ is the strong white noise. At the 5% significance level one cannot reject the hypothesis.

***Table 5*:**

Test	Value of test statsistic	p-value
Ljung-Box	$Q_{LB}(1) = 3.3112$	0.9156
	$Q_{LB}(2) = 3.2188$	0.1932
	$Q_{LB}(6) = 5.9297$	0.4311
	$Q_{LB}(12) = 11.3775$	0.4969
Turning point test	$Z = 655$	0.0521
Difference-sign test	$S = 1009$	0.9508

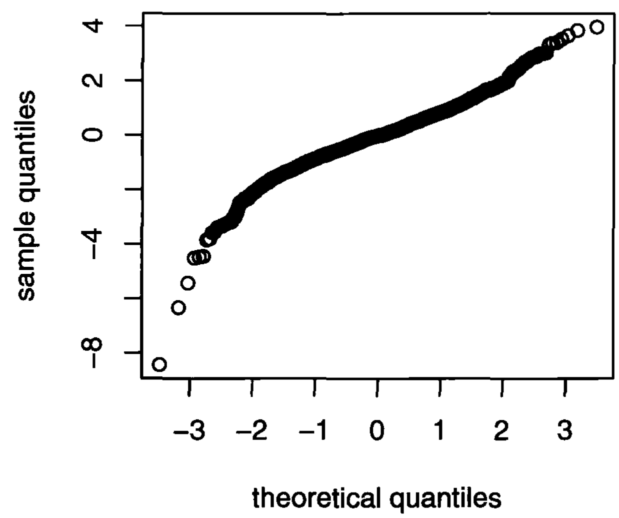

***Figure 6**: Elektrim – quantile plot of residuals of log returns AR(1) model.*

Figure 6 presents the normal quantile plot of $\{v_t\}$ which does not resemble a straight line that suggests not normal distribution, with fat tails. Indeed, values of test statistics of Kolmogorov and Shapiro–Wilks tests are 0.0481 and 0.9565, and corresponding p-values are 0.0001595 and 0.22E-15, respectively. The t-Student, hyperbolic and general error distributions are often used as models of heavy tailed distributions. The density of $\mathrm{GED}(\mu, \sigma, \nu)$ distribution:

$$f(x; \mu, \sigma, \nu) = \frac{\nu}{\lambda 2^{(1+1/\nu)}\Gamma(1/\nu)} \exp\left(-\frac{1}{2}\left|\frac{x-\mu}{\lambda\sigma}\right|^{\nu}\right),$$

where $\lambda = \left(\frac{2^{-2/\nu}\Gamma(1/\nu)}{\Gamma(3/\nu)}\right)^{1/2}$ and μ, $\sigma > 0$, $\nu > 0$ are location, scale and shape parameters, respectively. In particular GED(0,1,2) = N(0,1). If $\nu < 2$ ($\nu > 2$), the density has thicker (thinner) tails than the normal one.

For our sample residuals $\{v_t\}$, maximum likelihood estimators for normal, t-Student, GED, hyperbolic distributions have following values, respectively:

- $\widehat{\mu} = -0.03203267, \widehat{\sigma} = 1.002$
- $\widehat{\nu} = 11$
- $\widehat{\mu} = -0.01862165, \widehat{\sigma} = 0.991337337, \widehat{\nu} = 1.137376404$
- $\widehat{\alpha} = 1.612786198, \widehat{\beta} = -0.05965513, \widehat{\delta} = 0.481525055, \widehat{\mu} = 0.025483705.$

One may measure discrepancy between the fitted and empirical distributions using the functions:

$$h_K(x) = |F_n(x) - F_0(x)|$$

or

$$h_{AD}(x) = h_K(x) / \sqrt{\mathrm{F}_0(x) \cdot (1 - \mathrm{F}_0(x))}, \; x \in (-\infty, \infty),$$

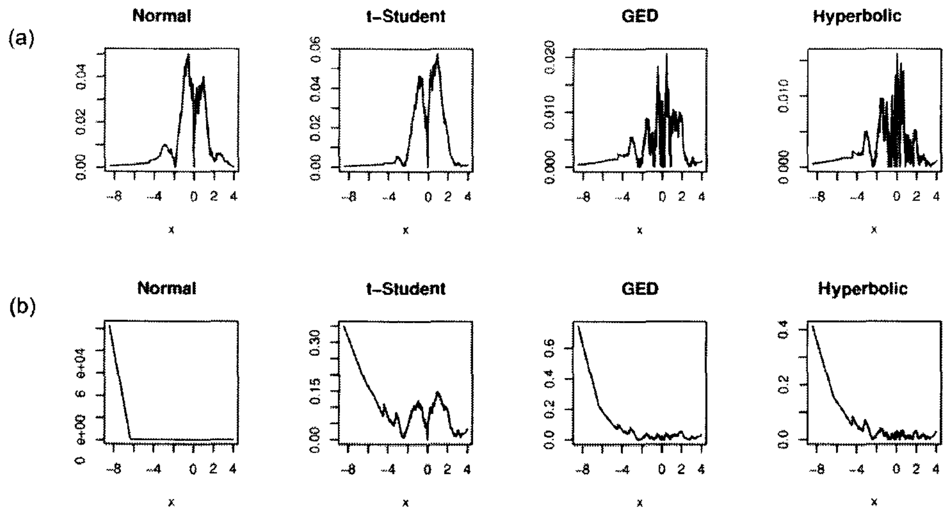

***Figure 7**: Elektrim: (a) Kolmogorow statistics $h_K(x)$ for log return residuals (b) Anderson-Darling statistics $h_{AD}(x)$ for log return residuals.*

where F_n and F_0 are empirical and fitted distribution functions, respectively. The function statistic h_{AD} measures discrepancy in tails of the distributions better than h_K. From graphs at Figure 7 we state that GED and t-Student distribution functions are closer to the empirical one, especially at the tails.

Table 6 presents values of Anderson-Darling $AD = \sqrt{n}\sup\{h_{AD}(x), x \in (-\infty, \infty)\}$ and Kolmogorow–Smirnow $K = \sqrt{n}\sup\{h_K(x), x\in(-\infty, \infty)\}$ statistics with the corresponding p-values. Only GED and hyperbolic distributions are not rejected.

***Table 6**: Elektrim.*

Distribution	Value of AD	Value of K	p-value
Normal	46.03884×10^6	2.269845	0.00001
t-Student	15.807146	2.607259	0.00001
GED	33.60154	0.932649	0.3530
Hyperbolic	18.630039	0.722223	0.6880

Figures 8 and 9 show the histogram with the fitted densities and empirical densities on a logarithmic scale.

In columns 1 of tables 7, 8, 9 (for Elektrim data) there are listed best (according to AIC or SBC criteria) AR(1) – xGARCH models of log returns found in two ways:

- two step procedure described so far
- one step maximum likelihood parameter estimation

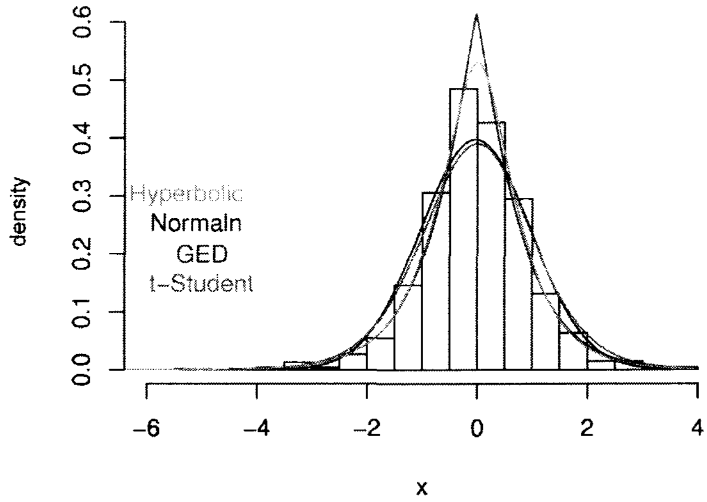

Figure 8: *Elektrim – Histogram of noise* $\{v_t\}$ *in GARCH model of log return residuals.*

Table 7: *Percentage of the observed log returns out of 99% prediction limits.*

No	Noise Log Return Model	Stand. Normal	Normal	GED	Hyper- bolic	t-Student
1	iid standardized return	1.7638	1.7148	0.8329		0.8329
2	1-step est. AR(1)-iid noise	1.7647	1.7647	0.8824		0.8824
3	1-step est. AR(1)-GARCH(1,1)	0.0000	0.6373	0.7843	1.4706	1.4216
4	1-step est AR(1)-GARCH(2,1)	2.3629	2.3529	1.2255	1.3235	1.2255
5	1-step est. AR(1)- station.GARCH(1,2)	2.4510	2.3529	1.2745	1.5196	1.3235
6	1-step est. AR(1)-EGARCH(1,2)	2.3529	2.3529	1.2745	1.2745	1.2745
7	2-step est. AR(1)- station.GARCH(1,3) with intercept	2.4510	2.4510	1.3235	1.4706	1.3235
8	2-step est. AR(1)- station.GARCH(1,3)	2.4510	2.4510	1.3235	1.5196	1.3235
9	2-step est. AR(1)-EGARCH(2,1) with intercept	2.3039	2.3039	1.3725	1.2255	1.2745
10	2-step est. AR(1)-EGARCH(2,1)	2.3039	2.3039	1.3725	1.2745	1.2745
11	GARCH(1,1)	2.3518	2.4008	1.2739	1.5189	1.2739
12	Classic Risk Metrics	1.4063	2.0090	0.9041	0.8036	1.2054

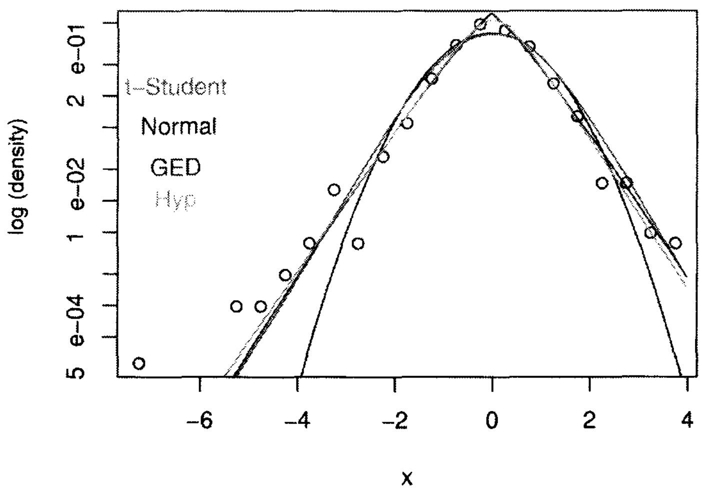

***Figure 9**: Observerd relative frequences and fitted densities from Figure 8 on logarithmic scale.*

***Table 8**: Percentage of the observed log returns out of 95% prediction limits.*

No	Noise Log return model	Stand. Normal	Normal	GED	Hyper-bolic	t-Student
1	iid standardized return	4.9976	4.9976	4.5076		3.2337
2	1-step est. AR(1)-iid noise	5.0490	5.0490	4.7549		3.2353
3	1-step est. AR(1)-GARCH(1,1)	1.8627	5.0000	5.0000	4.5098	3.8725
4	1-step est. AR(1)-GARCH(2,1)	4.6569	4.8529	4.2157	4.3137	3.7255
5	1-step est. AR(1)- station. GARCH(1,2)	5.0000	4.8039	4.2157	4.3137	3.6275
6	1-step est. AR(1)-EGARCH(1,2)	5.1471	5.1961	4.3137	4.6078	3.8725
7	2-step est. AR(1)- station.GARCH(1,3) with intercept	4.8529	4.8529	4.0686	4.4118	3.7254
8	2-step est. AR(1)- station.GARCH(1,3)	4.9020	4.8039	4.1667	4.2647	3.7255
9	2-step est. AR(1)-EGARCH(2,1) with intercept	5.0000	5.1471	4.3627	4.5098	3.7745
10	2-step est. AR(1)-EGARCH(2,1)	5.1471	5.1471	4.4118	4.5098	3.7745
11	GARCH(1,1)	5.0955	5.1445	4.2626	4.6056	3.7237
12	Classic RiskMetrics	4.9292	5.6755	5.0226	5.2235	4.6710

Table 9: *Percentage of the observed log returns out of 90% prediction limits.*

No	Noise Log return model	Stand. Normal	Normal	GED	Hyper- bolic	t-Student
1	iid standardized return	6.9084	6.9574	8.0843		5.6345
2	1-step est. AR(1)-iid noise	6.9608	6.9118	8.1373		5.6373
3	1-step est. AR(1)-GARCH(1,1)	4.9610	10.3431	10.441	8.8235	6.5686
4	1-step est. AR(1)-GARCH(2,1)	8.3333	8.3824	8.6275	9.2157	8.1275
5	1-step est. AR(1)- station.GARCH(1,2)	8.2843	8.2353	8.4314	9.0686	6.3725
6	1-step est. AR(1)-EGARCH(1,2)	8.2353	8.1863	8.6275	8.9706	6.6176
7	2-step est. AR(1)- station.GARCH(1,3) with intercept	8.4804	8.4804	8.6275	9.2157	6.3235
8	2-step est. AR(1)- station.GARCH(1,3)	8.5294	8.4804	8.5784	9.0686	6.3725
9	2-step est. AR(1)-EGARCH(2,1) with intercept	7.9902	8.0882	8.2843	8.9786	6.3725
10	2-step est. AR(1)-EGARCH(2,1)	8.0392	8.0392	8.2843	8.9216	6.3725
11	GARCH(1,1)	8.0843	8.3213	8.4762	9.0152	6.3204
12	Classic RiskMetrics	8.6891	9.6434	9.8945	9.9950	8.3877

References

Blaesid, P. and Sorensen, M. (1992). 'Hyp'—a computer program for analyzing data by means of the hyperbolic distribution, *Research Report*, 248, University Aarhus, Dept. Theor. Statist.

Eberlein, E. and Keller, U. (1995). Hyperbolic distributions in finance, *Bernoulli*, 1, 281-299.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation, *Econometrica*, 50, 987-1008.

Bollerslev, T. (1986). Generalised autoregressive conditional heteroskedasticity, *Journal of Econometrics*, 31, 307-327.

Gąsowski, M. (2002). *Modelling financial returns with nonlinear time series from ARCH family processes*, M. Sc. Thesis, Warsaw University of Technology, Warsaw.

Glosten, L. R. *et al.* (1993). Relationship between the expected value and the volatility of the nominal excess return on stocks, *Journal of Finance*, 48, 1779-1801.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: a new approach, *Econometrica*, 59, 347-370.

Resum

Les rendibilitats financeres es modelen sovint com a sèries temporals auto-regressives amb pertorbacions aleatòries amb variàncies condicionals heterocedàstiques, especialment amb processos de tipus GARCH. Els processos GARCH han estat intensament estudiats en la literatura financera i economètrica com a models de risc de moltes sèries financeres. Analitzant dos conjunts de dades de preus d'actius tractem d'ajustar processos AR(1) amb errors GARCH o EGARCH a les log-rendibilitats. A més, distribucions hiperbòliques o d'errors generalitzats resulten ser bons models de distribucions de soroll blanc.

MSC: Primary 62M10, 91B84; secondary 62M20

Paraules clau: processos auto-regressius, models GARCH i EGARCH, variància condicional heterocedàstica, log-rendibilitats financeres

SORT 28 (1) January-June 2004, 69-86

Improving both domain and total area estimation by composition*

Àlex Costa[1], Albert Satorra[2] and Eva Ventura[2]
Statistical Institute of Catalonia (Idescat) and Universitat Pompeu Fabra

Abstract

In this article we propose small area estimators for both the small and large area parameters. When the objective is to estimate parameters at both levels, optimality is achieved by a sample design that combines fixed and proportional allocation. In such a design, one fraction of the sample is distributed proportionally among the small areas and the rest is evenly distributed. Simulation is used to assess the performance of the direct estimator and two composite small area estimators, for a range of sample sizes and different sample distributions. Performance is measured in terms of mean squared errors for both small and large area parameters. Small area composite estimators open the possibility of reducing the sample size when the desired precision is given, or improving precision for a given sample size.

MSC: 62J07, 62J10, and 62H12

Keywords: regional statistics, small areas, mean square error, direct and composite estimators

1 Introduction

This study stems from a practical issue. The Institut d'Estadística de Catalunya (IDESCAT) had to develop an Industrial Production Index (IPI) for the Catalan autonomous community. The Instituto Español de Estadística (INE) did not produce any

* **Acknowledgements**: The authors are grateful to Xavier López of the Statistical Institute of Catalonia (IDESCAT) for his help in the elaboration of the figures of the paper, and to Nicholas T. Longford for his detail comments on a previous version of this paper. Eva Ventura acknowledges support from the research grants SEC2001-0769 and SEC2003-04476.

[1] *Address for correspondence:* Statistical Institute of Catalonia (Idescat). Via Laietana, 58 - 08003 Barcelona, Spain. E-mail: acosta@idescat.es.

[2] *Address for correspondence:* Department of Economics and Business. Universitat Pompeu Fabra, 08005 Barcelona, Spain E-mail: albert.satorra@econ.upf.es, eva.ventura@econ.upf.es.

Received: February 2004
Accepted: May 2004

regional IPI for Spain, just a national one. IDESCAT had no budget for conducting a Catalan monthly survey. Instead, IDESCAT estimated the IPI for Catalonia using the Spanish IPI of 150 industrial branches, weighted according to their relative importance in Catalonia. This Catalan IPI is a synthetic estimator and was accepted very well by the analysts of the Catalan economy.

The statisticians of IDESCAT had performed a test prior to publishing the new index. The Instituto Vasco de Estadística (EUSTAT) conducted its own regional survey in the Basque Country and published a Basque IPI. IDESCAT created a synthetic index for the Basque Country, using the methodology applied to the Catalan index. This index was compared to EUSTAT's IPI and the results of such comparison evidenced an acceptable performance for the new index. Both the level value of the synthetic IPI and its rate of variation were very useful in order to follow the Basque economic situation (see Costa and Galter 1994). Based on these results, IDESCAT produced a synthetic IPI for Catalonia. Following this, INE applied the same methodology to obtain a distinct IPI for each of the seventeen Spanish autonomous communities.

The method used by IDESCAT is by no means standard in the Spanish official statistics. The synthetic IPI was criticized within some fields, even when it worked well in Catalonia. Some studies (see Clar, Ramon and Surinach 2000) showed that the synthetic IPI works well in regions that possess a significant and quite diversified industry, such as Catalonia. But it fails in other Spanish regions. This observation encouraged the IDESCAT to investigate the theoretical basis of its synthetic IPI from the context of the small area methods.

There is a varied methodology on small area estimation. The reader can consult Platek, Rao, Särndal and Singh (1987), Isaki (1990), Ghosh and Rao (1994), and Singh, Gambino and Mantel (1994) to gain an overview of them. Some of the methods use auxiliary information from related variables in the estimation of area-level quantities. In Spain, recent work by Morales, Molina and Santamaría (2003) deals with small area estimation with auxiliary variables and complex sample designs. We concentrate on methods that use sample information solely from the target variable. These methods include direct and some indirect estimators. Traditional direct estimators use only data from the small area being examined. Usually they are unbiased, but they exhibit a high degree of variation. Indirect, composite and model-based estimators are more precise since they use also observations from related or neighboring areas. Indirect estimators are obtained using unbiased large area estimators. Based on them, estimators can be devised for smaller areas under the assumption that they exhibit the same structure as the large area. Composite estimators are linear combinations of direct and indirect estimators.

The research program on small area estimation carried out jointly by IDESCAT and researchers of the Universitat Pompeu Fabra is characterized by its focus on covariate-free models. An estimator that is based on using auxiliary information from other variables at hand will in general be more efficient, but introduces degrees of subjectivity.

We believe the covariate-free small area estimators are the only ones that are readily usable in the present stage of our official statistics framework.

Costa, Satorra and Ventura (2002) worked with a survey that included direct regional estimators of the Spanish work force. They studied three small area estimators: a synthetic, a direct, and a composite one. The study concluded that the composite estimator and the synthetic estimator were almost identical in Catalonia, because this region's economy is a large component of the whole Spanish economy. The bias of the synthetic estimator was found to be very small for Catalonia.

Costa, Satorra and Ventura (2003) used Monte Carlo methods (with both an empirical[1] and a theoretical population) to compare the performance of several small area estimators: a direct, a synthetic, and three composite estimators. These composite estimators differ in the way the direct and synthetic estimators are combined. One of the composite estimators used theoretical weights (based on known bias and variances). The other two use estimated weights assuming homogeneous or heterogeneous biases and variances across the small areas and concluded that, given the usual sample sizes used in official statistics, the composite estimator based on the assumption of heterogeneity of biases and variances is superior.

Often the statistician is interested on the estimation of both small and large area parameters. In this case, classical estimation methods use sample designs that vary according to the assignment of sample size to the small areas. The following sample designs are considered: a) a proportional design, in which the sample size of each area is proportional to the size of the area in the population; b) uniform design, in which all the areas share the same sample size, regardless of the size of the area, and c) the mixed design, that shares the strategies of a) and b). Clearly, design a) will be optimal when we focus on estimating accurately the large area parameter; while design b) will be chosen when we want to obtain accurate estimates of the small area quantities.

Using Monte Carlo methods on a real population (a labour force census of enterprises) and mixed designs with varying levels in the mixing of uniform and proportional sampling, in the present paper we show how small area estimation improves the estimation of both the small and large are parameters. It will be seen that by using composite small area estimates we can either reduce sample size when precision is given, or improve precision, when sample size is fixed.

The outline of the paper is as follows. Section 2 describes small area estimation. Section 3 describes the fixed, proportional and mixed sampling designs. Section 4 presents the Monte Carlo study. Section 5 describes the results of the Monte Carlo study concerning the direct versus composite estimates. Finally, Section 6 describes how composition improves both large and small area parameters.

1. The empirical population is the Labour Force Census of Enterprises affiliated with the Social Security system in Catalonia. The small areas are the forty-one Catalan counties.

2 Small area estimation without covariates

Suppose a large country area is divided into small area domains $j = 1, 2, \ldots, J$. Let N be the size of the population, and N_1, $N_2, \ldots, N_J$ be the sizes of the J small areas.

Let X be a scalar variable and that we are interested in estimating the mean (or the total) of X for each of the J regions, as well as the overall mean. Let θ_j be the mean of X in the region j, and θ_* be the mean of X in the population. The variance of X in region j is denoted as σ_j^2.

Suppose we have a direct estimator $\hat{\theta}_j$ of the mean of X in each domain, such that $\hat{\theta}_j \sim N\left(\theta_j, \sigma_j^2/n_j\right)$, $j = 1, 2, \ldots, J$, and an estimator $\hat{\theta}_*$ for the large area mean, with $\hat{\theta}_* \sim N\left(\theta_*, \sigma_*^2\right)$. Furthermore, assume a distribution for the mean of area j, $\theta_j \sim N\left(\theta_*, b_j^2\right)$ where b_j^2 is a variance parameter that (possibly) varies with the region.

The design of the survey attends usually to the objective of ensuring precision when estimating the parameters at the country level. For the sake of simplicity, assume that, $\hat{\theta}_*$ is unbiased for θ_* and σ_*^2 is very small. However, some sample surveys have secondary uses of providing information about the small areas. The sample size of most sub-domains is too small, may even be null, to draw accurate inferences about the mean of the small area on the basis of the direct estimate $\hat{\theta}_j$. That is, even though $\hat{\theta}_j$ is an unbiased estimate for θ_j, its variance σ_j^2 is too large to provide an accurate estimation of the small area level parameter.

In this context it is advisable to use composite estimators. They combine linearly the direct estimator and a synthetic (indirect) estimator. The best linear composite estimator of θ_j (in the sense of minimizing the mean squared error, or MSE) is

$$\tilde{\theta}_j = \pi_j \hat{\theta}_* + \left(1 - \pi_j\right) \hat{\theta}_j \tag{1}$$

with

$$\pi_j = \frac{\sigma_j^2/n_j - \gamma_j}{\left(\theta_j - \theta_*\right)^2 + \sigma_j^2/n_j + \sigma_*^2 - 2\gamma_j} \tag{2}$$

where γ_j denotes the covariance between the direct estimator and $\hat{\theta}_*$. For simplicity, assume that the covariance $\gamma_j = 0$ and σ_*^2 is negligible. The value of π_j that minimizes the MSE is

$$\pi_j = \frac{\sigma_j^2/n_j}{\left(b_j^2 + \sigma_j^2/n_j\right)} \tag{3}$$

where $b_j^2 = \left(\theta_j - \theta_*\right)^2$.

The values of the variance σ_j^2 and squared bias b_j^2 are usually unknown, and therefore they must be estimated if we wish to approach the optimal value of π_j in (3).

There are several procedures for estimating these population parameters. In the present study, we use two estimators, the "classic composite" and the "alternative composite", as described and investigated in Costa, Satorra and Ventura (2003).

1. Classic composite estimator

The classic composite estimator assumes that the small areas share the same within-area variance (of the baseline data) and a common estimate for the squared bias. Specifically, we assume $\hat{\theta}_j \sim N\left(\theta_j, \sigma^2/n_j\right)$, $j = 1,2,\ldots,J$, and $\theta_j \sim N\left(\theta_*, b^2\right)$. We obtain the base line variance by a weighted mean of the sample variances from each area as an estimate. Thus, we define the pooled within-area variance

$$\bar{s}^2 = \frac{\sum_{j=1}^{J}\left(n_j - 1\right)s_j^2}{(n - J)}, \tag{4}$$

where n is the size of the entire sample, n_j is the sample size of the small area and s_j^2 is the sample variance of the baseline data of the small area j. If we assume that $\sigma_j^2 = \sigma^2$ for all of j, the estimator of σ_j^2 is $\bar{s}^2$.

For the squared bias $\left(\theta_* - \theta_j\right)^2$, we define the common estimator

$$b^2 = \frac{1}{J}\sum_{j=1}^{J}\left(\hat{\theta}_j - \hat{\theta}_*\right)^2, \tag{5}$$

i.e., the mean squared difference of the direct and indirect estimators.

Thus, the estimator of π_j is:

$$\hat{\pi}_j^c = \frac{\bar{s}^2/n_j}{\bar{s}^2/n_j + b^2}, \tag{6}$$

and the composite estimator obtained by substituting π_j^c for π_j in (1)

$$\tilde{\theta}_j^c = \hat{\pi}_j^c\hat{\theta}_* + \left(1 - \hat{\pi}_j^c\right)\hat{\theta}_j \tag{7}$$

2. Alternative composite estimator

An alternative for the above classic composite estimator is based on direct estimators of each area's variance and bias. In this way the estimator of π_j is:

$$\hat{\pi}_j^a = \frac{s_j^2/n_j}{\left(\hat{\theta}_j - \hat{\theta}_*\right)^2} \tag{8}$$

Note that $\left(\hat{\theta}_j - \hat{\theta}_*\right)^2$ is biased for $(\theta_j - \theta_*)^2$, but is unbiased for $\sigma_j^2/n_j + b_j^2$, as

$$\begin{aligned} E\left(\hat{\theta}_j - \hat{\theta}_*\right)^2 &= E\left(\hat{\theta}_j - \theta_j + \theta_j - \hat{\theta}_*\right)^2 = \\ &= E\left(\hat{\theta}_j - \theta_j\right)^2 + E\left(\theta_j - \hat{\theta}_*\right)^2 + 2E\left(\hat{\theta}_j - \theta_j\right)\left(\theta_j - \hat{\theta}_*\right) = \\ &= \sigma_j^2/n_j + b_j^2 \end{aligned}$$

which leads to the alternative composite estimator

$$\tilde{\theta}_j^a = \hat{\pi}_j^a \hat{\theta}_* + \left(1 - \hat{\pi}_j^a\right) \hat{\theta}_j \tag{9}$$

If necessary, the weight $\hat{\pi}_j^a$ is truncated to one.

3 Survey design with small areas

$$\sum_{j=1}^{J} n_j = n \tag{10}$$

and

$$\frac{n_j}{n} = \frac{N_j}{N} \quad j = 1, 2. \ldots, J \tag{11}$$

where n_j is the size of the sample belonging to area j.

A **purely fixed** survey design assigns the same sample size to each small area. Therefore

$$n_j = \frac{n}{J} \quad \text{for } j = 1, 2, \ldots, J \text{ and } \sum_{j=1}^{n} n_j = n \tag{12}$$

A **mixed** survey design distributes a fraction of the whole sample in a proportional way among the different areas, with the rest of the sample distributed evenly among the areas (Sing, Mantel and Thomas, 1994). Let k be the fraction of the sample to be assigned to the proportional design.

Then

$$n_j = k \frac{N_j}{N} n + (1 - k) \frac{n}{J} \text{ for } j = 1, 2, \ldots, J \text{ and } \sum_{j=1}^{J} n_j = n \tag{13}$$

A pure proportional sample design minimizes MSE for the estimate of the country-level quantity, while a pure fixed design minimizes the MSE of the estimates at the small area level. In the present paper we use simulation to explore the performance of a mixed design strategy when the interest is in minimizing the MSE of estimates of both the country and region level quantities. For that we consider different sample sizes, different survey design strategies, and different estimators.

4 Monte Carlo study

In this section we conduct a Monte Carlo study in which we extract multiple samples from a known population. We use data from the Labour Force Census of Enterprises affiliated with the Social Security system in Catalonia. This census contains data on the number of employees from each surveyed enterprise who are registered with the Social

Table 1: Population characteristics: size, county-mean, square bias, and variance.

	Population size	θ_j	$(\theta_j - \theta_*)^2$	$\sigma_j^2(x)$
Alt Camp	1282	8,73[a]	0,09	3250,37
Alt Empordà	4712	5,28	14,11	294,27
Alt Penedès	3052	8,91	0,02	1686,24
Alt Urgell	745	4,71	18,7	158,25
Alta Ribagorça	140	4,59	19,73	205,38
Anoia	3264	7,86	1,37	801,64
Bages	5698	8,24	0,63	1356,9
Baix camp	5530	6,47	6,59	479,54
Baix Ebre	2237	6,31	7,41	534,4
Baix Empordà	4634	5,44	12,92	425,17
Baix Llobregat	20541	9,73	0,48	1642,46
Baix Penedès	2197	5,26	14,23	171,82
Barcelonès	88331	10,63	2,55	10314,88
Berguedà	1397	5,44	12,9	196,15
Cerdanya	788	3,71	28,34	71,93
Conca de Barberà	611	8,29	0,56	1388,95
Garraf	3466	6,28	7,62	685,91
Garrigues	516	5,24	14,42	96,89
Garrotxa	1909	7,51	2,33	419,72
Gironès	6369	9,82	0,62	2037,47
Maresme	11718	6,46	6,64	605,07
Montsià	1918	5,61	11,73	246
Noguera	1128	5,12	15,3	93,29
Osona	5494	7,09	3,77	774,65
Pallars Jussà	410	4,37	21,76	130,37
Pallars Sobirà	272	4,06	24,76	55,46
Pla d'Urgell	1106	6,59	5,95	271,85
Pla de l'Estany	1160	6,07	8,79	143,37
Priorat	254	4,11	24,26	180,17
Ribera d'Ebre	620	5,71	11,07	418,72
Ripollès	959	7,87	1,35	875,92
Segarra	594	10,87	3,35	8171,41
Segrià	7096	7,74	1,69	714,23
Selva	4586	7,11	3,7	610,2
Solsonès	508	5,58	11,93	157,58
Tarragonès	7440	9,42	0,15	1675,66
Terra Alta	297	4,25	22,87	40,28
Urgell	1178	6,28	7,59	312,25
Val d'Aran	503	5,28	14,08	270,11
Vallès Occidental	26683	10,34	1,71	3026,89
Vallès Oriental	11795	8,45	0,34	832,68

The mean of the affiliates for the whole of Catalonia is 9.04

Security. The census was carried out in each of the four quarters between the years 1992 and 2000 (inclusive). We limit the analysis to one year, 2000.

The database contains 243,184 observations from year 2000, divided into 12 groups according to the economic sector, and 41 counties (Catalan «comarques»). A few enterprises were excluded from the analysis because their locations were not established.

We eliminated the sector-based classification and focused solely on the division to counties. Table1 shows the number of enterprises per county and the mean and variance of the number of employees per enterprise. The distribution of enterprises is quite uneven, as they are concentrated in a few populous areas.

We consider four sample sizes and five alternative survey designs. The smallest sample size is 2,050 observations. We then repeatedly double the sample and use 4,100, 8,200 and 16,400 observations. For each sample size, we consider a purely proportional sample, a 75%, 50% and 25% mixed sample design (respectively), and a purely fixed sample design, that is, combinations with $k = 1, 0.75, 0.50, 0.25$ and 0.

For the overall sample size of $n = 4100$, Table A1 in the Appendix shows the small area sample sizes in each of those 20 (4 by 5) scenarios. As some of the counties have very small population, we have used sampling with replacement. The number of Monte Carlo replications is 1,000. We evaluate the direct, classic composite and alternative composite estimators for each of the 41 counties as well as for the whole of Catalonia.

5 Direct vs. composite estimators

We computed the MSE of each small area estimator. Table 2 shows a summary of descriptive statistics. The mean, median, variance, minimum and maximum values of the MSE of the small area estimators across the 1,000 replications are presented. Table 3 evaluates the relative performance of the three alternative estimators differently. In that table we calculate the percentage of counties for which the MSE of a particular estimator (in the leftmost column) is lower than the MSE of the other two estimators.

The performance of the small area estimations can be evaluated by several criteria. We have observed in Costa, Satorra and Ventura (2003) that the distribution of the MSE across the Catalan counties is asymmetric. It also exhibits extreme values and it is very dispersed. This is a consequence of the extremely uneven distribution of the population and economic activity in the region. This is a drawback of the simple average of the MSE of the counties. The median is affected less by the presence of extreme values. On the other hand, we may want to put one upper limit on the MSE for each small area. Looking at the maximum MSE of the counties is then appropriate. To keep things simple we present the results graphically using the median evaluation criterion. Tables 2 and 3 and Figure 4 show the results using other evaluation criteria. Those criteria are the average of the MSE of the counties, the maximum and minimum values of the MSE of the counties, and the percentage of counties for which a particular type of estimator

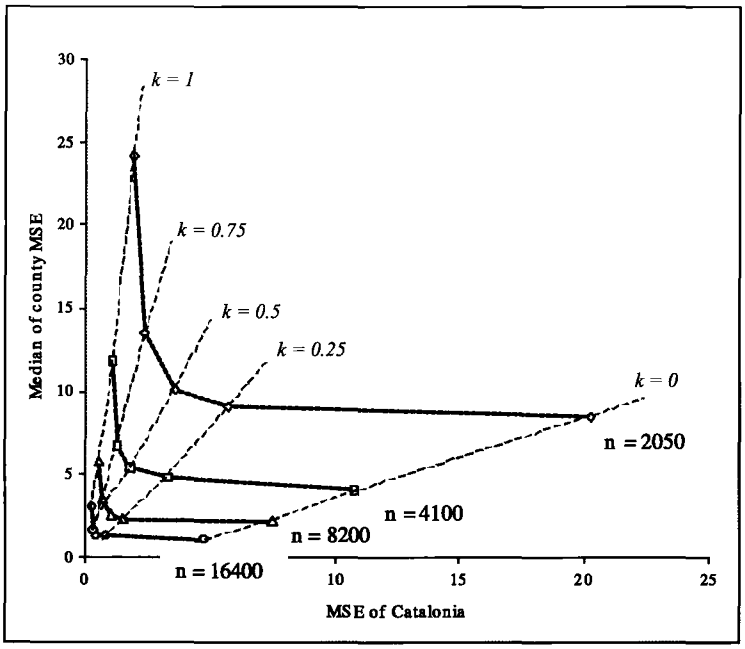

***Figure 1**: MSE of direct estimator for various combinations of sample size (n) and sampling design (k).*

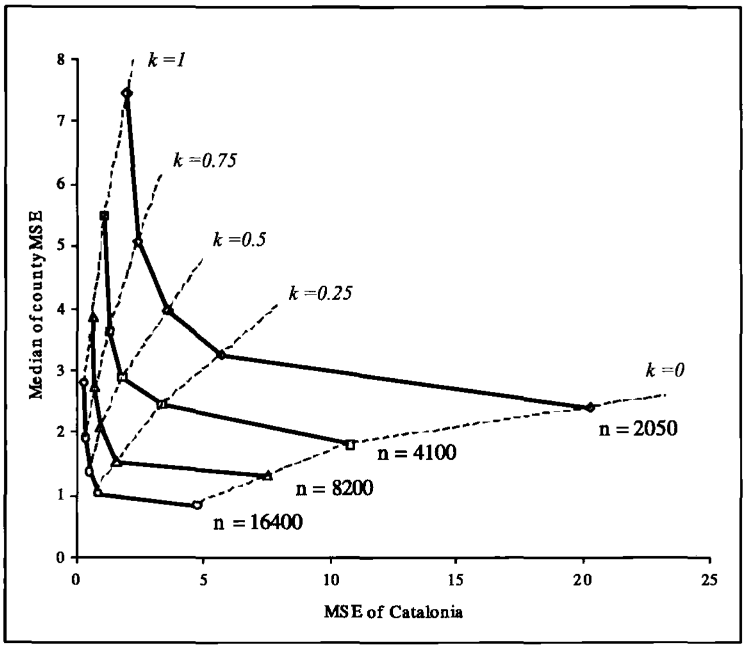

***Figure 2**: MSE of composite alternative estimator for various combinations of sample size (n) and sampling design (k).*

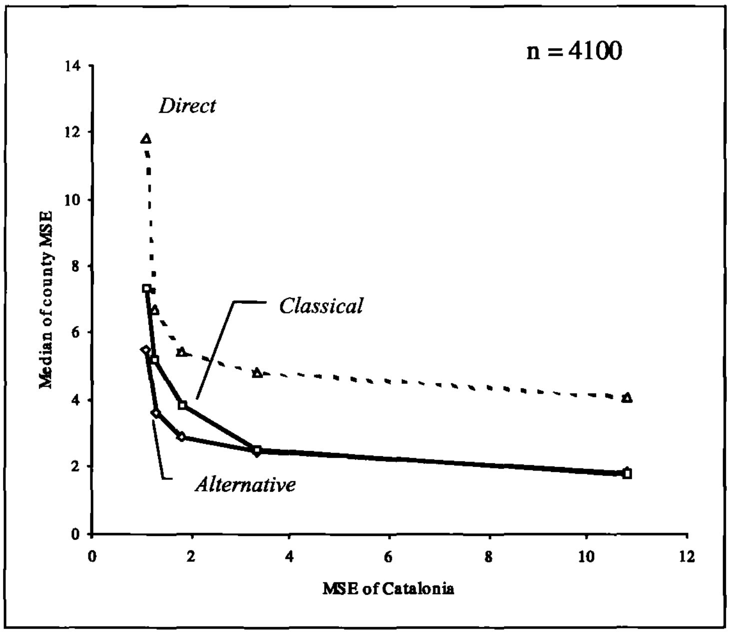

Figure 3: *Comparing three estimators for sample n = 4100.*

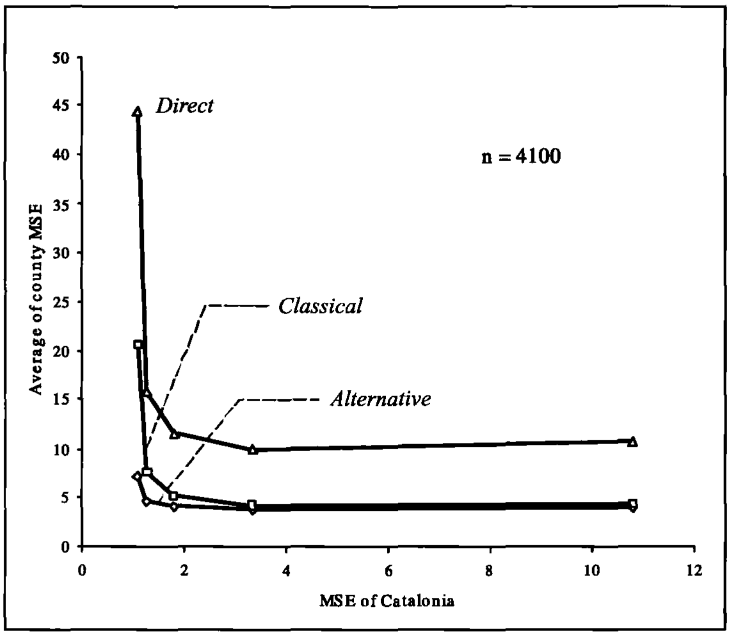

Figure 4: *Comparing three estimators for sample n = 4100 by the criteria of the average county MSE.*

performs better than the other two estimators. Figures 1 to 4 summarize the results visually.

Table 2: *Descriptive statistics of the MSE of the small area estimators, by sample size and sampling choice.*

n = 2050

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Mean	20,45	7,19	5,32	19,20	6,83	5,24	23,22	8,89	5,67	31,05	13,19	6,62	79,69	37,93	8,82
Median	8,43	2,91	2,41	9,12	4,32	3,26	10,04	5,30	3,96	13,49	7,63	5,07	24,20	11,12	7,45
Variance	1141,20	122,53	43,41	1006,35	78,88	27,18	2275,40	177,32	30,53	4737,04	577,39	41,89	48928,68	16967,77	43,10
Min	0,80	1,10	1,34	1,05	1,84	1,76	1,58	2,83	2,66	2,87	3,93	2,97	5,79	4,53	2,99
Max	158,64	55,38	36,97	194,64	56,89	31,89	299,66	87,70	37,24	423,58	154,48	44,76	1382,42	837,40	41,17

n = 4100

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Mean	10,62	4,35	3,90	9,91	4,25	3,78	11,49	5,26	4,06	15,74	7,58	4,75	44,39	20,61	7,23
Median	4,07	1,80	1,82	4,83	2,50	2,46	5,41	3,86	2,90	6,70	5,19	3,63	11,84	7,34	5,51
Variance	298,21	40,74	26,25	260,30	24,03	19,18	484,49	28,65	15,75	1218,45	71,44	17,25	20158,91	4111,94	73,94
Min	0,41	0,57	0,61	0,54	1,21	0,80	0,77	1,50	1,26	1,42	1,94	1,56	2,88	2,27	1,70
Max	78,71	30,84	28,32	94,56	30,84	27,74	133,82	34,52	26,57	212,70	53,80	28,13	904,32	417,57	56,99

n = 8200

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Mean	5,70	2,83	2,83	4,88	2,73	2,64	6,01	3,72	2,94	8,39	5,36	3,53	21,21	10,93	5,24
Median	2,26	1,21	1,32	2,28	1,63	1,54	2,53	2,89	2,10	3,54	3,64	2,74	5,82	5,41	3,88
Variance	112,27	19,88	17,66	66,30	9,59	13,59	145,38	11,38	14,13	408,74	21,91	16,75	3998,63	323,97	34,33
Min	0,19	0,37	0,24	0,28	0,68	0,33	0,40	1,00	0,51	0,74	1,34	0,96	1,60	1,51	1,02
Max	54,53	21,02	23,29	49,34	20,06	23,42	73,93	21,20	25,23	125,17	26,32	27,95	397,33	115,13	38,37

n = 16400

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Mean	2,85	1,75	1,95	2,45	1,78	1,75	2,89	2,45	1,87	4,28	4,03	2,45	10,96	7,38	3,86
Median	1,07	0,72	0,83	1,28	1,16	1,02	1,32	1,87	1,36	1,59	2,50	1,92	2,97	3,76	2,80
Variance	30,51	7,57	9,54	17,80	3,81	7,38	32,95	4,58	7,51	125,96	13,68	13,57	1295,79	70,52	26,49
Min	0,11	0,24	0,12	0,12	0,42	0,13	0,18	0,59	0,20	0,36	0,67	0,43	0,68	0,74	0,62
Max	30,92	13,42	16,78	26,17	12,75	17,36	35,43	12,68	18,26	70,87	17,97	24,85	230,29	42,83	34,26

Table 3: *Comparing estimators under the percentage criterion.*

n = 2050

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Direct		7,32%	7,32%		17,07%	12,20%		19,51%	12,20%		24,39%	7,32%		12,20%	0,00%
Classical	92,68%		34,15%	82,93%		19,51%	80,49%		12,20%	75,61%		7,32%	87,80%		7,32%
Alternative	92,68%	65,85%		87,80%	80,49%		87,80%	87,80%		92,68%	92,68%		100,00%	92,68%	

n = 4100

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Direct		14,63%	21,95%		24,39%	21,95%		34,15%	24,39%		39,02%	17,07%		34,15%	7,32%
Classical	85,37%		56,10%	75,61%		39,02%	65,85%		29,27%	60,98%		21,95%	65,85%		12,20%
Alternative	78,05%	43,90%		78,05%	60,98%		75,61%	70,73%		82,93%	78,05%		92,68%	87,80%	

n = 8200

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Direct		21,95%	31,71%		34,15%	36,59%		41,46%	39,02%		41,46%	39,02%		41,46%	26,83%
Classical	78,05%		68,29%	65,85%		60,98%	58,54%		43,90%	58,54%		41,46%	58,54%		17,07%
Alternative	68,29%	31,71%		63,41%	39,02%		60,98%	56,10%		60,98%	58,54%		73,17%	82,93%	

n = 16400

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative	Direct	Classical	Alternative
Direct		21,95%	43,90%		39,02%	46,34%		43,90%	53,66%		53,66%	60,98%		53,66%	43,90%
Classical	78,05%		60,98%	60,98%		58,54%	56,10%		48,78%	46,34%		43,90%	46,34%		26,83%
Alternative	56,10%	39,02%		53,66%	41,46%		46,34%	51,22%		39,02%	56,10%		56,10%	73,17%	

The horizontal axis corresponds to the MSE of Catalonia $MSE(\hat{\theta}_*)$. The vertical axis corresponds to the median of the 41 counties MSE. The figures show the behaviour of the three estimators considered for different total sample sizes and alternative sample designs. In Figure 3 we overlay the curves of the estimators. We only draw the curves for sample sizes 4,100 and 8,200, to avoid excessive clutter.
The results can be summarized as follows:

- From Table 2 and figures 1 to 3 we see that the MSE for Catalonia is smaller when $k = 1$ and larger when$k = 0$. In contrast, the median and mean county MSE is smaller when $k = 0$ and larger when$k = 1$. This result holds for all three estimators.
- Both the MSE for Catalonia and the median county MSE are reduced as the total sample size is increased, for each estimator. The same result holds for other summaries, such as the average county MSE, or the maximum values of MSE (see Figure 4).
- On average, the alternative composite estimator is the best estimator, as assessed by the MSE of Catalonia and median county MSE, for any sample size and values of k of the mixed design. There are some exceptions, when the total sample sizes are large and we use a fixed sample survey design. This is seen in Figure 3, where, to avoid clutter, we only show the sample size $n = 4100$.
- In Table 3 we see that the two composite estimators are the best in almost all settings. There are some exceptions when the total sample size is large. As expected, the direct estimator improves its behavior as the total sample size increases, independently of the survey design.
- Table 3 shows also that the alternative composite estimator is better than the classical one except for sampling designs with $k = 0$ or $k = 0.25$ (nonproportional survey designs).
- To achieve a particular combination of a small MSE for the large area jointly with small median or average small area MSE, a mixed design strategy is recommended. The desired combination will depend on the preferences about how to use the estimators.

6 Improving survey design by composition

The results in the previous section suggest some clear guidelines for how to improve both the sampling design and the estimation. We examine them now.

Assume that we start with a predetermined sample size and a mixed-design allocation, partly proportional and partly of equal sample sizes (the same sample size in all the counties). More specifically, suppose the budget allows to extract a sample of size $n = 8{,}200$. A fraction of these observations (for example, 35%) is distributed proportionally among the small areas and the remainder is distributed evenly. That means that each Catalan county would have at least 0.65 * 8200/41 = 130 observations. Some counties with a large population would have up to 500 more observations, while the sample size of others would not surpass 150.

If we decide to use a direct estimator for each small area as well as the county, we will obtain the MSE for large and small areas that corresponds approximately to point A (k = 0.35) in Figure 5.

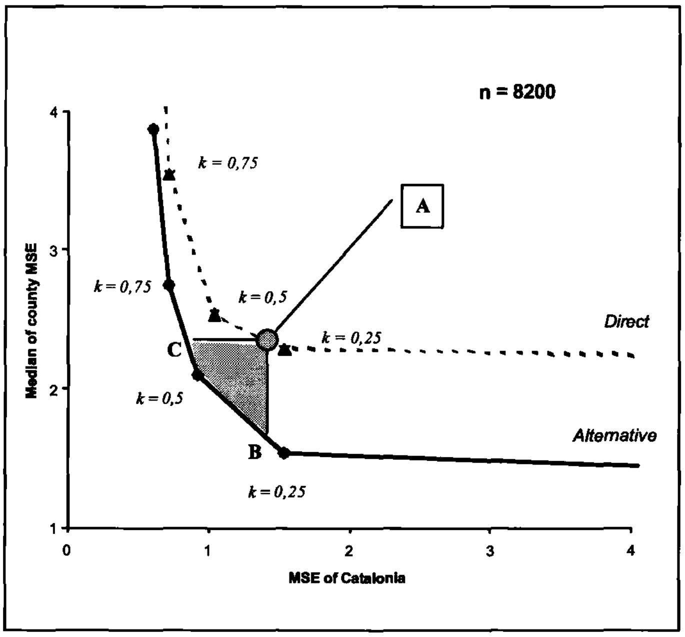

Figure 5: *Opportunities for improvement using composite estimators.*

This figure shows that using the alternative composite estimator (the same would occur if we choose the classical composite estimator) improves estimation over the use of a direct estimation (i.e., non composition). There is a continuum of choices between points B and C (that correspond to varying values of k) that achieves lower or equal MSE for the estimate of both the large and small area parameters. The point B, which has k= 0.35, shows that the use of a composite estimator reduces the median of county MSE, while keeping the MSE of the estimate of Catalonia at the same level than the direct estimate. This point B, with k = 0.35, is the limit that we can move toward egalitarian sampling design, without loosing (by composition) in the estimation of the large area parameter. As we move towards point C, thus adopting a more proportional sampling design, composition improves both the median of county MSE and the MSE of Catalonia over direct estimation. Point C, with k = 0.65, represents the limit we can move on a more proportional design, without loosing (by composition) in the estimation of the small area parameters (i.e., without increasing the median of county MSE over the value obtained with the direct estimate).

That is, the adoption of a small area estimate such as the alternative composite estimator (the classical composite estimator would lead also to the same phenomena) brings room for improvement in the precision of the estimates of both the large and small area parameters, over the use of the simple direct estimation. The main aim of this paper was to illustrate this issue using Monte Carlo data on a real population. We left for further work the development of specific tables to be used for the choice of the required sample sizes and values of k required for attaining a priori specified precision.

References

Clar, M., Rarmos and Suriñach, J. (2000). Avantatges i inconvenients de la metodologia del INE per elaborar indicadors de la producció industrial per a les regions espanyoles, *Qüestiió*, 24, 1, 151-186.

Costa, A. and Galter, J. (1994). L'IPPI, un indicador molt valuós per mesurar l'activitat industrial catalana, *Revista d'Indústria*, 3, Generalitat de Catalunya, 6-15.

Costa, A., Satorra, A. and Ventura, E. (2002). Estimadores compuestos en estadística regional: Aplicación para la tasa de variación de la ocupación en la industria, *Qüestiió*, 26, 1-2, 213-243.

Costa, A., Satorra, A. and Ventura, E. (2003). An empirical evaluation of small area estimators, *SORT (Statistics and Operations Research Transactions)*, 27, 1, 113-135.

Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: an appraisal, *Statistical Science*, 9, 1, Statistics Canada, 55-93.

Isaki, C.T. (1990). Small-Area Estimation of Economic Statistics, *Journal of Business & Economic Statistics*, 8, 4, 435-441.

Morales, D., Molina, I. and Santamaría, L. (2003). Modelos estándar para estimaciones en áreas pequeñas. Extensión a diseños complejos. *Proceedings of the XXVII Congreso Nacional de Estadística e Investigación Operativa* (Lleida, Abril de 2003).

Platek, R., Rao, J.N.K., Särndal, C.E. and Singh, M.P. (eds.) (1987). *Small Area Statistics: An International Symposium*. New York. John Wiley and Sons.

Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and Strategies for Small Area Data, *Survey Methodology*, 20, Statistics Canada, 3-22.

Resum

Els estimadors de petita àrea poden ser utilitzats no només per aproximar els paràmetres d'una petita àrea, sinó també per estimar els paràmetres de l'àrea gran. Quan l'objectiu és estimar un paràmetre en totes dues àrees, l'estratègia òptima s'aconsegueix mitjançant un disseny mostral en dues parts: una part que es distribueix proporcionalment entre les petites àrees i una altra part que es distribueix fixa. S'utilitza un mètode de simulació per avaluar el comportament tant de l'estimador directe com de dos estimadors compostos de petita àrea. La bondat de les estimacions es valora en termes de l'error quadràtic mitjà dels estimadors dels paràmetres de les dues àrees, la gran i la petita. Els estimadors compostos de petita àrea obren la possibilitat bé de reduir la mida mostral quan el nivell de precisió està donat, bé de millorar la precisió quan la mida mostral està donada.

MSC: 62J07, 62J10, and 62H12

Paraules clau: estadística regional, petites àrees, error quadràtic mitjà, estimadors directe i compost

APPENDIX

***Table** A.1: Sample sizes of the small areas, by sampling choice, for n = 4100.*

	k = 0			k = 0.25			k = 0.5			k = 0.75			k = 1		
	Proportional	**Fixed**	**Total**	**Proportional**	**Fixed**	**Total**	**Proportional**	**Fixed**	**Total**	**Proportional**	**Fixed**	**Total**	**Proportional**	**Fixed**	**Total**
Alt Camp	0	100	100	6	74	80	10	50	60	16	26	42	22	0	22
Alt Empordà	0	100	100	20	74	94	40	50	90	58	26	84	80	0	80
Alt Penedès	0	100	100	14	74	88	26	50	76	38	26	64	52	0	52
Alt Urgell	0	100	100	4	74	78	6	50	56	10	26	36	12	0	12
Alta Ribagorça	0	100	100	2	74	76	2	50	52	2	26	28	2	0	2
Anoia	0	100	100	14	74	88	28	50	78	40	26	66	56	0	56
Bages	0	100	100	26	74	100	48	50	98	72	26	98	96	0	96
Baix Camp	0	100	100	24	74	98	46	50	96	70	26	96	94	0	94
Baix Ebre	0	100	100	10	74	84	18	50	68	28	26	54	38	0	38
Baix Empordà	0	100	100	20	74	94	40	50	90	58	26	84	78	0	78
Baix Llobregat	0	100	100	88	74	162	174	50	224	254	26	280	346	0	346
Baix Penedès	0	100	100	10	74	84	18	50	68	28	26	54	38	0	38
Barcelonès	0	100	100	386	74	460	742	50	792	1100	26	1126	1488	0	1488
Berguedà	0	100	100	6	74	80	12	50	62	18	26	44	24	0	24
Cerdanya	0	100	100	4	74	78	6	50	56	10	26	36	14	0	14
Conca de Barberà	0	100	100	2	74	76	6	50	56	8	26	34	10	0	10
Garraf	0	100	100	16	74	90	30	50	80	44	26	70	58	0	58
Garrigues	0	100	100	2	74	76	4	50	54	6	26	32	8	0	8
Garrotxa	0	100	100	8	74	82	16	50	66	24	26	50	32	0	32
Gironès	0	100	100	28	74	102	54	50	104	80	26	106	108	0	108
Maresme	0	100	100	52	74	126	98	50	148	146	26	172	198	0	198
Montsià	0	100	100	8	74	82	16	50	66	24	26	50	32	0	32
Noguera	0	100	100	6	74	80	10	50	60	14	26	40	20	0	20
Osona	0	100	100	24	74	98	46	50	96	68	26	94	92	0	92
Pallars Jussà	0	100	100	2	74	76	4	50	54	6	26	32	8	0	8
Pallars Sobirà	0	100	100	2	74	76	2	50	52	4	26	30	4	0	4
Pla d'Urgell	0	100	100	4	74	78	10	50	60	14	26	40	18	0	18
Pla de l'Estany	0	100	100	6	74	80	10	50	60	14	26	40	20	0	20
Priorat	0	100	100	2	74	76	2	50	52	4	26	30	4	0	4
Ribera d'Ebre	0	100	100	2	74	76	6	50	56	8	26	34	10	0	10
Ripollès	0	100	100	4	74	78	8	50	58	12	26	38	16	0	16
Segarra	0	100	100	2	74	76	6	50	56	8	26	34	10	0	10
Segrià	0	100	100	32	74	106	60	50	110	88	26	114	120	0	120
Selva	0	100	100	20	74	94	38	50	88	58	26	84	78	0	78
Solsonès	0	100	100	2	74	76	4	50	54	6	26	32	8	0	8
Tarragonès	0	100	100	32	74	106	62	50	112	92	26	118	126	0	126
Terra Alta	0	100	100	2	74	76	2	50	52	4	26	30	6	0	6
Urgell	0	100	100	6	74	80	10	50	60	14	26	40	20	0	20
Val d'Aran	0	100	100	2	74	76	4	50	54	6	26	32	8	0	8
Vallès Occidental	0	100	100	116	74	190	226	50	276	332	26	358	448	0	448
Vallès Oriental	0	100	100	50	74	124	100	50	150	148	26	174	198	0	198
TOTAL	0	4100	4100	1066	3034	4100	2050	2050	4100	3034	1066	4100	4100	0	4100

SORT 28 (1) January-June 2004, 87-108

Incorporating patients' characteristics in cost-effectiveness studies with clinical trial data: a flexible Bayesian approach

F. J. Vázquez-Polo and M. A. Negrín-Hernández*

University of Las Palmas de G.C., Spain

Abstract

Most published research on the comparison between medical treatment options merely compares the results (effectiveness and cost) obtained for each treatment group. The present work proposes the incorporation of other patient characteristics into the analysis. Most of the studies carried out in this context assume normality of both costs and effectiveness. In practice, however, the data are not always distributed according to this assumption. Altervative models have to be developed.

In this paper, we present a general model of cost-effectiveness, incorporating both binary effectiveness and skewed cost. In a practical application, we compare two highly active antiretroviral treatments applied to asymptomatic HIV patients.

We propose a logit model when the effectiveness is measured depending on whether an initial purpose is achieved. For this model, the measure to compare treatments is the difference in the probability of success. Besides, the cost data usually present a right skewing. We propose the use of the log-transformation to carry out the regression model. The three models are fitted demonstrating the advantages of this modelling. The cost-effectiveness acceptability curve is used as a measure for decision-making.

MSC: 62F15, 62H12, 62P10

Keywords: Bayesian analysis, cost-effectiveness, Markov Chain Monte Carlo (MCMC), skewed cost distributions

1 Introduction

The frequentist approximation is the one most commonly adopted to compare different treatment options (Laska *et al.*, 1997, Stinnett and Mullahy, 1998, Tambour *et al.*

* *Address for correspondence:* Dept. of Quantitative Methods. University of Las Palmas de G.C. Fac. CC. Económicas. Campus de Tafira. 35017 Las Palmas de Gran Canaria, Spain.
E-mail: {fjvpolo or mnegrin}@dmc.ulpgc.es. Phone: +34 928 45 18 00 Fax: +34 928 45 82 25
Received: September 2003
Accepted: March 2004

1998, Van Hout *et al.*, 1994, Wakker and Klaassen, 1995, Willam and O'Brien, 1996). However, clinical research is fundamentally a dynamic process in which any study must be considered in the context of continual updating of the state of the art. The Bayesian method is of a dynamic nature in which initial beliefs, determined on the basis of a prior distribution, are modified by new data, using Bayes' theorem. A large body of literature has been published on Bayesian methods, chief among which are texts by Berry (1996), Box and Tiao (1973) and Gelman *et al.* (1995).

Spiegelhalter *et al.* (1994) and Jones (1996) were the first to discuss Bayesian approximation for statistical inference in the comparison of health technologies. Since then, many studies have proposed the Bayesian approach to compare treatment options by means of cost-effectiveness analysis (Al and Van Hout, 2000, Briggs, 1999, 2001, Heitjan, 1997, Heitjan *et al.*, 1999, O'Hagan *et al.*, 2001, O'Hagan and Stevens, 2001*a*, 2001*b*, 2002).

Most studies carried out in this field compare the effectiveness and the costs of the different treatment options analysed. This type of analysis assumes that the patients sampled and subjected to a particular treatment option present similar characteristics or, at least, that the differences between samples are not relevant to the analysis of cost and effectiveness, and so the variations between the treatment groups are only caused by the type of treatment applied. In the present paper, the above assumption is not made and so, in order to obtain the true effect of the type of treatment applied on costs and effectiveness a regression model is proposed. The use of regression models in cost-effectiveness analysis has recently been proposed by Hoch *et al.* (2002) and Willan *et al.* (2004) under a frequentist point of view. This paper presents the Bayesian solution, offering a more flexible framework for different measures of effectiveness and cost.

Sometimes effectiveness is not measured quantitatively but in a discrete way, depending on whether or not a particular objective has been attained. Therefore, we have developed two alternative regression models, a multiple linear regression model to be used when the effectiveness is measured by means of a continuous variable, and a logit discrete choice model when effectiveness is defined by a categorical variable.

Most published studies on cost-effectiveness analysis assume normality of the cost generation distribution (Laska, 1997, Stinnet and Mullahy, 1998, Tambour *et al.*, 1998, Willam and O'Brien, 1996, Heitjan *et al.*, 1999, O'Hagan *et al.*, 2001). In practice, however, costs usually present a high degree of skewness, and so the normality assumption is not valid. O'Hagan and Stevens (2001*b*) determined, from a practical application, the importance of dealing with skewed cost data, obtaining different results from those achieved under the assumption of normality.

The standard measure used to compare the cost and effectiveness of treatments is the incremental cost-effectiveness ratio (ICER). Nevertheless, this measure presents severe interpretational problems, as well as difficulties in estimating the confidence or credibility intervals. The incremental net benefit (INB) has been proposed as an alternative to ICER (Mullahy and Stinnett, 1998, among others). The INB of treatment

1 (new) versus treatment 0 (actual, or control) is defined as

$$INB(R_c) = R_c \cdot (\mu_1 - \mu_0) - (\gamma_1 - \gamma_0) = R_c \cdot (\Delta\mu) - (\Delta\gamma), \tag{1.1}$$

where μ's and γ's are the mean effectiveness and cost of the respective treatments. The value R_c is interpreted by O'Hagan and Stevens (2001a) as the cost that decision-makers are willing to accept in order to increase the effectiveness of the treatment applied by one unit. Thus, analysing whether the alternative treatment is more cost-effective than the control treatment is equivalent to determining whether $INB(R_c)$ is positive. In practice, it is not a simple matter for the decision-maker to determine a single R_c, and so a cost-effectiveness acceptability curve (CEAC) is constructed (Löthgren and Zethraeus, 2000). This curve provides a graphical representation of the probability of the alternative treatment being preferred ($\Pr(INB(R_c) > 0)$) for each value R_c. This interpretation of the CEAC, in terms of probability, is only possible when the Bayesian approach is adopted (Briggs, 1999).

Section 2 presents the regression models used in this study. These are selected depending on how the effectiveness is to be measured (qualitatively or quantitatively) and on the cost patterns generated. Section 3 provides a comparison of the different models created by means of a practical application using real data from a clinical trial comparing two alternative treatments for asymptomatic HIV patients. Section 4 presents a discussion of the results obtained and draws some conclusions.

2 Bayesian cost-effectiveness regression models incorporating covariates

2.1 Assumed normality of effectiveness and costs

Given a sample of N individuals participating in a clinical trial, we obtained effectiveness data (E_i) and cost data (C_i) for each patient i, $i = 1 \ldots N$. These N patients were given two different types of treatment, termed the control treatment and the new, or alternative treatment.

The results of the clinical trial, in terms of effectiveness and costs, are not determined only by the type of treatment received (X_T), and so it is necessary to consider a series of possible covariates that may influence the above results. Such covariates include the patient's age, state of health at the time of the clinical trial, gender and other characteristics that depend on the type of clinical trial under analysis (X). We define X as an $n \times (k+1)$ matrix of covariates, where each column (X_i) refers to one covariate. The first column is a column of ones referring to the constant.

We seek, therefore, to explain the results obtained (E_i and C_i), as a linear combination of the k covariates considered (the patient's individual characteristics and the type of treatment received). For this purpose, we propose a Bayesian multiple linear regression model in which the perturbation term (u_i or v_i) is assumed to be Gaussian, independent and identically distributed (i.i.d) with a mean of 0 and variances of σ_1^2 and σ_2^2 respectively.

$$E_i = \beta_0 + \beta_1 \cdot X_{1,i} + \beta_2 \cdot X_{2,i} + \ldots + \beta_{k-1} \cdot X_{k-1,i} + \beta_T \cdot X_{T,i} + u_i, \tag{2.1}$$

$$C_i = \delta_0 + \delta_1 \cdot X_{1,i} + \delta_2 \cdot X_{2,i} + \ldots + \delta_{k-1} \cdot X_{k-1,i} + \delta_T \cdot X_{T,i} + v_i, \tag{2.2}$$

where the vectors $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{k-1}, \beta_T)'$, $\delta = (\delta_0, \delta_1, \delta_2, \ldots, \delta_{k-1}, \delta_T)'$, and the accuracy values $\tau_1 = 1/\sigma_1^2$ and $\tau_2 = 1/\sigma_2^2$ are the parameters of the model.

The k covariates considered for which data are available need not be explicative of both the effectiveness and the costs, and so the above general model could be corrected by eliminating those covariates that do not explain effectiveness and cost.

The first step to be taken in estimating the parameters is to determine the likelihood function, both of the effectiveness $\ell_e(E|\beta, \tau_1)$ and of the costs $\ell_c(C|\delta, \tau_2)$, where $E = (E_1, \ldots, E_N)'$ and $C = (C_1, \ldots, C_N)'$. In this stage both costs and effectiveness are assumed to present a normal distribution, and so the likelihood functions are represented by the following expressions:

$$\ell(E, C|\beta, \delta, \tau_1, \tau_2) = \ell_e(E|\beta, \tau_1) \cdot \ell_c(C|\delta, \tau_2), \tag{2.3}$$

where

$$\ell_e(E|\beta, \tau_1) \propto \tau_1^{\frac{N}{2}} \exp\left\{-\frac{\tau_1}{2}(E - X\beta)'(E - X\beta)\right\},$$

and

$$\ell_c(C|\delta, \tau_2) \propto \tau_2^{\frac{N}{2}} \exp\left\{-\frac{\tau_2}{2}(C - X\delta)'(C - X\delta)\right\}.$$

Assuming model (2.1)-(2.2) from a Bayesian point of view, we must specify the prior distribution for the $2 \cdot k + 4$ parameters of the model. The prior distribution represents expert information about the set of model parameters before the sample observations are analysed. We propose a normal/gamma form for the base prior and assume independence between the coefficients (β, δ) and precision terms (τ_1, τ_2). Obviously, the prior distributions used here are not the only possible choices and indeed, their independent conditional conjugate form is a suitable property to be considered by an expert.

$$\pi(\beta, \tau_1) = \pi_{e,1}(\beta) \cdot \pi_{e,2}(\tau_1), \tag{2.4}$$

$$\pi(\delta, \tau_2) = \pi_{c,1}(\delta) \cdot \pi_{c,2}(\tau_2), \tag{2.5}$$

where

$$\pi_{e,1}(\beta) \sim \mathcal{N}(\beta^0, V_1^{-1}), \qquad \text{and} \qquad \pi_{c,1}(\delta) \sim \mathcal{N}(\delta^0, V_2^{-1}),$$

and,

$$\pi_{e,2}(\tau_1) \sim \mathcal{G}(a_1, b_1), \qquad \text{and} \qquad \pi_{c,2}(\tau_2) \sim \mathcal{G}(a_2, b_2).$$

The symbols $\mathcal{N}$ and $\mathcal{G}$ denote the normal and gamma distributions, respectively, and the parameters β^0, V_1^{-1}, δ^0, V_2^{-1}, a_1, b_1, a_2 and b_2, which determine the prior distribution, are defined on the basis of the information available when the analysis begins. Thus, the eliciting process plays an important role, by modelling the available empirical or historical evidence by means of the prior distribution (Chaloner and Duncan, 1983, Chaloner, 1995, Chaloner and Rhame, 2001, Freedman and Spiegelhalter, 1983, Kadane, 1980, Kadane and Wolfson, 1995, Kadane and Wolfson, 1998, Winkler, 1967, Wolpert, 1989).

The joint posterior distribution of the parameters $(\beta, \delta, \tau_1, \tau_2)$, given the data (E, C), can be calculated from equations (2.3-2.5), using Bayes' theorem.

$$\pi(\beta, \tau_1|E) \propto \tau_1^{\frac{N+2a_1}{2}-1} \exp\left\{-\tfrac{1}{2}\left[\tau_1(E - X\beta)'(E - X\beta) + (\beta - \beta^0)'V_1^{-1}(\beta - \beta^0) + 2b_1\tau_1\right]\right\}, \quad (2.6)$$

$$\pi(\delta, \tau_2|C) \propto \tau_2^{\frac{N+2a_2}{2}-1} \exp\left\{-\tfrac{1}{2}\left[\tau_2(C - X\delta)'(C - X\delta) + (\delta - \delta^0)'V_2^{-1}(\delta - \delta^0) + 2b_2\tau_2\right]\right\}. \quad (2.7)$$

Inferences about quantities of interest must be based on these posterior distributions. Unfortunately, these are not straightforward, thus the Gibbs sampling algorithm, in the context of the Markov Chain Monte Carlo (MCMC) simulation seems to be the most appropriate (Gelman *et al.*, 1995, Geman and Geman, 1984, Gilks *et al.*, 1996, Tweedie, 1998).

The treatment received is defined by means of a dichotomous variable (X_T) that is assigned a value of 0 for the control treatment and a value of 1 when the treatment received is a new treatment. The parameters corresponding to the latter variable are simple to interpret. The coefficient of the treatment variable in the effectiveness regression model (β_T) is interpreted as the mean increment in effectiveness derived from the new treatment in comparison with the control treatment. To obtain the cost increment corresponding to the new treatment, it is only necessary to estimate the coefficient δ_T.

The posterior cost-effectiveness acceptability curve describes the probability of the net benefit presenting positive values, that is, the posterior probability of the new treatmente being preferred to the control treatment, for each of the R_c considered:

$$Q(R_c) = \Pr(INB(R_c) > 0|E, C).$$

2.2 Binary effectiveness

On many occasions, the effectiveness data are not determined by a quantitative variable. An example of this is binary effectiveness, which is measured from a dichotomous variable $\{0, 1\}$ depending on whether or not a certain positive event has occurred.

Let us assume N binary random independent variables and that $Y_i, \ldots, Y_N$ are observed, where Y_i follows a Bernoulli distribution with a probability p_i of the event occurring. This probability p_i depends on a series of covariates that may be continuous or discrete. Let us define a binary regression model in a general way as $p_i = \mathcal{H}(X_i'\beta)$, $i = 1, \ldots, N$, where β is a vector of unknown parameters with dimension $(k + 1) \times 1$, and $X_i = (1, X_{1,i}, X_{2,i}, \ldots, X_{k,i})'$ is the vector of the known covariates. The logit model is obtained when we assume that $\mathcal{H}$ is the logistic distribution. For a classical description of binary models, see Cox (1971), Nelder and McCullagh (1989), Maddala (1983) and McFadden (1974).

We now present the application of the logit model to cost-effectiveness studies. We describe the model corresponding to effectiveness, the cost model being identical to that analysed in Section 2.1.

We examined a sample of N individuals who took part in a clinical trial involving two alternative treatments, in which the effectiveness (E_i) of each was known, $i = 1, \dots, N$.

$$E_i \sim Be(p_i), \tag{2.8}$$

where

$$p_i = \frac{e^{X_i' \cdot \beta}}{1 + e^{X_i' \cdot \beta}}.$$

The first step in the Bayesian analysis requires us to consider a likelihood function for the data, which in this case is the effectiveness. We apply the logit model, and so the likelihood function is specified as follows:

$$\ell_e(E|\beta) = \prod_{i=1}^{N} p_i^{E_i}(1 - p_i)^{1-E_i} = \prod_{i=1}^{N} \left(\frac{\exp[X_i' \cdot \beta]}{1 + \exp[X_i' \cdot \beta]} \right)^{E_i} \left(1 - \frac{\exp[X_i' \cdot \beta]}{1 + \exp[X_i' \cdot \beta]} \right)^{1-E_i}. \tag{2.9}$$

Having defined the likelihood function, we now propose a flexible model for the prior distribution. The normal multivariate distribution for the β parameters is flexible enough to include a large number of possible prior situations,

$$\pi(\beta) \sim \mathcal{N}(\beta^0, V_1^{-1}). \tag{2.10}$$

Estimation of the above binary response model was carried out using Gibbs sampling (Carlin and Polson, 1992, Albert and Chib, 1993).

We propose the use of the difference in the probability of success between treatments (Δp) as the measure to analyse the effectiveness. In a logit model the effect of a covariate on the probability of success depends on the level of the independents. Under the assumption that the sample is representative of the population, we can estimate the difference in probabilities of success between control and new treatment for each patient. The mean incremental effectiveness is estimated as the mean of the increase in the probability of success for the sample. The INB can be calculated as in the previous section where the value R_c is interpreted as the cost that decision-makers are willing to accept in order to increase the probability of success in 1%.

2.3 Skewed cost data: the log-normal model

The cost data obtained from the data of individual patients in health-care economic studies present, for the most part, a strongly asymmetrical distribution. Another characteristic of many cost-effectiveness studies is the small sample size employed. These circumstances frequently oblige us to reject the normality assumption described in Section 2.1.

We now describe a model that reflects this skewed cost, using a non-normal likelihood function. In this sense, Al and Van Hout (2000) described a Bayesian approach to cost-effectiveness analysis showing how costs can be modelled under the assumption of a log-normal distribution. Such a distribution is a much more appropriate way of reflecting possible cost asymmetries.

It is now necessary to reformulate the cost model using a log-normal likelihood function, by which the cost model described in Section 2.1 is expressed as follows:

$$\log(C_i) = \delta_0 + \delta_1 \cdot X_{1,i} + \delta_2 \cdot X_{2,i} + \ldots + \delta_{k-1} \cdot X_{k-1,i} + \delta_T \cdot X_{T,i} + v_i, \tag{2.11}$$

where the vector $\delta = (\delta_0, \delta_1, \delta_2, \ldots, \delta_{k-1}, \delta_T)'$ and $\tau_2 = 1/\sigma_2^2$ are the parameters to be estimated.

The likelihood function of the logarithm of the costs $\ell_c(\log(C)|\delta, \tau_2)$ is:

$$\ell_c(\log(C)|\delta, \tau_2) \propto \tau_2^{\frac{N}{2}} \exp\left\{-\frac{\tau_2}{2}(\log(C) - X\delta)'(\log(C) - X\delta)\right\}.$$

A conditional-conjugate prior distribution is thus the normal-gamma distribution defined above:

$$\pi(\delta, \tau_2) = \pi_{c,1}(\delta) \cdot \pi_{c,2}(\tau_2), \tag{2.12}$$

where

$$\pi_{c,1}(\delta) \sim \mathcal{N}(\delta^0, V_2^{-1}) \qquad \text{and} \qquad \pi_{c,2}(\tau_2) \sim \mathcal{G}(a_2, b_2).$$

Under the assumption of lognormality, the parameter δ_T cannot be interpreted as the incremental cost and it is necessary to search for another means of comparing the two treatment options. In this case the ratio of the costs of the new treatment and those of the control treatment can be described by a simple expression, one that does not depend on the patients' individual characteristics,

$$\frac{C_i^1}{C_i^0} = \exp(\delta_T) \tag{2.13}$$

where C_i^1 is the cost of a patient i who has received the new treatment, and C_i^0 is the cost of the same patient i when the control treatment is applied.

Therefore, values greater than 1 for $\exp(\delta_T)$ indicate that the new treatment is more costly than the control treatment. Thus, $(\exp(\delta_T) - 1) \cdot 100\%$ shows the percentage increase in costs arising from the new treatment.

In comparison with the model described in Section 2.1, the INB presents the following expression:

$$INB = (R_c) \cdot \beta_T - (\exp(\delta_T) - 1), \tag{2.14}$$

where R_c is interpreted as the proportion of the cost increase that the decision-maker is willing to accept in order to increase effectiveness by one unit. Positive INB values show a preference for the alternative treatment. As in the previous sections, we can construct a posterior cost-effectiveness acceptability curve for each value of R_c.

3 Practical application

The data used in this section were obtained from a real clinical trial in which a comparison was made of two highly active antiretroviral treatment protocols applied to asymptomatic HIV patients (COSTVIR study, Pinto *et al.*, 2000).

We obtained data on the direct costs (of drugs, medical visits and diagnostic tests), on the effectiveness, based on clinical variables (percentage of patients with no detectable virus load) and on health-related life-quality variables, using EuroQol-5D.

EuroQol–5D is an instrument for the self-evaluation of personal health, consisting of five questions that investigate five aspects of health-related life quality, based on a visual analogue scale (VAS) (Brooks, 1996).

In this exercise we compared two three-way treatment protocols. The first of these (d4T + 3TC + IND) combines the drugs estavudine (d4T), lamivudine (3TC) and indinavir (IND); the second treatment protocol (d4T + ddl + IND) combines estavudine (d4T), didanosine (ddl) and indinavir (IND).

Two alternative measures of effectiveness were employed. The first of these was the improvement in the patient's life quality, measured as the improvement on a visual analogue scale (VAS). This scale simulates a thermometer with a minimum of 0 and a maximum of 100. The 0 represents the worst health state imaginable, and the 100, the best.

The second effectiveness measure considered was the percentage of patients who, at the end of the treatment programme, presented undetectable levels of viral load. The effectiveness, therefore, can only be expressed as one of two values, either 1 if the viral load is undetectable, otherwise 0.

Table 1 summarises the statistical data obtained. The d4T + ddl + IND treatment is more costly than the d4T + 3TC + IND treatment, by an average of 164.82 euros. When the VAS variation is used as the measure of effectiveness, the d4T + ddl + IND treatment is more effective because, on average, the patients who received this treatment experienced an improvement in their life quality of 4.94 units, while those who were given the d4T + 3TC + IND treatment only experienced a VAS improvement of 4.56 units. However, if the percentage of patients experiencing a reduction of the viral load to undetectable levels is used as the measure of effectiveness, then a better result is obtained for the d4T + 3TC + IND group (68%) than for those who received the alternative treatment (66%).

Table 1: *Statistical summary of costs (in euros) and effectiveness (change in VAS and percentage of patients with undetectable viral load).*

	d4T + 3TC + IND			d4T + ddl + IND		
Statistical measure	Cost	Change in VAS	% with undetectable VL	Cost	Change in VAS	% with undetectable VL
Mean	7142.44	4.56	0.68	7307.26	4.94	0.66
Stan. Devn.	1573.98	15.17	0.47	1720.96	13.98	0.48
N	N_0 = 268			N_1 = 93		

3.1 Assumption of normality in effectiveness and in costs

In this section, the increase in the VAS is used as the measure of the effectiveness of each treatment protocol. For this purpose, we applied the model described in Section 2.1, taking into account the effectiveness and cost of the treatment given to each patient, the individual characteristics of each patient and his/her clinical situation at the moment of the clinical trial.

The model's explanatory variables are the *age*, the *gender* (value 0 if the patient is male and value 1 for a female) and the existence of any concomitant illness (*cc1* with a value of 1 if a concomitant illness is present, otherwise 0; and *cc2* with a value of 1 if two or more concomitant illnesses are present, otherwise 0). The concomitant illnesses considered were hypertension, cardiovascular disease, allergies, asthma, diabetes, gastrointestinal disorders, urinary dysfunction, previous kidney pathology, high levels of cholesterol and/or triglycerides, chronic skin complaints and depression/anxiety. Also included in the model was the time (in months) elapsed since the *start* of the illness until the moment the clinical trial was performed. Finally, we included a dichotomous variable (*trat*) that was assigned a value of 1 if the patient received the (d4T + ddl + IND) treatment protocol and a value of 0 if the (d4T + 3TC + IND) treatment was applied. The linear model of the effectiveness and the costs, for the i-th patient is

$$E_i = \beta_0 + \beta_1 \cdot age_i + \beta_2 \cdot gender_i + \beta_3 \cdot cc1_i + \beta_4 cc2_i + \beta_5 \cdot start_i + \beta_T \cdot trat_i + u_i, \quad (3.1)$$

$$C_i = \delta_0 + \delta_1 \cdot age_i + \delta_2 \cdot gender_i + \delta_3 \cdot cc1_i + \delta_4 cc2_i + \delta_5 \cdot start_i + \delta_T \cdot trat_i + v_i. \quad (3.2)$$

3.1.1 Priors

For a fully Bayesian analysis, we must specify priors for the parameters of interest. The COSTVIR study was carried out in 1999 and it is not practical now to try to elicit the prior information. For the purpose of our illustrative analysis, we look at the reasoning behind the design of the study as an indication of what prior information we can use. For HAART regimens, there were no indications of differences in effectiveness because of age or gender. However, they showed better results for patients with concomitant illnesses and for patients in the early stages of the illness. The d4T + ddl + IND treatment was expected to be on average more effective than the d4T + 3TC + IND treatment but with a prior interval of probability large enough to include negative values.

In cost terms, it was expected that the age, the fact to be female and the months of illness increase cost of HAART therapies. No effect of the existence of concomitant illnesses in cost was expected. Higher cost was expected for the treatment d4T + ddl + IND.

Mean and interval of probability were asked to the experts in an elicitation process to obtain the prior mean and variance of the parameters of interest. Diffuse information is assumed for the precision terms. Then, the prior elicitation is implemented by using the following parameter assignments:

$$\beta^0 = (0, 0, 0, 5, 10, -0.5, 2), \qquad V_1 = \text{diag}(10^{10}, 1, 1, 6.25, 6.25, 0.01, 2.25),$$

$$\delta^0 = (0, 10, 200, 0, 0, 5, 200), \qquad V_2 = \text{diag}(10^{10}, 25, 2500, 625, 625, 6.25, 2500),$$

$$a_1 = 0.5, \quad b_1 = 0, \quad a_2 = 0.5, \text{ and } b_2 = 0.$$

3.1.2 Results

For all models, simulations were done using WinBUGS (Spiegelhalter *et al.*, 1999). A total of 50000 iterations were carried out, after a burn-in period of 10000 iterations. The codes are available from authors upon request. Table 2 shows the posterior estimation of the parameters.

***Table 2**: Posterior statistics and symmetrical interval of probability at 95% (normal model).*

	Mean	Standard deviation	95% CI
β_0	0.9514	3.9213	(-6.6991, 8.6842)
β_1	0.05458	0.1072	(-0.1549, 0.2629)
β_2	-0.3023	0.8701	(-2.0084, 1.3882)
β_3	3.5431	1.4382	(0.7186, 6.3673)
β_4	9.5387	1.7963	(6.0183, 13.0518)
β_5	-0.005698	0.008184	(-0.02176, 0.01038)
β_T	1.4080	1.1471	(-0.8494, 3.6707)
δ_0	6673.4	194.9	(6287.3, 7052.7)
δ_1	9.1532	4.6151	(0.0483, 18.2076)
δ_2	199.31	48.36	(103.84, 293.35)
δ_3	2.4683	24.7677	(-46.2720, 50.9167)
δ_4	-1.0110	24.9816	(-49.7648, 48.3937)
δ_5	1.0614	0.8412	(-0.5928, 2.7221)
δ_T	198.80	48.56	(103.39, 293.91)

Let us begin by analysing the effectiveness model. The age and gender coefficients (β_1 and β_2) are not statistically relevant, which means that these covariates do not affect the final results for effectiveness. The existence of concomitant illnesses favours an increase in the patient's VAS, as shown by the positive signs of the corresponding coefficients. The months elapsed between the start of the illness and the moment of the clinical trial do not seem to affect the final effectiveness results.

The β_T coefficient indicates the incremental effectiveness of the new treatment. The coefficient has a value of 1.4080, which indicates that the patients who received the three-way treatment (d4T + ddl + IND), under conditions of *ceteris paribus*, reported an increase in their health state evaluation an average of 1.4080 units greater than the patients who were given the alternative treatment. Nevertheless, the 95% probability interval includes both positive and negative values, and so we cannot claim that the difference between the two treatment protocols, with regard to effectiveness, is statistically relevant. From the posterior marginal distribution of the β_T coefficient, it can be said that there exists a probability of 88.8% that the (d4T + ddl + IND) treatment is more effective than the (d4T + 3TC + IND) treatment.

With regard to costs, we found that the (d4T + ddl + IND) treatment is more expensive than the alternative, by an average of 198.80 euros, with an interval of probability of (103.39, 293.91).

The incremental cost-effectiveness ratio is calculated as the ratio of the increases in cost and effectiveness (δ_T/β_T). In the present study, the ICER was found to be 290.21. Figure 1 shows the joint posterior distribution of the incremental costs and effectiveness measured.

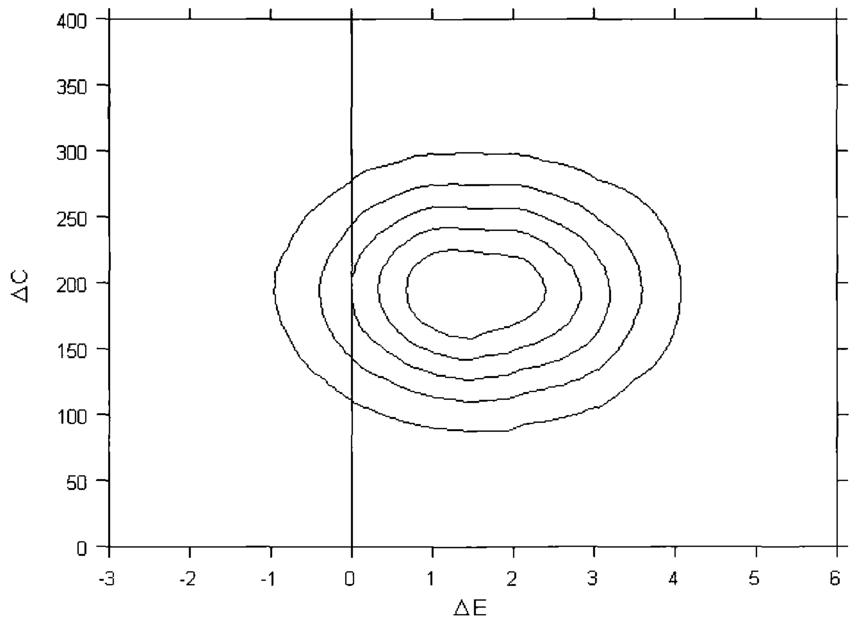

Figure 1: *Joint posterior distribution of costs and incremental effectiveness (normal model).*

In addition to the ICER, we obtained the value of the incremental net benefit (INB). Figure 2 shows the probability that the INB is positive for every possible value of R_c, that is, the cost-effectiveness acceptability curve.

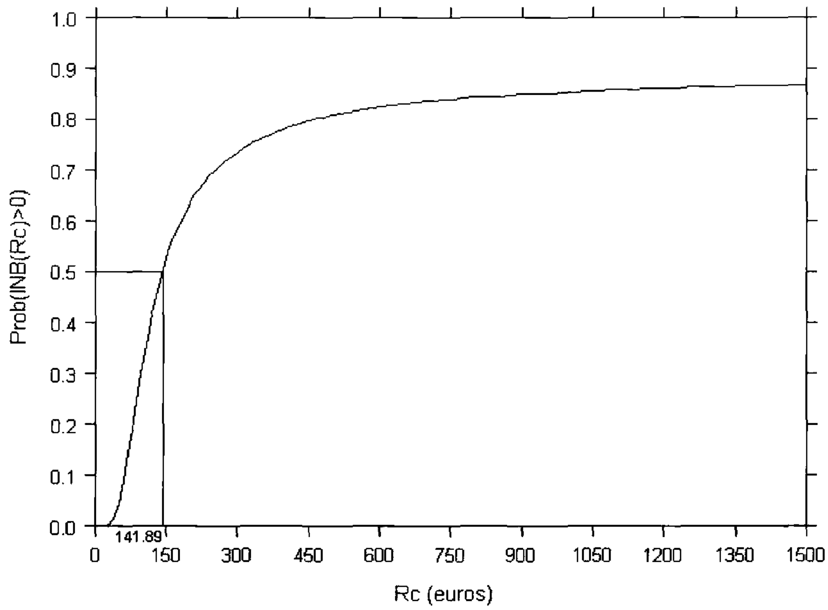

Figure 2: *Cost-effectiveness acceptability curve (normal model).*

At a willingness to pay of 141.89 euros or more, the decision-maker prefers the alternative treatment (d4T + 3TC + IND), because the probability of this preference is greater than 50%.

3.2 Binary effectiveness

We now consider the possibility of the effectiveness being measured by means of a binary variable, that is, the percentage of patients who, given a certain treatment option, achieve undetectable levels of viral load.

Table 1 shows that 68% of the patients achieved undetectable levels of viral load with the (d4T + 3TC + IND) treatment, versus 66% of those given the (d4T + ddl + IND) treatment. We now apply the logit regression model described in Section 2.2. This model enables us to determine whether the differences between the two treatment groups are due to the treatment itself or to individual characteristics of the patients.

The odds ratio (OR) is the most common measurement used to compare the probability of success between two categories of a qualitative variable in a logit model

(Deeks, 1998). Its main advantage over alternative measurements comparing treatments is its ability to measure independently of individual patient characteristics. Thus, when two categories 1 and 0, of a dichotomous variables are compared, indicating here the type of treatment received, the odds ratio is obtained as the relative probability of the success ratio between categories. Thus the final value obtained does not depend on the remaining individual patient characteristics:

$$\mathrm{OR} = \frac{\dfrac{p_i^1}{1 - p_i^1}}{\dfrac{p_i^0}{1 - p_i^0}} = \exp(\beta_T), \tag{3.3}$$

where p_i^1 is the probability of success of a patient i who has received the new treatment, and p_i^0 is the probability of success of the same patient i who has received the control.

Values greater than 1 for the odds ratio reflect a preference for the new treatment, as the relative probability of improvement is greater than in the case of the control treatment. The odds ratio has a very intuitive practical consideration, and the decision-maker who has a good statistical training should have no problem to assess it. We propose to use this feature in the elicitation process as shown in the following.

3.2.1 Priors

We include prior information about the value of the coefficients of the logit model. However, the coefficients have not a natural interpretation to be elicited. For that reason we asked the experts the prior beliefs about the mean and variance of the odds ratio for each covariate.

Assuming that the prior distribution of the vector of coefficients β is normal, the prior distribution of the odds ratio is log-normal. Thus, we can elicit the prior mean and variance using the following relationship:

$$\beta_k \sim \mathcal{N}(\beta_k^0, V_{1_{k,k}}^{-1}) \qquad \Longleftrightarrow \qquad \mathrm{OR}_k = \exp(\beta_k) \sim \text{log-}\mathcal{N}(\mathrm{OR}_k^0, V_{\mathrm{OR}_{k,k}}^{-1}),$$

where $\log -\mathcal{N}$ denotes the log-normal distribution and the two first moments are:

$$\mathbb{E}[\mathrm{OR}_k] = \mathrm{OR}_k^0 = \exp(\beta_k^0 + V_{1_{k,k}}^{-1}/2),$$

$$\mathrm{Var}[\mathrm{OR}_k] = V_{\mathrm{OR}_{k,k}}^{-1} = \exp(2 \cdot \beta_k^0 + V_{1_{k,k}}^{-1}) \cdot (\exp(V_{1_{k,k}}^{-1}) - 1).$$

The experts have prior information about the mean and variance of odds ratios. Solving the previous system of equations we can obtain prior information about the coefficients β.

Before the study was carried out, the experts expected lower probabilities to achieve undetectable viral load for women (odds ratio of 0.8), patients with concomitant illnesses (odds ratios of 0.7 and 0.5 for *cc1* and *cc2*) and for each additional month of illness (odds ratio of 0.8). It is necessary to comment on the different signs of the coefficient for concomitant illnesses for the two measures of effectiveness considered. The HAART regimens improve the quality of life of the patients with concomitant illnesses attenuating the effect of these illnesses. However, the existence of these concomitant illnesses supposes an inconvenience in the goal of achieving undetectable viral load. There was no prior information about the difference between treatments. A small value of 0.01 was assigned to the prior variance for all the odds ratios. Then, the prior elicitation is implemented by using the following parameter assignments:

$$\beta^0 = (0, 0, -0.2301, -0.3667, -0.7124, -0.2301, 0),$$

and

$$V_1 = \text{diag}(10^{10}, 10^{10}, 0.0154, 0.02, 0.0385, 0.0154, 10^{10}).$$

3.2.2 Results

Table 3 shows some posterior moments of the parameters for the effectiveness regression estimated by means of MCMC simulation techniques.

Table 3: *Posterior statistics and symmetrical interval of probability at 95% (binary effectiveness).*

	Mean	Standard deviation	95% CI
β_0	1.5281	0.6020	(0.3998, 2.7534)
β_1	-0.0127	0.0162	(-0.0454, 0.0181)
β_2	-0.3174	0.1109	(-0.5359, -0.0972)
β_3	-0.3728	0.1244	(-0.6102, -0.1276)
β_4	-0.7451	0.1709	(-1.0760, -0.4078)
β_5	-0.000402	0.001304	(-0.002807, 0.002327)
β_T	-0.0367	0.2589	(-0.5392, 0.4722)
$\exp(\beta_T)$	0.9968	0.2631	(0.5832, 1.6041)
Δp	-0.002285	0.014085	(-0.030341, 0.024540)

The relative risk measure is usually employed to compare categories in logit discrete choice models. This measure is obtained by determining the ratio of the relative probabilities of success and failure of two categories. With regard to the type of treatment received, a patient given the (d4T + ddl + IND) treatment has an odds ratio of reducing the viral load to undetectable levels of 99.7% with respect to another, with the same characteristics, who receives the (d4T + 3TC + IND) treatment. There is a probability of 44.7% that the new treatment (d4T + ddl + IND) is more effective than the first-named one (d4T + 3TC + IND). The regression coefficients corresponding to the costs are the same as in the previous section.

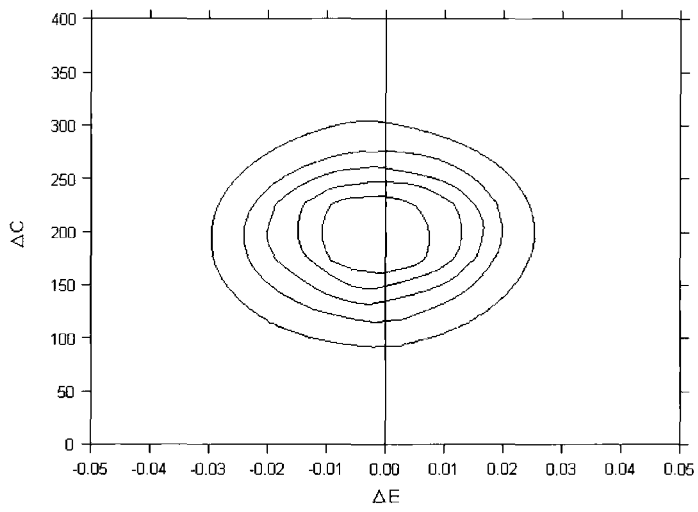

Figure 3: *Joint posterior distribution of costs and relative risk (binary effectiveness).*

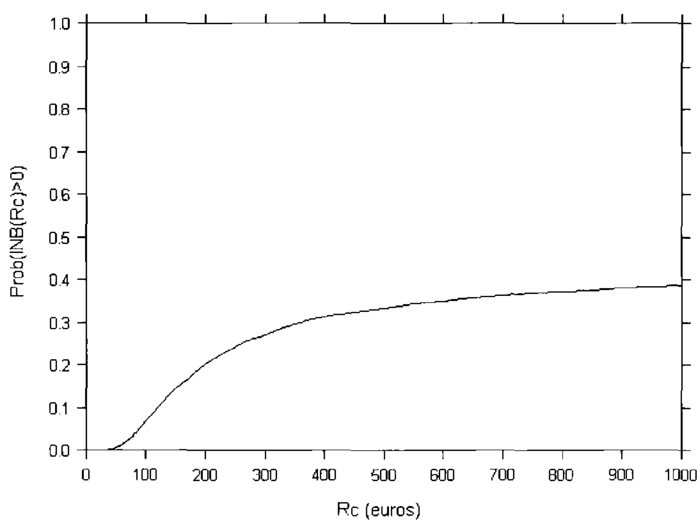

Figure 4: *Cost-effectiveness acceptability curve (binary effectiveness).*

Besides the odds ratio, we estimate the mean difference in the probability of success between treatments. The mean incremental change in probability is estimated as -0.229%, with a Bayesian interval of $(-3.03\%, 2.45\%)$.

Figure 3 shows the joint posterior distribution of the increase in probability and of the incremental cost.

The cost-effectiveness acceptability curve is shown in Figure 4. From the cost-effectiveness acceptability curve, we see that the new treatment (d4T + ddl + IND) is not preferred, in all cases, to the control treatment (d4T + 3TC + IND).

3.3 Cost asymmetry: log-normal model

Most statistical models assume normality in effectiveness and in costs (O'Hagan *et al.*, 2001, O'Hagan and Stevens, 2002). In practice, however, costs tend to present severe asymmetry, and this should be taken into account in the analysis. Evidence of skewing is shown in Figure 5, which contains a histogram of the residuals from the normal model of Section 3.1. Due to this skewness, it is more appropriate to consider a log-transformation. The analysis of the effectiveness is similar to that of Section 3.1.

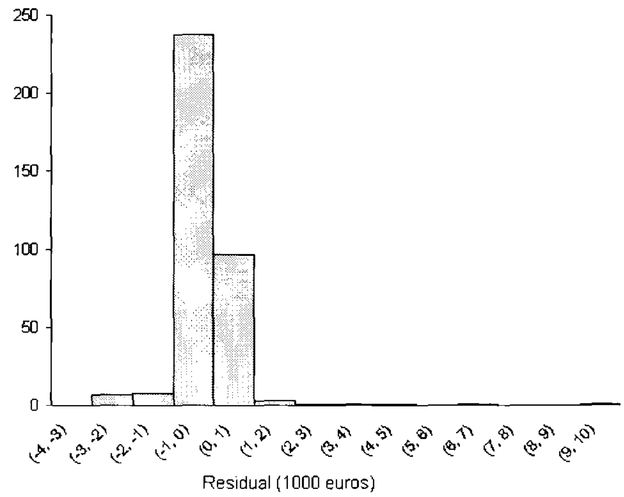

***Figure 5**: Histogram of residuals of the normal model.*

The coefficients of the log-normal model does not have a natural interpretation and it is necessary to search for another means of comparing the effect of a covariate. In this case, the ratio of the costs of having or not a characteristic can be described by the exponential of the coefficient. We use this property to define our prior information:

$$\exp(\delta_k) = \frac{C(X_k = 1)}{C(X_k = 0)} = \frac{C(X_k = 1) - C(X_k = 0)}{C(X_k = 0)} + 1 = \frac{\Delta C}{C(X_k = 0)} + 1,$$

where $C(X_k = 1)$ is the cost of a patient in the treatment group and $C(X_k = 0)$ is the cost of a patient in the reference group.

3.3.1 Priors

We can elicit the prior mean and variance of the exponential of each coefficient β using the prior information shown in Section 3.1.

$$\mathbb{E}[\exp(\delta_k)] = \frac{E(\Delta C)}{C(X_k = 0)} + 1 \qquad \text{and} \qquad \text{Var}[\exp(\delta_k)] = \frac{\text{Var}(\Delta C)}{(C(X_k = 0))^2}.$$

Using as $C(X_k = 0)$ the sample mean of the cost for the reference group we obtain the prior mean and variance for the exponential of the coefficients. For continuous covariates as *age* or *start* we use as reference group the total of the sample. With this information and similarly to the previous section we can obtain the prior information about the coefficients:

$$\beta^0 = (0, 1.39060 \cdot 10^{-3}, 2.76073 \cdot 10^{-2}, -6.12301 \cdot 10^{-6}, -6.02103 \cdot 10^{-6},$$
$$\cdot 6.93488 \cdot 10^{-3}, 2.75936 \cdot 10^{-2}),$$

and

$$V_1 = \text{diag}(10^{10}, 4.80952 \cdot 10^{-7}, 4.64127 \cdot 10^{-5}, 1.22460 \cdot 10^{-5}, 1.20421 \cdot 10^{-5},$$
$$\cdot 1.19405 \cdot 10^{-7}, 4.63492 \cdot 10^{-5}).$$

3.3.2 Results

The new treatment is 2.78% more expensive than the control one, with an interval of probability of 95% of (1.48%, 4.06%).

***Table 4**: Posterior statistics and symmetrical interval of probability at 95% (log-normal model).*

	Mean	Standard deviation	95% CI
δ_0	8.785	0.02294	(8.74, 8.829)
δ_1	0.000961	0.000585	(-0.000190, 0.002107)
δ_2	0.02548	0.006382	(0.01285, 0.03784)
δ_3	0.000862	0.003448	(-0.005887, 0.007592)
δ_4	-0.000138	0.003429	(-0.00685, 0.006563)
δ_5	0.000423	0.0000858	(0.000259, 0.000596)
δ_T	0.027411	0.006416	(0.0147, 0.03977)
$\exp(\delta_T) - 1$	0.02781	0.006594	(0.01481, 0.04057)

Figure 6 shows the joint posterior distribution of the incremental effectiveness and the relative incremental cost ($\exp(\delta_T) - 1$).

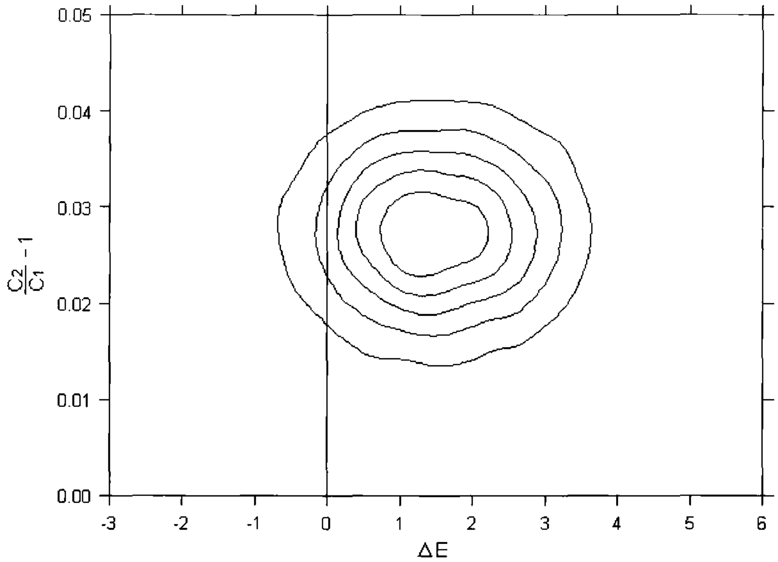

Figure 6: *Joint posterior distribution of incremental effectiveness and of the ratio between costs (log-normal model).*

The cost-effectiveness acceptability curve is shown in Fig. 7.

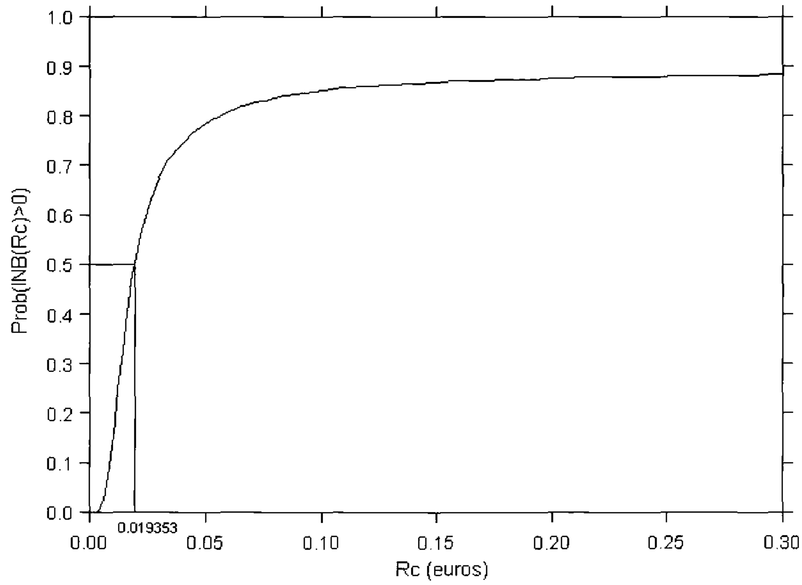

Figure 7: *Cost-effectiveness acceptability curve (log-normal model).*

The critical value is 0.019353. When the decision-maker is prepared to increase costs by 1.9353% or more in order to increase effectiveness by one unit, then the new treatment (d4T + ddl + IND) will be preferred. If we take the cost of the control treatment as its mean value (7142.44 euros), an increase of 1.9353% is equivalent to 138.23 euros. In Section 3.1, with the assumption of normality, this critical value was calculated to be 141.89 euros. The greater the degree of asymmetry in costs, the greater is the divergence between the results obtained by the normality assumption and the log-normal assumption.

4 Conclusions

This paper presents a flexible methodology to carry out cost-effectiveness analysis, developed from a Bayesian perspective. The assumption common to all models is that the effectiveness and cost differences between alternative treatment options may not be due solely to the type of treatment received. Sample differences between the groups given one or other of the two treatments may be relevant and influence the final results for effectiveness and cost. Therefore, a valid comparison of two alternative treatments is only possible if we are able to isolate the effect of the type of treatment received on the variables of interest (effectiveness and cost). In order to achieve this, we must create a regression model that includes the other explanatory variables and a dichotomous variable that is assigned a value of 0 or 1 depending on the type of treatment received. On the basis of these models, we can generate the different cost-effectiveness decision-making measures described in the literature.

The initial model is normal-normal, in which both effectiveness and costs are assumed to follow a normal distribution. This assumption may be justified by the central limit theorem, in the case of large sample sizes.

However, on some occasions the effectiveness measure is not determined by a quantitative variable. For example, effectiveness may be measured by whether or not a certain objective has been achieved. Taking this into account, we have developed an alternative model that uses the difference in the probability of success as measure of effectiveness.

Moreover, costs often present severely asymmetrical distributions, or the sample size may be limited, which would invalidate the assumption of normality. In such cases, it is necessary to assume an alternative cost distribution, one that is flexible to the existence of extreme values. Such a requirement is met by the log-normal distribution, and the ratio of costs is then used to compare different treatments.

All the models described here have been developed from a Bayesian perspective, which enables us to incorporate prior information (if it exists) in a natural, flexible way, and to interpret the results in terms of probability. For the purposes of our illustrative analysis, we obtained prior information from the consensus of the experts who participated in the study. A different elicitation process is proposed for each model, and this process plays an important role in the analysis of the results. For future research more efforts have to be carried out to elicit the prior information and to analyse the robustness of the models. The cost-effectiveness acceptability curve is shown to be a natural measure and one that is easy for the decision-maker to interpret.

Acknowledgments

The authors are very grateful to an executive editor and two anonymous referees for their constructive and helpful comments, which have significantly improved the paper. This research is partially funded by grants Beca Bayer (Asociación de Economía de la Salud and Química Farmacéutica Bayer, S.A.) and MCyT (Ministerio de Ciencia y Tecnología, Spain, Project BEC2001/3774). We would like to thank Professor Badía (Health Outcomes Research Europe, Spain) for allowing us to use the COSTVIR data archive. Any responsability for the further analysis of the data is ours.

References

Al, M.J. and Van Hout, B.A. (2000). A Bayesian approach to economic analysis of clinical trials: the case of Stenting versus Balloon Angioplasty, *Health Economics*, 9, 599-609.

Albert, C.H. and Chib, S. (1993). Bayesian analysis of binary and polychotomus response data, *Journal of the American Statistical Association*, 88, 669-679.

Berry, D.A. (1996). *Statistics: a Bayesian Perspective*. Belmont, CA: Duxbury Press.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

Briggs, A.H. (1999). A Bayesian approach to stochastic cost-effectiveness analysis, *Health Economics*, 8, 257-261.

Briggs, A.H. (2001). A Bayesian approach to stochastic cost-effectiveness analysis. An illustration and application to blood pressure control in type 2 diabetes, *International Journal of Technology Assessment in Health Care*, 17, 69-82.

Brooks R., with the EuroQol Group. (1996). EuroQol the current state of play, *Health Policy*, 37, 53-72.

Carlin B.P. and Polson N.G. (1992). Monte Carlo Bayesian methods for discrete regression models and categorical time series, *Bayesian Statistics*, 4, 577-586.

Chaloner, K. and Duncan, G.T. (1983). Assessment of a Beta prior distribution: PM Elicitation, *Journal of the Royal Statistical Society, D (The Statistician)*, 32, 174-180.

Chaloner, K. (1995). The elicitation of prior distributions. In Berry, D. and Stangl, D. (eds). *Case Studies in Bayesian Biostatistics*. Marcel Dekker, New York, 141-156.

Chaloner, K. and Rhame, F.S. (2001). Quantifying and documenting prior beliefs in clinical trials, *Statistics in Medicine*, 20, 581-600.

Cox, D.R. (1971). *The Analysis of Binary Data*. London: Mehtuen.

Deeks, J. (1998). Odds ratios should be used only in case-control studies and logistic regression analysis, *Bristish Medical Journal*, 317, 1155.

Freedman, L.S. and Spiegelhalter, D.J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials, *Journal of the Royal Statistical Society, D (The Statistician)*, 33, 153-160.

Gelman, A., Carlin, J., Stern, H.S. Rubin, D.B. *et al.* (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Geman, R., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Heitjan, D.F. (1997). Bayesian interim analysis of phase II cancer clinical trials, *Statistics in Medicine*, 16, 1791-1802.

Heitjan, D.F., Moskowitz, A.J. and William, W. (1999). Bayesian estimation of cost-effectiveness ratios from clinical trials, *Health Economics*, 8, 191-201.

Hoch, J.S., Briggs, A.H. and Willan, A. (2002). Something olds, something new, something borrowed, something BLUE: a framework for the marriage of health econometrics and cost-effectiveness analysis, *Health Economics*, 11, 415-430.

Jones, D.A. (1996). A Bayesian approach to the economic evaluation of health care technologies. In Spiker (eds). *Quality of Life and Pharmaeconomics in Clinical Trials, second edition*. Philadelphia: Lippincott-Raven, 1189-1196.

Kadane, J.B. (1980). Predictive and structural methods for eliciting prior distributions. In Zellner, A. (eds.) *Bayesian Analysis in Econometrics and Statistics*. Amsterdam: North-Holland.

Kadane, J.B. and Wolfson, L.J. (1995). Priors for the design and analysis of clinical trials. In Berry, D. and Stangl, D. (eds.) *Bayesian Biostatistics*. New York: Marcel Dekker, 157-186.

Kadane, J.B. and Wolfson, L.J. (1998). Experiences in elicitation, *Journal of the Royal Statistical Society, D (The Statistician)*, 47, 3-19.

Laska, E.M., Meisner, M. and Siegel, C. (1997). Statistical inference for cost-effectiveness ratios, *Health Economics*, 6, 229-242.

Löthgren, M. and Zethraeus, N. (2000). Definition, interpretation and calculation of cost-effectiveness acceptability curves, *Health Economics*, 9, 623-630.

Maddala, G.S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In Zarembka, P. (eds.) *Frontiers of Econometrics*. New York: Academic Press, 105-142.

Mullahy, J. and Stinnett, A. (1998). Net health benefits: a new framework for the analysis of uncertainty in cost effectiveness analysis, *Medical Decision Making*, 18, S68-S80.

Nelder, J.A. and McCullagh, P. (1989). *Generalized Linear Models*. New York: Chapman & Hall.

O'Hagan, A. and Stevens, J.W. (2001*a*). Framework for cost-effectiveness analysis from clinical trial Data, *Health Economics*, 10, 303-315.

O'Hagan, A. and Stevens, J.W. (2001*b*). Bayesian assessment of sample size for clinical trials of cost-effectiveness, *Medical Decision Making*, 21, 219-230.

O'Hagan, A., Stevens, J.W., and Montmartin, J. (2001). Bayesian cost-effectiveness analysis from clinical trial data, *Statistics in Medicine*, 20, 733-753.

O'Hagan, A. and Stevens, J.W. (2002). Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality?, *Health Economics*, 12, 33-49.

Pinto, J.L., López C., Badía, X., Coma, A. *et al.* (2000). Análisis coste-efectividad del tratamiento antirretroviral de gran actividad en pacientes infectados por el VIH asintomáticos, *Medicina Clínica*, 114, 62-67.

Spiegelhalter, D.J., Feedman, L.S. and Parmar, M.K.B. (1994). Bayesian approaches to randomized trials (with discussion), *Journal of the Royal Statistical Society, Series A*, 157, 357-416.

Spiegelhalter, D.J., Thomas, A. and Best, N. (1999). *WinBUGS, version 1.2.* MRC Biostatistics Unit: Cambridge, U.K.

Stinnett, A.A. and Mullahy, J. (1998). Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis, *Medical Decision Making*, 18, S65-S80.

Tambour, M., Zethraeus, N. and Johannesson, M. (1998). A note on confidence intervals in cost-effectiveness analysis, *International Journal of Technology Assessment in Health Care*, 14, 467-471.

Tweedie, R.L. (1998). *Markov chains: structure and applications.* Technical Report, Department of Statistics, Colorado State University.

Van Hout, B.A., Al, M.J., Gordon, G.S. and Rutten, F.F. (1994). Costs, effects and C/E-ratios alongside a clinical trial, *Health Economics*, 3, 309-319.

Wakker, P. and Klaassen, M.P. (1995). Confidence intervals for cost-effectiveness ratios, *Health Economics*, 4, 373-381.

Willam, A.R. and O'Brien, B.J. (1996). Confidence intervals for cost-effectiveness ratios: an application of Fieller's theorem, *Health Economics*, 5, 297-305.

Willan, A.R., Briggs, A.H., Hoch, J.S. (2004). Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data, *Health Economics*, (in press).

Winkler, R.L. (1967). The assessment of prior distributions in Bayesian analysis, *Journal of the American Statistical Association*, 63, 776-800.

Wolpert, R.L. (1989). Eliciting and combining subjective judgements about uncertainty, *International Journal of Technology Assessment in Health Care*, 5, 537-557.

Resum

La major part de les publicacions que comparen diverses opcions de tractament, es redueixen a comparar els resultats (eficàcia i cost) obtingudes per cada grup. Aquest treball proposa la incorporació d'altres característiques dels pacients en l'anàlisi. La major part dels estudis duts a terme en aquest context suposen que tant el cost com l'eficàcia són normals. A la pràctica les dades no sempre es distribueixen d'acord amb aquesta hipòtesi. Cal desenvolupar models alternatius. En aquest article presentem un model general que incorpora una mesura de l'eficàcia binària i un cost asimètric. En un aplicació pràctica, comparem dos tractaments antiretrovirals altament actius donats a pacients VIH asimptomàtics. Proposem un model logit on l'eficàcia es mesura d'acord amb si s'ha aconseguit un determinat propòsit inicial. Per a aquest model, la mesura per comparar els tractaments es la diferència en la probabilitat d'èxit.

A més, les dades de cost són usualment asimètriques cap a la dreta. Proposem usar la transformació logarítmica per a dur a terme el model de regressió. Els tres models es condueixen demostrant els avantatges d'aquest model. La corba d'acceptabilitat cost-eficàcia s'utilitza com a mesura per prendre les decisions.

MSC: 62F15, 62H12, 62P10

Paraules clau: anàlisis baiesiana, eficàcia-Cost, Markov chain Monte Carlo (MCMC), distribucions de cost asimètriques

Book reviews

STATISTICAL INFERENCE FOR ERGODIC DIFFUSION PROCESSES

Yury A. Kutoyans

Springer-Verlag, London, 2004
481 pages

This book presents an introduction to the large sample theory of statistical inference for one-dimensional homogeneous ergodic diffusion processes. That is, given the model

$$dX_t = S(X_t)dt + \sigma(X_t)dW_t$$

the book considers several problems of estimation of the trend coefficient $S(x)$. The diffusion coefficient σ is supposed to be always known and positive.

Then it is a book on statistical inference for continuous time stochastic processes. This has to be seen as a branch of Mathematical Statistics, and clearly attracts every year more and more attention of researchers, interested in its wide range of applications that covers biomedical sciences, economics, genetic analysis, mechanics, physics and especially financial mathematics.

This book contains many recent results and most of them appear for the first time in a book. The statements of the problems are in the spirit of classical mathematical statistics and special attention is paid to asymptotically efficient procedures.

In practice, in estimating continuous time models we have two types of errors. One due to the statistical nature of continuous time observations and the other is a consequence of approximation of continuous data by discrete data. If discrete observations are taken sufficiently dense, the second error is negligible with respect to the first one. Then this fact justifies this inference theory.

Chapter 1, Diffusion Processes and Statistical Problems is a summary of the required basic concepts of Diffusion Processes, Limit Theorems and Statistical Inference.

Chapter 2, Parameter Estimation is devoted to parametric estimation: maximum likelihood, bayesian, minimum distance, moments,...

Chapter 3, Special Models, studies in detail partial observed systems and cusp, delay and change point estimation.

Chapter 4, Nonparametric estimation, is devoted to distribution function and density estimation, and also semiparametric estimation.

Chapter 5, Hypothesis Testing, is devoted to parametric and nonparametric tests, and includes also goodness of fit test.

The book also contains historical remarks and a good list of references.

Josep Vives
Departament de Matemàtiques
Universitat Autònoma de Barcelona
Spain

PARTIAL IDENTIFICATION ON PROBABILITY DISTRIBUTIONS

Charles F. Manski

Springer Series in Statistics, 2003
178 pages

This book deals with statistical inference based on data and assumptions that only partially identify population parameters. Its origin is the research of the author on partial identification of probability distributions, started in the late 1980s on nonparametric regression analysis with missing outcome data, and continued by investigating more general incomplete data situations.

The chosen approach to statistical inference is nonparametric analysis, that enables us to learn from the available data without imposing additional assumptions on the population distribution (assumptions that are not often well motivated) and to know about the limitations of the data in order to support inferences about the population parameters.

This book complements an earlier book of the same author, entitled *"Identification Problems in the Social Sciences" (Manski, 1995)*, that introduces in an accessible way the principles of partial identification to students and researchers in the social sciences. The present book develops this subject in a more rigorous manner, with the aim of providing the foundations for further studies by statisticians and econometricians.

The background needed to follow the contents of the book is only elementary probability theory, especially the Law of Total Probability and Bayes Theorem. At the end of each chapter there is some complements and endnotes to place it in context and to provide historical perspective. The first endnote of each chapter cites its sources, basically research articles of the author, alone or with co-authors, until 2002.

There is a common structure in every chapter: first of all, the sampling processes are specified which generate the available data; then the question is considered of what can be said of population parameters without imposing restrictive assumptions on the population distribution, by obtaining the set-valued identification region containing the parameters. Finally, it is studied the possibility that these identification regions may shrink if certain assumptions on the population distribution are imposed, such as statistical independence and monotonicity assumptions. The complementary approach which begins with some point-identifying assumption and then examine how identification becomes more partial as the assumption is weakened, is not considered here. This last approach is referred to as *sensitivity, perturbation* or *robustness* analysis, and has been followed by Rosenbaum (1995) and Robins (1999) among others.

Chapter 1, "Missing Outcomes", deals with the problem of identification by using only empirical evidence, when the data are generated by random sampling and some outcome realizations are not observable at all. It is also considered the generalization to cases in which data from multiple sampling processes are available, and where outcomes that are observable under some sampling processes may be missing under others; the objective is then to combine data generated by the sampling processes to learn as much as possible about the population distribution. Sometimes the real situation for empirical researchers is the intermediate one, corresponding to the partial knowledge that the realization belongs to a set-valued identification region. This case is studied at the end of the chapter.

Chapter 2, "Instrumental Variables", treats the use of instrumental variables in the formulation of distributional assumptions that help to identify the distribution of outcomes. Some of such assumptions imply point identification, whereas others have less identifying power and, possibly, more credibility. The supposition that data are missing-at-random (MAR) is one of such assumptions, that is weakened to the mean-missing-at-random assumption (MMAR). Another interesting assumption is the mean independence of outcomes of an instrumental variable (MI).

A large part of statistical practice aims to predict outcomes conditional on covariates. In practice it is common to have missing outcomes and/or covariates. While analysis of chapters 1 and 2 extends immediately to inference on conditional outcome distributions when the conditioning event is always observed, in Chapter 3, entitled "Conditional Prediction with Missing Data", the case of data on outcomes and/or conditioning events missing is considered. Therefore, Chapters 1, 2 and 3 form a unit on prediction with missing outcome and/or covariates.

Chapters 4, "Contaminated Outcomes", and 5, "Regressions, Short and Long", form a unit on decomposition of finite mixtures. Inference on the components of finite probability mixtures has application in distinct areas, as contaminated sampling, ecological inference and regression with missing covariate data, that is the problem that originally motivates the author in his research. In fact, Chapter 4 deals with the mixture model of data errors, that presents the available data as realizations of a probability mixture of an error-free realization and a data-error (that imperfectly measures the variable of interest). On the other hand, Chapter 5 studies the problem of ecological inference, that is well known by the social scientists who aim to predict outcomes conditional on (two) covariates. It uses the terminology *short regression* (respectively, *long regression*) from Goldberger (1991), that corresponds to the conditional expectation of the outcome with respect to one covariate (respectively, with respect both covariates).

The analysis of the response-based sampling, often motivated by practical considerations as cost reduction, is treated in Chapter 6, entitled "Response-based Sampling". The response-based sampling consists on divide the population into some sub-populations or strata, according to the values of the outcome (response) and sample at random within each stratum. It is particularly effective, for instance, in generating observations of serious diseases, as ill persons are clustered in treatment centres.

Chapter 7, "Analysis of Treatment Response", and the next chapters form a unit on the study of the problem of missing outcomes given by the non-observability of *counterfactual outcomes* in empirical analysis of treatment response. In studies of treatment response, treatments are mutually exclusive, so it is not possible to observe the outcomes that an experimental unit would experience under other treatment that its own. This study starts with Chapter 7 and is continued in Chapter 8, "Monotone Treatment Response", that considers the situation in which there exist consistent reasons to believe that outcomes vary monotonically with the intensity of the treatment. Chapter 9, "Monotone Instrumental Variables", studies identification of mean treatment response under distributional assumptions weaker (and more credible) that the assumption of independence between outcomes and instrumental variables. The last chapter is Chapter 10, "The Mixing Problem", and studies prediction of outcomes when treatment, unlike the three previous chapters, may vary within the group of experimental units who share the same value of the covariates. In this sense, it is a extrapolation from classical randomized experiments (that does not point-identify outcome distributions under rules in which treatment may vary within groups).

Rosario Delgado
Departament de Matemàtiques
Universitat Autònoma de Barcelona
Spain

Information for authors and subscribers

Information for authors and subscribers

Submitting articles to SORT

Guidelines for submitting articles

SORT accepts for publication only original articles that have not been submitted to any other journal in the areas of Statistics, Operations Research, Official Statistics and Biometrics. Articles may be either theoretical or applied, and may include computational or educational elements. Publication will be exclusively in English, with an abstract in Catalan (abstract translation into Catalan can be done by the journal's staff).

All articles submitted to thematic sections will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board, except for those articles specifically invited by the journal or reprinted with permission. Reviewer comments will be sent to the primary author if changes are requested in form or content.

When an original article is received, the journal sends the author a dated receipt; this date will appear in the publication as "reception date". The date in which its final version is received will be the "acceptance date".

To submit an article, the author must send it in PDF format to the electronic address of the journal (sort@idescat.es) especifying his e-mail address and/or a phone number. Once the article has been approved, the author must send a final electronic version, following the instructions of the editor responsible for that article. It is recommended that this final version be submitted using a $\LaTeX 2_\varepsilon$.

In any case, upon request the journal secretary will provide authors with $\LaTeX 2_\varepsilon$ templates and appropriate references to the MSC2000 classification of the American Mathematical Society. For more information about the criteria for using this classification system, authors may consult directly the Mathematics Subject Classification 2000 (MSC2000) at the website of the Institut d'Estadística de Catalunya (http://www.idescat.es/sort/Normes.stm).

Publishing rights and authors' opinions

Authored articles represent the opinions of the authors; the journal does not necessarily agree with the opinions expressed in authored articles.

Formal guidelines for articles

The first page of the article must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (100-200 words) followed by MSC2000 classification of the American Mathematical Society and the keywords. Before submitting an article, the author(s) would be well advised to ensure that the text uses correct English.

Bibliographic references have to be listed alphabetically at the end of the article, according to the following examples:

Citations
Mahalanobis (1936), Rao (1982b)

Journal articles
Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics*, 9, 73-84.

Books
Platek, R., Rao, J. N. K., Särndal, C. E. and Singh, M. P. (eds.) (1987). *Small Area Statistics: an International Symposium.* New York: John Wiley and Sons.

Parts of books
Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J.P. Klein and P.K. Goel (eds.), Kluwer Academic Publishers, 237-247.

Web files or "pages"
Nielsen, S. F. (2001). *Proper and improper multiple imputation* http://www.stat.ku.dk/~feodor/publications/ (10th May 2003).

Multiple publications by a single author are to be listed chronologically. Explanatory notes should be numbered sequentially and placed at the bottom of the corresponding page. Tables and figures should also be numbered sequentially and will be reproduced directly from the submitted originals if it is impossible to include them in the electronic text.

SORT *(Statistics and Operations Research Transactions)*

Institut d'Estadística de Catalunya (Idescat)

Via Laietana, 58

08003 Barcelona (Spain)

Tel.: +34-93 412 09 24 or +34-93 412 00 88

Fax: +34-93 412 31 45

E-mail: sort@idescat.es

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form
SORT *(Statistics and Operations Research Transactions)*

Name ____
____
Organisation ____
____
Street Address ____
____
Zip/Postal code ____________________ City ____________________
State/Country ____________________ Tel. ____________________
Fax ______________ NIF/VAT Registration Number ______________
E-mail ____
Date ____________________

Signature

I wish to subscribe to **SORT *(Statistics and Operations Research Transactions)*** for the year 2003 (volume 27)

Annual subscription rates:
- Spain: €22 (VAT included)
- Other countries: €25 (VAT included)

Price for individual issues (current and back issues):
- Spain: €9/issue (VAT included)
- Other countries: €11/issue (VAT included)

Method of payment:

☐ Bank transfer to account number 2013-0100-53-0200698577

☐ Automatic bank withdrawal from the following account number
☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

☐ Check made payable to the Institut d'Estadística de Catalunya

Please send this subscription form (or a photocopy) to:

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

Bank copy

Authorisation for automatic bank withdrawal in payment for
SORT *(Statistics and Operations Research Transactions)*

The undersigned ____

authorises Bank/Financial institution ______________________________

located at (Street Address) ___________________________________

Zip/postal code ____________________ City ____________________

Country ____

to draft the subscription to **SORT *(Statistics and Operations Research Transactions)*** from my account

number ☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

Date ________________________

Signature

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

Institut d'Estadística de Catalunya

The Institut d'Estadística de Catalunya (Idescat) is the statistical office of the Government of Catalonia. Its main duty is the management of the Catalan Statistical System by planning, coordinating and standardizing the statistical activity as well as providing statistical technical assistance.

Secretary and subscriptions to SORT

Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel. +34 - 93 412 09 24 - 93 412 00 88
Fax. +34 - 93 412 31 45
E-mail: sort@idescat.es

Publisher: Institut d'Estadística de Catalunya (Idescat)

ISSN 1696-2281
DL B-46.085-1977
Key title: SORT
Numbering: 1 (december 1977)
http://www.idescat.net /sort

SORT

Volume 28 (1), January - June 2004
Formerly Qüestiió

Contents

Articles

ISSN: 1696-2281
http://www.idescat.net/sort