ESTADÍSTICA ESPAÑOLA
Vol. 39, Núm. 142, 1997, págs. 5 a 21

# Regresión sobre componentes principales de un proceso estocástico con funciones muestrales escalonadas(*)

por
A. M. AGUILERA
Departamento de Estadística e I.O. Universidad de Granada
Tlf: 958 243 718. Fax: 958 243 267. e-mail: aaguiler@goliat.ugr.es
Campus de Fuentenueva
18071.Granada

F. A. OCAÑA
Departamento de Estadística e I.O. Universidad de Granada
Tlf: 958 243 878. Fax: 958 249 046. e-mail: focana@platon.ugr.es
Campus de Cartuja
18071.Granada

M. J. VALDERRAMA
Departamento de Estadística e I.O. Universidad de Granada
Tlf: 958 243 908. Fax: 958 249 046. e-mail: valderra@platon.ugr.es
Campus de Cartuja
18071.Granada

## RESUMEN

El objetivo de este trabajo es desarrollar un procedimiento para estimar observaciones faltantes de un proceso estocástico con funciones muestrales escalonadas a partir de su evolución en el pasado. El

(*) Esta investigación ha sido financiada por el Proyecto PB96-1436 de la Dirección General de Enseñanza Superior del Ministerio de Educación y Cultura, España.

modelo que se propone está basado en regresión lineal múltiple en términos de las componentes principales asociadas al proceso en el pasado. Los factores principales de un proceso de este tipo se aproximan mediante los del proceso cuyas trayectorias se obtienen como proyección de las del proceso original en el subespacio de las funciones constantes sobre los subintervalos de una partición previamente fijada en el pasado. Finalmente, se incluye una aplicación con datos simulados.

*Palabras Clave*: Función de Covarianza, Componentes Principales, Estimación Lineal Mínimo Cuadrática, Proyección Ortogonal, Modelo ECP.

*Clasificación AMS:* 60G12, 65D30, 62H25.

## 1. INTRODUCCIÓN

Este trabajo se enmarca dentro del contexto de Análisis de Datos Funcionales intensamente estudiado en los últimos años (ver, por ejemplo, Ramsay y Silverman (1997)). Su principal objetivo es la estimación lineal de un proceso estocástico en tiempo continuo a partir de su pasado reciente. El modelo que vamos a desarrollar es una extensión de la técnica de regresión en componentes principales al caso en que se quiere estimar una variable en función de las componentes principales asociadas a un número infinito de predictores que serían las variables del proceso en el pasado.

Las componentes principales asociadas a un proceso estocástico de segundo orden son las variables aleatorias de su desarrollo de Karhunen-Loève (Deville (1974)). El problema de estimación de las componentes principales del proceso, a partir de funciones muestrales independientes, fue resuelto por Deville (1973). Más recientemente, Dauxois *et al.* (1982) han extendido la teoría asintótica del ACP de un vector aleatorio con distribución normal al caso de un proceso estocástico gaussiano.

Los factores principales muestrales son las funciones propias de una ecuación integral de núcleo la función de covarianza muestral del proceso. Desafortunadamente, obtener soluciones exactas de ecuaciones de este tipo es una tarea muy difícil, que se complica, aún más, cuando en la práctica disponemos sólo de observaciones discretas del proceso en el tiempo. En este caso el ACP clásico de observaciones discretas no igualmente espaciadas en el tiempo proporcionaría resultados erróneos (Castro *et al.* (1986)) y no permitiría, además, la reconstrucción de las

funciones muestrales entre los tiempos de observación. Por ello, este problema se resuelve recurriendo a técnicas numéricas eficientes. Para un estudio completo y riguroso sobre la solución numérica de ecuaciones integrales puede verse Baker (1977).

El método numérico más simple consiste en aproximar dicha ecuación integral mediante una fórmula de cuadratura compuesta. Así, Aguilera *et al.* (1992) han aplicado la fórmula de cuadratura del trapecio obteniendo muy buenas aproximaciones de los factores principales en los nodos de la partición elegida. Un método más sofisticado es el de proyección ortogonal que consiste en aproximar los factores principales en un subespacio de dimensión finita y resulta particularmente adecuado cuando se dispone de información a priori sobre la naturaleza de la solución exacta. Para el caso de procesos con trayectorias regulares, Aguilera *et al.* (1995) han resuelto el problema proyectando el proceso original sobre un subespacio finito dimensional de funciones trigonométricas. Además, ha sido contrastado mediante simulación (Aguilera *et al* (1996a)) que una interpolación spline cúbica de las funciones muestrales proporciona aproximaciones óptimas de los factores principales en este caso. Otros autores, como por ejemplo Besse y Ramsay (1986), resuelven este problema imponiendo que las trayectorias pertenezcan a un espacio de Sobolev adecuado. En el caso de procesos con funciones muestrales escalonadas los factores principales se aproximan proyectando sobre el subespacio de las funciones constantes en los intervalos de una partición previamente fijada en el periodo de observación (Aguilera *et al.* (1996b)). Además, para estimar el ACP de un proceso mediante estos métodos de aproximación, los autores han desarrollado el programa computacional PCAP, que ha sido codificado en Turbo Pascal usando programación orientada a objeto y matrices dinámicas (Aguilera *et al.* (1994)).

## 2. ACP DE UN PROCESO ESTOCÁSTICO

Como el modelo que vamos a desarrollar está basado en las componentes principales asociadas a un proceso estocástico en tiempo continuo, comenzaremos resumiendo brevemente las nociones básicas del ACP de un proceso de este tipo y abordando el problema de su estimación y aproximación a partir de funciones muestrales independientes.

### 2.1. Teoría Básica

Consideremos un proceso aleatorio $\{X(t): t \in [T_1, T_2]\}$ definido sobre el espacio probabilístico $(\Omega, A, P)$, con funciones muestrales en el espacio de Hilbert separable

$L^2[T_1, T_2]$, de las funciones de cuadrado integrable sobre $[T_1, T_2]$, con producto escalar definido por

$$< f|g >= \int_{T_1}^{T_2} f(t)g(t)dt, \quad \forall f, g \in L^2[T_1, T_2]$$

Supongamos, además, que el proceso $\{X(t)\}$ es continuo en media cuadrática y de segundo orden, con función media $\mu(t)$, y función de covarianza C(t,s), como consecuencia continua $\forall t, s \in [T_1, T_2]$.

Bajo estas condiciones de regularidad, el teorema de Mercer da la siguiente representación uniformemente convergente para la función de covarianza (ver, por ejemplo, Riesz y Sz-Nagy (1990))

$$C(t, s) = \sum_i \lambda_i f_i(t) f_i(s), \quad \forall t, s \in [T_1, T_2], \qquad [2.1]$$

siendo $\{\lambda_i\}$ la sucesión decreciente de valores propios del operador de covarianza asociado al proceso y $\{f_i\}$ la sucesión de funciones propias asociadas a los $\{\lambda_i\}$. Es decir, las funciones $\{f_i\}$ forman un conjunto ortonormal completo, en $L^2[T_1, T_2]$, de soluciones de la ecuación

$$\int_{T_1}^{T_2} C(t, s) f_i(s) ds = \lambda_i f_i(t) \qquad [2.2]$$

La extensión natural de la definición de Hötelling del ACP lleva a definir la *i-ésima* componente principal (c.p.) asociada al proceso $\{X(t): t \in [T_1, T_2]\}$ como la v.a.

$$\xi_i = \int_{T_1}^{T_2} (X(t) - \mu(t)) f_i(t) dt \qquad [2.3]$$

Las componentes principales así definidas tienen las mismas propiedades de optimalidad que en el caso finito. De hecho, la *i-ésima* c.p., $\xi_i$, es una combinación lineal generalizada de las variables del proceso que es centrada y tiene varianza máxima, $\lambda_i$, de entre todas aquellas que son incorreladas con $\{\xi_j\}_{j=1}^{i-1}$. A la varianza $\lambda_i$ se le llama *i-ésimo* valor principal y a $f_i$ *i-ésimo* factor principal.

Si denotamos por V a la varianza total del proceso definida por

$$V = E\left[\int_{T_1}^{T_2} (X(t) - \mu(t))^2 dt\right] = \sum_i \lambda_i < \infty, \qquad [2.4]$$

la cantidad $V_i=\lambda_i/V$ es llamada varianza explicada por la *i-ésima* c.p.

Entonces, el proceso admite la siguiente descomposición en componentes principales (ver, por ejemplo, Todorovic (1992)), que es conocida en el contexto probabilístico como desarrollo de Karhunen-Loève:

$$X(t) - \mu(t) = \sum_i \xi_i f_i(t), \quad t \in [T_1, T_2], \qquad [2.5]$$

donde la serie del segundo miembro converge uniformemente en media cuadrática en el intervalo $[T_1, T_2]$.

La representación ortogonal del proceso en términos de sus componentes principales es óptima debido a que la serie [2.5] truncada en el *q-ésimo* término es el mejor modelo lineal de dimensión *q* para $\{X(t)\}$ en el sentido de mínimos cuadrados (ver, por ejemplo, Fukunaga (1990)), de modo que $\sum_{i=1}^{q} \lambda_i$ es la varianza explicada por dicho modelo, y $\sum_{i=q+1}^{\infty} \lambda_i$ es el error cuadrático medio mínimo.

## 2.2. Estimación

A continuación abordaremos el problema de la estimación de los factores y componentes principales a partir de la información proporcionada por *N* funciones muestrales independientes del proceso aleatorio $\{X(t)\}$ a las que denotaremos

$$\{X_\omega(t): t \in [T_1, T_2], \quad \omega = 1, 2, \ldots, N\}$$

El estimador natural de C(t,s) es la función de covarianza muestral

$$\hat{C}(t,s) = \frac{1}{N-1} \sum_{\omega=1}^{N} \left(X_\omega(t) - \overline{X}(t)\right)\left(X_\omega(s) - \overline{X}(s)\right), \qquad [2.6]$$

siendo $\overline{X}$ el estimador insesgado de la media $\mu$ definido por

$$\overline{X}(t) = \frac{1}{N} \sum_{\omega=1}^{N} X_\omega(t) \qquad [2.7]$$

Como consecuencia de las propiedades de estos estimadores, que han sido estudiadas detalladamente por Deville (1973), los valores propios y funciones propias, $(\lambda_i, f_i)$, del núcleo *C(t,s)* se estiman mediante los correspondientes valores propios y funciones propias, $(\hat{\lambda}_i, \hat{f}_i)$, del núcleo $\hat{C}(t,s)$. Por lo tanto, los factores

principales muestrales, $\hat{f}_i$, en el intervalo $[T_1, T_2]$ son las soluciones de la siguiente ecuación integral de segundo orden:

$$\int_{T_1}^{T_2} \hat{C}(t,s)\hat{f}_i(s)ds = \hat{\lambda}_i \hat{f}_i(t), \quad t \in [T_1, T_2]. \qquad [2.8]$$

En el caso, poco usual, de valores propios $\lambda_i$ múltiples su estimación muestral se define promediando los correspondientes valores propios $\hat{\lambda}_i$ de $\hat{C}$ y los factores principales no estarían determinados de forma única.

Finalmente, el estimador natural de la varianza explicada por la *i-ésima* componente principal es el cociente $\hat{\lambda}_i / \hat{V}$, siendo $\hat{V} = \sum_i \hat{\lambda}_i$ un estimador insesgado y consistente para la varianza total V.

Una vez estimados los factores principales muestrales, la *i-ésima* c.p muestral asociada al proceso en el intervalo $[T_1, T_2]$ viene definida por

$$\xi_i = \int_{T_1}^{T_2} \left(X(t) - \hat{X}(t)\right)\hat{f}_i(t)dt. \qquad [2.9]$$

## 2.3. Aproximación de los Factores Principales de Procesos con Funciones Muestrales Escalonadas

Desafortunadamente, resolver la ecuación integral [2.8] es una tarea muy difícil que solo tiene solución exacta para núcleos muy especiales reduciendo la ecuación integral a una ecuación diferencial.

El punto de partida de este trabajo es aquellos procesos $\{X(t)\}$ cuyas trayectorias permanecen constantes en intervalos aleatorios, como, por ejemplo, los procesos puntuales y de recuento. En este caso, un método numérico eficiente consiste en aproximar los factores principales mediante los de la proyección de las funciones muestrales originales sobre el subespacio de las funciones constantes en los intervalos fijos de una partición previamente elegida en $[T_1, T_2]$. A continuación presentamos un breve resumen de este procedimiento de aproximación de los factores principales muestrales que ha sido desarrollado en el trabajo de Aguilera *et al.* (1996b).

Fijemos en el intervalo $[T_1, T_2]$ una partición $\pi_n$ definida por los nodos

$$T_1 = a_0 < a_1 < \ldots < a_n = T_2$$

verificando

$$\Delta_n = \max_{j=1,\ldots,n}\left\{\left(a_j - a_{j-1}\right)\right\} \to 0 \text{ cuando } n \to \infty$$

Sea $E_n$ el subespacio de las funciones constantes sobre cada uno de los intervalos $\left(a_{j-1}, a_j\right]$ $(j = 1,\ldots,n)$. Una base ortonormal de $E_n$ viene dada por las funciones

$$\delta_j(t) = \left(a_j - a_{j-1}\right)^{-1/2} I_j(t),$$

siendo $I_j$ la función indicadora en el intervalo $\left(a_{j-1}, a_j\right]$.

Para cada realización particular $X_\omega(t)$ del proceso en la muestra, su proyección ortogonal sobre este subespacio de funciones constantes es de la forma

$$X_\omega^{(n)}(t) = P_n\left[X_\omega(t)\right] = \sum_{j=1}^{n} Y_{\omega j}\delta_j(t) = \sum_{j=1}^{n} M_{\omega j} I_j(t) \qquad [2.10]$$

definiendo, para cada $\omega = 1, \ldots, N$ y $j = 1, \ldots, n$

$$Y_{\omega j} = < X_\omega \left| \delta_j > = \int_{T_1}^{T_2} X_\omega(t)\delta_j(t)dt = \left(a_j - a_{j-1}\right)^{1/2} M_{\omega j} \qquad [2.11]$$

siendo $M_{\omega j}$ el valor medio de la trayectoria muestral $\omega$ sobre el intervalo $\left(a_{j-1}, a_j\right]$ definido por

$$M_{\omega j} = \left(a_j - a_{j-1}\right)^{-1} \int_{a_{j-1}}^{a_j} X_\omega(t)dt$$

Entonces, se demuestra que los valores principales muestrales del proceso proyectado $X^{(n)}(t)$ son los valores propios de la matriz **R** de dimensión n x n con elementos

$$R_{ij} = \int_{T_1}^{T_2}\int_{T_1}^{T_2} \hat{C}(t,)\delta_i(t)\delta_j(s)dtds = \frac{1}{N-1}\sum_{\omega=1}^{N}\left(Y_{\omega i} - \overline{Y}_i\right)\left(Y_{\omega j} - \overline{Y}_j\right) \qquad [2.12]$$

definiendo,

$$\overline{Y}_j = \frac{1}{N}\sum_{\omega=1}^{N} Y_{\omega j} = \int_{T_1}^{T_2} \overline{X}(t)\delta_j(t)ds \qquad [2.13]$$

Además, los factores principales muestrales son funciones escalonadas dadas por

$$\hat{f}_i^{(n)} = \sum_{j=1}^{n} z_{ji}\delta_j,$$

donde el vector columna $z_i$ es el *i-ésimo* vector propio de la matriz **R** asociado a su *i-ésimo* valor propio $\hat{\lambda}_i^{(n)}$. Una vez obtenidos los factores principales aproximados $\hat{f}_i^{(n)}$, bajo la condición de normalización $\sum_{j=1}^{n}(z_{ji})^2 = 1$ las correspondientes componentes principales muestrales, a las que denotaremos $\hat{\xi}_i^{(n)}$, vienen dadas por:

$$\hat{\xi}_{\omega i}^{(n)} = \sum_{j=1}^{n} z_{ji}\left(Y_{\omega j} - \overline{Y}_j\right), \quad \omega = 1,\ldots,N \qquad [2.14]$$

Es decir, las componentes principales aproximadas son claramente las componentes principales asociadas a la matriz de datos **Y** de dimensión N x n con elementos $Y_{\omega j}$ definidos en [2.11].

## 3. CONSTRUCCIÓN DEL MODELO

Retomando el objetivo de este trabajo, nos planteamos ahora el problema de estimar la variable $X(s)\,(s > T_2)$ a partir de las variables $\{X(t): T_1 < t < T_2\}$ que, como se pondrá de manifiesto a la hora de estimar el modelo, resuelve el problema de estimación de datos faltantes y de filtrado en el instante *s* para distintas realizaciones muestrales del proceso.

Es conocido que, para cada $s>T_2$, el mejor estimador en el sentido de mínimos cuadrados es la esperanza condicional de la variable X(s) a las variables $\{X(t): t \in [T_1, T_2]\}$. Ante la dificultad de obtener solución explícita a este problema, nos centraremos en predición de tipo lineal.

Supongamos, sin pérdida de generalidad, que el proceso es centrado $(\mu(t) = 0)$ y denotemos por $L_x^2$ el subespacio cerrado de $L^2(\Omega)$ engendrado linealmente por las variables $\{X(t): t \in [T_1, T_2]\}$.

En esta situación el estimador lineal de mínimos cuadrados de X(s) $(s > T_2)$ es la v.a. $\tilde{X}(s)$ de $L_x^2$ que verifica

$$E\left[\left|\tilde{X}(s) - X(s)\right|^2\right] = \inf\left\{E\left[\left|Z - X(s)\right|^2\right]: Z \in L_x^2\right\},$$

es decir, $\tilde{X}(s)$ es, en terminología de espacios de Hilbert, la proyección ortogonal de X(s) sobre el subespacio $L_X^2$ y la cantidad $\delta^2(s) = E\left[\left|\tilde{X}(s) - X(s)\right|^2\right]$ recibe el nombre de error cuadrático medio de la estimación lineal.

Como las componentes principales $\{\xi_i\}$ forman una base ortogonal de $L_X^2$ (ver, por ejemplo, Todorovic (1992)), el conjunto $\{Z_i\}$ de componentes principales normalizadas definidas como $Z_i = \lambda_i^{-1/2}\xi_i$ constituyen una base ortonormal del subespacio $L_X^2$. Por lo tanto, como consecuencia del teorema de proyección ortogonal, el estimador lineal mínimo cuadrático $\tilde{X}(s)$ admite, para cada $s>T_2$, la siguiente representación convergente en media cuadrática en términos de las componentes principales (Deville (1978)):

$$\tilde{X}(s) = \sum_{i=1}^{\infty} \beta_i(s)\xi_i, \qquad [3.1]$$

definiendo

$$\beta_i(s) = \frac{E[X(s)\xi_i]}{\lambda_i}$$

El coeficiente de correlación lineal múltiple entre X(s) y las variables $\{X(t): T_1 < t < T_2\}$ viene dado por

$$R^2 = \sum_{i=1}^{\infty} r_i^2(s), \qquad [3.2]$$

siendo $r_i(s)$ el coeficiente de correlación lineal entre X(s) y $\xi_i$.

Entonces, truncando la serie infinita en la ecuación [3.1], podemos construir la siguiente estimación lineal aproximada:

$$\tilde{X}^p(s) = \sum_{i=1}^{p} \beta_i(s)\xi_i, \qquad [3.3]$$

que llamaremos Estimación en Componentes Principales (ECP) y verifica

$$\lim_{p\to\infty} E\left[\left(\tilde{X}(s) - \tilde{X}^p(s)\right)^2\right] = 0, \quad s > T_2$$

Observemos que la ecuación [3.3] representa la estimación lineal mínimo cuadrática de X(s) sobre las p primeras componentes principales como predictores.

Una vez definido el modelo ECP nos planteamos su identificación y estimación a partir de la información proporcionada por N funciones muestrales escalonadas independientes del proceso a las que denotaremos

$$\{X_\omega(t): t \in [T_1, T_2]; \quad \omega = 1, \ldots, N\}$$

y una muestra aleatoria de tamaño N de X(s), denotada por

$$\{X_\omega(s): \omega = 1, \ldots, N\}$$

El problema que se plantea en la identificación del modelo ECP es elegir aquellas componentes principales que serán introducidas como los predictores óptimos del futuro X(s). La práctica usual consiste en borrar automáticamente como explicativas aquellas componentes principales asociadas a valores propios pequeños. Sin embargo, algunos autores, como por ejemplo Hötelling y Joliffe (Jackson (1992)), han demostrado que no hay razón para que sean las cc.pp. con mayor varianza los mejores predictores. De hecho, puede ocurrir que alguna de las cc.pp. menos explicativas en el pasado estén altamente correladas con el proceso en el futuro.

Teniendo en cuenta esta última observación, el procedimiento de identificación y estimación del modelo ECP para un proceso con trayectorias escalonadas se puede resumir en los siguientes pasos:

1. Elección en el intervalo $[T_1, T_2]$ de una partición de nodos

$$T_1 = a_0 < a_1 < \ldots < a_n = T_2,$$

y estimación de las componentes principales aproximadas $\left\{\xi_{\omega i}^{(n)}\right\}$ como se expuso en la subsección 2.3. Observemos que la elección de la partición estará en función de la densidad de cambios de valor del proceso en el tiempo observado. De este modo los subintervalos no tienen que tener la misma amplitud y serán más finos en aquellos periodos en los que se producen más cambios.

2. Estimación de las correlaciones lineales $\hat{r}_i^{2(s)}$ y reordenación de las cc.pp. en orden decreciente a la correlación estimada, seleccionando como predictores óptimos aquellas *p* primeras cc.pp. $\xi_i^{(n)}$, según este nuevo orden, cuya correlación lineal con X(s) sea significativamente alta, verificando

$$\hat{R}^2 = \sum_{i=1}^{p} \hat{r}_i^2(s) \cong 1.$$

3. Estimación en la forma usual del modelo de regresión lineal de X(s) sobre las p cc.pp. seleccionadas

$$\hat{X}^p(s) = \overline{X}(s) + \sum_{i=1}^{p} \hat{\beta}_i(s)\hat{\xi}_i^{(n)}$$

4. Para cada nuevo individuo $\omega$ del que se disponga solo de su trayectoria en $[T_1, T_2]$, estimación de $X_\omega(s)$ sin más que evaluar sus cc.pp. $\left\{\hat{\xi}_{\omega i}\right\}_{i=1}^{p}$ mediante la expresión [2.14], y sustituir en el modelo ECP previamente estimado.

## 4. APLICACIÓN CON DATOS SIMULADOS

En esta sección vamos a ilustrar el modelo ECP desarrollado anteriormente mediante una aplicación con datos simulados de un proceso con funciones muestrales escalonadas. El proceso considerado corresponde a una cola del tipo $M^X|M|1$ donde las llegadas múltiples están generadas por una variable *X* con distribución uniforme discreta con valores {1,2,3,4 } y el tiempo entre llegadas se ajusta a una distribución exponencial de media 6.

De este proceso estocástico se han simulado un total de 50 trayectorias,de las cuales 40 han sido utilizadas para la estimación de los parámetros del modelo y las 10 restantes para evaluar su capacidad predictiva. Dicha simulación se ha realizado sobre el intervalo de tiempo [0,220] y el objetivo consiste en estimar el valor del proceso en el instante 220 a partir de su evolución en el intervalo [0,200].

Dado que las trayectorias de este proceso son claramente constantes en intervalos aleatorios, se ha considerado adecuado, para estimar las componentes principales, su proyección sobre el subespacio de las funciones constantes en 20 subintervalos de amplitud 10, cubriendo así el intervalo pasado. En la Tabla 1 aparecen las varianzas explicadas por las componentes principales estimadas mediante el programa PCAP, observándose que entre las tres primeras el porcentaje de varianza acumulada es superior al 95%.

En la Figura 1 aparece la proyección de dos de las trayectorias muestrales superpuesta con su reconstrucción mediante las tres primeras componentes principales.

**Tabla 1**

VARIANZA EXPLICADA POR LAS COMPONENTES PRINCIPALES Y CUADRADO DE SUS CORRELACIONES CON EL NÚMERO ACUMULADO DE LLEGADAS X(220)

| $\hat{\xi}_i$ | $V_i$ | $\sum_{j=1}^{i} V_j$ | $\hat{r}_i^2$ |
|---|---|---|---|
| 1 | 81.46990 | 81.46990 | 0.778806 |
| 2 | 10.42800 | 91.89790 | 0.016874 |
| 3 | 3.11408 | 95.01198 | 0.068225 |
| 4 | 1.98729 | 96.99927 | 0.009781 |
| 5 | 0.85287 | 97.85214 | 0.000576 |
| 6 | 0.66297 | 98.51512 | 0.002992 |
| 7 | 0.40316 | 98.91827 | 0.002809 |
| 8 | 0.27581 | 99.19409 | 0.000750 |
| 9 | 0.19662 | 99.39071 | 0.011578 |
| 10 | 0.14481 | 99.53551 | 0.004610 |
| 11 | 0.12224 | 99.65775 | 0.020909 |
| 12 | 0.08460 | 99.74235 | 0.021815 |
| 13 | 0.07188 | 99.81423 | 0.002294 |
| 14 | 0.04576 | 99.85999 | 0.000005 |
| 15 | 0.03222 | 99.89221 | 0.001096 |
| 16 | 0.03033 | 99.92254 | 0.000079 |
| 17 | 0.02694 | 99.94947 | 0.000449 |
| 18 | 0.01961 | 99.96908 | 0.000012 |
| 19 | 0.01656 | 99.98565 | 0.001116 |
| 20 | 0.01435 | 100.00000 | 0.000534 |

Una vez estimadas las componentes principales se han calculado las correlaciones lineales entre las cc.pp. y el número acumulado de llegadas en el instante 220. Los cuadrados de estas correlaciones lineales aparecen recogidos también en la Tabla 1.

Observemos que la correlación lineal con la primera componente principal es muy alta, siendo las restantes correlaciones lineales considerablemente inferiores. No obstante, combinando el hecho de que la segunda y tercera componentes principales explican una parte importante de la varianza total y que sus correlaciones lineales con respecto a la variable dependiente han resultado significativas aplicando regresión stepwise, se han ajustado los tres modelos ECP siguientes, en

términos de la primera, dos primeras y tres primeras componentes principales del pasado, respectivamente:

$$\text{ECP(1):} \quad \tilde{X}^1(220) = 91.6250 + 0.1026\,\hat{\xi}_1^{(20)},$$

$$\text{ECP(1,2):} \quad \tilde{X}^2(220) = 91.6250 + 0.1026\,\hat{\xi}_1^{(20)} - 0.0422\,\hat{\xi}_2^{(20)},$$

$$\text{ECP(1,2,3):} \quad \tilde{X}^3(220) = 91.6250 + 0.1026\,\hat{\xi}_1^{(20)} - 0.0422\,\hat{\xi}_2^{(20)} + 0.1554\,\hat{\xi}_3^{(20)},$$

con los siguientes coeficientes de correlación múltiple estimados: $\hat{R}^2(1) = 0.7788$, $\hat{R}^2(1,2) = 0.7956$, $\hat{R}^2(1,2,3) = 0.8638$. Los cálculos relacionados con el ajuste lineal mínimo cuadrático han sido realizados con los programas 1R y 2R de BMDP.

En la Tabla 2 figuran los valores simulados en t=220 para las 10 últimas trayectorias del proceso acumulado de llegadas múltiples, sus estimaciones mediante los tres modelos anteriores, los residuos asociados y la raiz cuadrada del error cuadrático medio de estimación definido por

$$\varepsilon_p^2 = \frac{1}{10}\sum_{i=1}^{10}\left(X(220) - \tilde{X}^p(220)\right)^2, \quad p = 1,2,3.$$

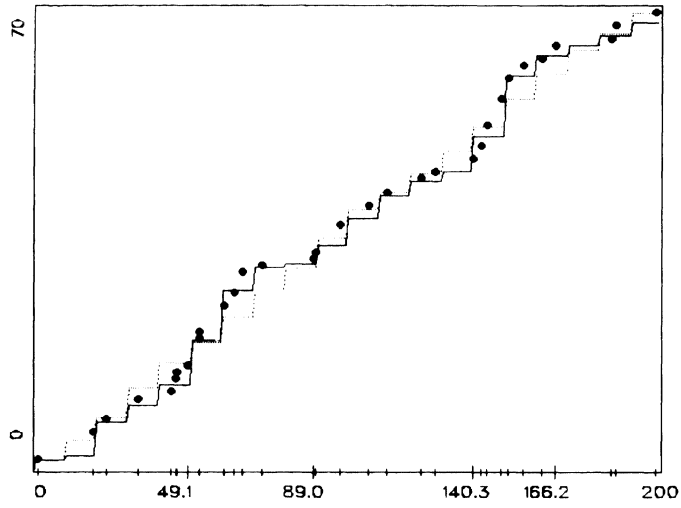
**Tabla 2**

VALORES SIMULADOS X(220) Y SUS ESTIMACIONES MEDIANTE LOS MODELOS ECP CON LOS RESIDUOS ASOCIADOS ENTRE PARÉNTESIS Y LOS ERRORES DE ESTIMACIÓN

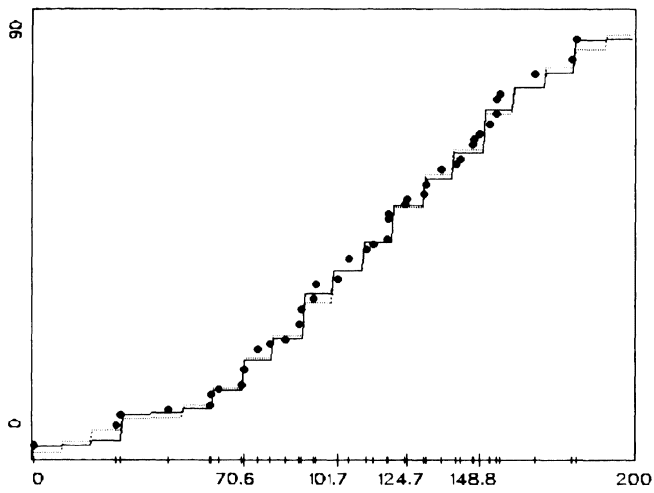
| *Tray.* | *X(220)* | *ECP(1)* | | *ECP(1,2)* | | *ECP(1,2,3)* | |
|---|---|---|---|---|---|---|---|
| 41 | 105 | 93.00 | (-12.00) | 91.90 | (-13.10) | 95.51 | (-9.49) |
| 42 | 80 | 74.45 | (-5.55) | 76.83 | (-3.17) | 78.85 | (-1.15) |
| 43 | 68 | 73.95 | (5.95) | 71.51 | (3.51) | 67.39 | (-0.61) |
| 44 | 75 | 83.32 | (8.32) | 80.62 | (5.62) | 74.68 | (-0.32) |
| 45 | 119 | 121.10 | (2.10) | 120.20 | (1.20) | 127.20 | (8.20) |
| 46 | 93 | 75.40 | (-17.60) | 80.56 | (-12.44) | 84.89 | (-8.11) |
| 47 | 83 | 77.03 | (-5.97) | 76.99 | (-6.01) | 76.35 | (-6.65) |
| 48 | 108 | 111.10 | (3.10) | 109.20 | (1.20) | 107.00 | (-1.00) |
| 49 | 100 | 100.20 | (0.20) | 100.80 | (0.80) | 99.86 | (-0.14) |
| 50 | 96 | 93.18 | (-2.82) | 92.66 | (-3.34) | 100.50 | (4.50) |
| | | $\varepsilon_1$=*8.04* | | $\varepsilon_2$=*6.57* | | $\varepsilon_3$=*5.39* | |

**Figura 1**

NUBE DE PUNTOS PARA EL NÚMERO ACUMULADO DE LLEGADAS EN [0,200], PROYECCIÓN SOBRE EL SUBESPACIO DE LAS FUNCIONES CONSTANTES EN INTERVALOS DE AMPLITUD 10 (LÍNEA CONTINUA) Y SU RECONSTRUCCIÓN MEDIANTE LAS TRES PRIMERAS COMPONENTES PRINCIPALES (LÍNEA DISCONTINUA) PARA LAS TRAYECTORIAS MUESTRALES (A) Y (B)




(a)




(b)

## 5. CONCLUSIONES

A la vista de los resultados que proporciona la Tabla 2 se observa que los modelos ECP proporcionan *buenas* estimaciones. Observemos que, a pesar de que las correlaciones de la variable dependiente con la segunda y tercera componentes principales no son muy significativas, los modelos ECP(1, 2) y ECP(1, 2, 3) disminuyen considerablemente el error de predicción y proporcionan estimaciones más precisas en la mayoría de los casos.

Resumiendo podemos concluir que los modelos ECP introducidos en este tienen las siguientes ventajas:

1. La estimación $\tilde{X}^p(s)$ converge en media cuadrática al estimador lineal mínimo cuadrático $\tilde{X}(s)$ y ha sido obtenida sin imponer propiedades restrictivas como la estacionariedad.

2. Si el proceso que queremos predecir fuese superposición de una señal y un ruido, la eliminación de aquellas componentes principales poco explicativas filtraría la señal original proporcionando una estimación óptima del futuro.

3. La incorrelación de las componentes principales evita el problema de multicolinealidad de la regresión lineal múltiple.

4. La reducción de dimensión que proporciona el ACP lleva a modelos ECP extremadamente sencillos. Además, si las componentes principales fuesen fácilmente interpretables, las ecuaciones de regresión serían más significativas y facilmente estimables.

5. Permiten la recuperación de datos faltantes (*missing*) de las trayectorias de un proceso estocástico.

## REFERENCIAS

AGUILERA, A.M., VALDERRAMA, M.J., Y DEL MORAL M.J. (1992). «Un Método para la Aproximación de Estimadores en ACP. Aplicación al Proceso de Ornstein-Uhlenbeck», *Revista de la SOCHE*, 9, 57-77.

AGUILERA, A.M., OCAÑA F.A., Y VALDERRAMA, M.J. (1994). «A computational Algorithm for PCA of Random Processes», *Proceedings in COMPSTAT. Software Descriptions,* Ed. Dutter, R. y Grossmann, W., pp. 39-40. Berlin: Physica-Verlag.

AGUILERA, A.M., GUTIÉRREZ, R. OCAÑA, F.A. Y VALDERRAMA, M.J.(1995). «Computational approaches to estimation in the principal component analysis of a stochastic process». *Applied Stochastic Models and Data Analysis,* 11(4), 279-299.

AGUILERA, A.M., GUTIÉRREZ, R. AND VALDERRAMA, M.J. (1996a). «Approximation of estimators in the PCA of a stochastic process using B-Splines». *Communications in Statistics,* 25(3), 671-690.

AGUILERA, A.M., OCAÑA F.A., Y VALDERRAMA, M.J. (1996b). «Análisis en Componentes Principales de un Proceso Estocástico con Funciones Muestrales Escalonadas», *Qüestiio,* 20, 7-28..

BAKER, C.T.H. (1977). «The Numerical Treatment of Integral *Equations»,* London: Oxford University Press.

BESSE, P. Y RAMSAY, J.O. (1986). «Principal Components Analysis of Sample Functions», *Psychometrika,* 51, 285-311.

CASTRO, P.E., LAWTON, W.H., Y SYLVESTRE, E.A. (1986). «Principal Modes of Variation for Processes With Continuous Sample Curves», *Technometrics,* 28, 329-337.

DAUXOIS, J., POUSSE, A. Y ROMAIN, Y. (1982). «Asymptotic Theory for the Principal Component Analysis of a Vector Random Function: Some Applications to Statistical Inference», *Journal of Multivariate Analysis,* 12, 136-154.

DEVILLE, J.C. (1973). «Estimation of the Eigenvalues y of the Eigenvectors of a Covariance Operator», *Note Interne de l'INSEE.*

DEVILLE, J.C. (1974). «Méthodes Statistiques et Numériques de l'Analyse Harmonique», *Annales de l'INSEE,* 15, 3-101.

DEVILLE, J.C. (1978). «Analyse et Prevision des Series Chronologiques Multiples Non Stationnaires», *Statistique et Analyse des Données,* 3, 19-29.

FUKUNAGA, K. (1990). «Introduction to Statistical Pattern Recognition», 2ª ed., San Diego: Academic Press.

JACKSON, J.E. (1991). «A User's Guide to Principal Components», New York: Wiley.

RAMSAY, J.O. Y SILVERMAN, B.W. (1997). «Functional Data Analysis». New York: Springer-Verlag.

RIESZ, F. Y SZ-NAGY, B. (1990). «Lecons d'Analyse Fonctionnelle», Reimpresión de la 3ª edición publicada por Gauthier-Villars y Akadémiai Kiadó en 1955, Sceaux: Ediciones Jacques Gabay.

TODOROVIC, P. (1992). «An Intoduction to Stochastic Processes y Their Applications», New York: Springer-Verlag.

## REGRESSION ON PRINCIPAL COMPONENTS OF A STOCHASTIC PROCESS WHOSE SAMPLE PATHS ARE PIECEWISE CONSTANT FUNCTIONS


### SUMMARY

The objective of this paper is to develop a procedure for estimating missing observations of a stochastic process whose sample paths are piecewise constant functions, in terms of its evolution in the past. This model is based on multiple linear regression against the principal components in the past. The sample principal factors of such a process are approximated by projection of the original sample paths on the subspace of the piecewise constant functions over the subintervals of a partition previously fixed. Finally, an application with simulated data is included.


*Key Words:* Covariance Function, Principal Components, Least Squares Linear Estimation, Orthogonal Projection.

*AMS Classification:* 60G12, 65D30, 62H25.