QÜESTIIÓ, Vol 20, 1 pp. 7-28, 1996

ANÁLISIS EN COMPONENTES PRINCIPALES DE UN PROCESO ESTOCÁSTICO CON FUNCIONES MUESTRALES ESCALONADAS

ANA Mª AGUILERA DEL PINO

FRANCISCO A. OCAÑA LARA

MARIANO J. VALDERRAMA BONNET

Universidad de Granada*

El ACP de un número finito de variables puede ser generalizado para manejar datos que evolucionan en el tiempo. El objetivo de este trabajo es la estimación de los factores principales de procesos aleatorios con funciones muestrales escalonadas. Ante la imposibilidad de obtener una solución exacta a este problema, proponemos estimar el ACP de un proceso de este tipo a partir del ACP del proceso cuyas trayectorias se obtienen como proyección de las originales en el subespacio de las funciones constantes sobre los subintervalos de una partición previamente fijada. Finalmente, incluimos una aplicación con datos reales estudiando el error cuadrático medio de las reconstrucciones del proceso proporcionadas por el ACP así aproximado.

Principal Component Analysis of a Stochastic Process whose Sample Paths are Piecewise Constant Functions.

Key words: Operador de covarianza, componentes principales, autovalores y autofunciones, proyección ortogonal.

Clasificación AMS: 60G12, 62H25.

* Departamento de Estadística e Investigación Operativa. Universidad de Granada. Campus de Fuentenueva. 18071. Granada, España.

- Article rebut el juliol de 1994.

- Acceptat el juny de 1995.

1. INTRODUCCIÓN

El Análisis en Componentes Principales (ACP) de un conjunto finito de variables aleatorias correladas tiene como finalidad la reducción de dimensión mediante la obtención de combinaciones lineales con varianza máxima e incorreladas de las variables originales. En el campo de las ciencias experimentales, económicas y sociales, es frecuente encontrar individuos caracterizados por una curva, que corresponde a las observaciones de una variable a lo largo del tiempo, en lugar de un vector de datos como ocurre en análisis multivariante. El estudio estadístico de datos de este tipo se enmarca clásicamente en la teoría de los procesos estocásticos de segundo orden en tiempo continuo imponiendo, generalmente, que se verifiquen hipótesis bastante restrictivas como estacionariedad o su pertenencia a una clase general de procesos como, por ejemplo, procesos de Markov.

Ante la existencia de una amplia gama de aplicaciones que no se ajustan a estos modelos, Church (1966) y Deville (1974) desarrollaron una extensión del ACP clásico para la reducción de dimensión en el caso de procesos estocásticos en tiempo continuo. La base de esta teoría es una propiedad probabilística clásica, llamada desarrollo de Karhunen-Loéve, que proporciona una representación ortogonal del proceso como suma de funciones ortogonales ponderadas por variables aleatorias incorreladas que son las componentes principales. Dichas funciones ortogonales son las autofunciones del operador de covarianza del proceso. En el trabajo de Gutiérrez *et al.* (1992) se estudian dos métodos numéricos para el cálculo aproximado de estas autofunciones, los métodos de Rayleigh-Ritz y de colocación, contrastando su eficiencia mediante comparación con las funciones propias de covarianzas conocidas como, por ejemplo, la del proceso de Wiener-Lévy.

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 autofunciones 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. 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. En el caso en que el proceso es observado sólo en un conjunto finito de instantes de tiempo, Pardoux (1989) propone interpolar linealmente las funciones muestrales entre los valores observados. Un método más sofisticado es el de proyección ortogonal que resulta particularmente adecuado cuando se dispone de información a priori sobre la naturaleza de la solución exacta (Deville (1974)). Para el caso de procesos con funciones muestrales regulares, Aguilera *et al.* (1993) han resuelto el problema proyectando el proceso original sobre un subespacio finito-dimensional de funciones trigonométricas. En este trabajo nos proponemos aplicar el método de proyección ortogonal para calcular de forma aproximada los factores principales de procesos cuyas funciones muestrales permanecen constantes en intervalos aleatorios (por ejemplo, el proceso de Poisson).

2. ACP DE UN PROCESO ESTOCÁSTICO

Consideraremos un proceso aleatorio $\{X(t) : t \in [0,T]\}$ definido sobre el espacio probabilístico $(\Omega, \mathcal{A}, P)$, con funciones muestrales, denotadas por $X(w) = X(.,w)$ para cualquier $w \in \Omega$ fijo, en el espacio de Hilbert separable $L^2[0,T]$ de las funciones de cuadrado integrable sobre $[0,T]$, con producto escalar definido por:

$$< f \mid g >= \int_0^T f(t)g(t)dt, \quad \forall f,g \in L^2[0,T].$$

Supondremos, además, que el proceso $\{X(t)\}$ es de segundo orden y continuo en media cuadrática. Llamaremos $\mu(t)$ a su función media y $C(t,s)$ a su función de covarianza.

2.1. Teoría Básica

Definición 2.1

Bajo las condiciones de regularidad anteriores, el operador de covarianza, C, asociado al proceso $\{X(t)\}$ es un operador de Hilbert-Schmidt, sobre $L^2[0,T]$, con núcleo la función de covarianza C(t,s), dado por:

$$C(f(t)) = \int_0^T C(t,s)f(s)ds, \quad \forall f \in [0,T].$$

El operador C es compacto, autoadjunto y positivo (Deville (1974)). Como consecuencia de estas propiedades, el teorema de Mercer da la siguiente representación

uniformemente convergente para la función de covarianza (ver, por ejemplo, Riesz y Sz-Nagy (1990), p. 242):

$$C(t,s) = \sum_i \lambda_i f_i(t) f_i(s), \quad \forall t,s \in [0,T],$$

siendo $\{\lambda_i\}$ la sucesión decreciente de valores propios de C y $\{f_i\}$ la sucesión de funciones propias de C asociadas a los $\{\lambda_i\}$ que constituyen una base ortonormal en $L^2[0,T]$. Es decir, forman un conjunto ortonormal completo de soluciones de la ecuación:

$$\int_0^T C(t,s) f_i(s) ds = \lambda_i f_i(t).$$

Entonces se obtiene la siguiente representación ortogonal del proceso (ver, teorema de Kac y Siegert, Shorack and Wellner (1986), p. 210).

Lema 2.2

Sea $\{\xi_i\}$ la familia de variables aleatorias definida por:

$$\xi_i = \int_0^T f_i(t)(X(t) - \mu(t)) dt.$$

Entonces,

1. El proceso $\{X(t)\}$ admite la siguiente representación ortogonal

$$X(t) - \mu(t) = \sum_i \xi_i f_i(t) \tag{2.1}$$

   donde la serie del segundo miembro converge uniformemente en media cuadrática en $[0,T]$.

2. $E[\xi_i] = 0$, $Cov[\xi_i, \xi_j] = \lambda_i \delta_{ij}$, siendo δ_{ij} la Delta de Kronecker.

La variable aleatoria ξ_i se llama i-ésima componente principal puesto que, análogamente al caso finito, es una combinación lineal generalizada de las variables del proceso que tiene varianza máxima, λ_i, de entre todas aquellas que son incorreladas con ξ_j $\forall j = 1 \ldots i-1$. Del mismo modo, f_i recibe el nombre de i-ésimo factor principal y la representación ortogonal (2.1) se llama descomposición en componentes principales del proceso $\{X(t)\}$.

Si denotamos por V a la varianza total del proceso,

$$V = E\left[\int_0^T (X(t) - \mu(t))^2 dt\right] = tr(C) = \sum_i \lambda_i < \infty,$$

entonces la cantidad $V_i = \lambda_i / V$ es la varianza explicada por la i-ésima componente principal.

Además, la serie (2.1) truncada en el m-ésimo término es el mejor modelo lineal de dimensión m para $\{X(t)\}$ en el sentido de mínimos cuadrados (ver, por ejemplo, Fukunaga (1990), p. 149), de modo que $\sum_{i=1}^{m} \lambda_i$ es la varianza explicada por dicho modelo, y siendo

$$E_m = V - \sum_{i=1}^{m} \lambda_i$$

el error cuadrático medio mínimo.

2.2. Estimación Muestral

A continuación nos proponemos estimar los factores y componentes principales a partir de la información proporcionada por N funciones muestrales independientes del proceso aleatorio $\{X(t)\}$ a las que notaremos $\{X_k(t) : t \in [0,T], \quad k = 1,2,\ldots,N\}$.

Definición 2.3

Llamaremos operador de covarianza muestral $\hat{C}$ al operador de Hilbert-Schmidt de núcleo la función de covarianza muestral $\hat{C}(t,s)$ definida por

$$\hat{C}(t,s) = \frac{1}{N-1} \sum_{k=1}^{N} \left(X_k(t) - \bar{X}(t)\right) \left(X_k(s) - \bar{X}(s)\right),$$

donde $\bar{X}$ es el estimador natural de la media μ definido por

$$\bar{X}(t) = \frac{1}{N} \sum_{k=1}^{N} X_k(t).$$

Es decir,

$$\hat{C}(f(t)) = \int_0^T \hat{C}(t,s) f(s) ds \quad \forall f \in L^2([0,T]).$$

Las propiedades fundamentales del operador de covarianza muestral son las siguientes (Deville (1973)):

1. $\hat{C}$ es un estimador insesgado en C.
2. $\hat{C}$ es convergente en media cuadrática a C.

3. Suponiendo que los valores propios λ_i de C son simples y denotando por $\hat{\lambda}_i$ a los autovalores de $\hat{C}$ y por $\hat{f}_i$ a las correspondientes autofunciones, se verifica que, para cada i, $\hat{\lambda}_i$ y $\hat{f}_i$ convergen en media cuadrática a los correspondientes elementos propios λ_i y f_i de C, respectivamente, cuando $N \to \infty$.

Como consecuencia de estas propiedades tomaremos como estimadores naturales de los factores principales, f_i, las correspondientes funciones propias, $\hat{f}_i$, de $\hat{C}$ a las que llamaremos factores principales muestrales. En el caso, poco usual, de autovalores λ_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 $\hat{V}$ el estimador insesgado y consistente para la varianza total V, definido como:

$$\hat{V} = tr(\hat{C}) = \sum_i \hat{\lambda}_i.$$

3. OBTENCIÓN APROXIMADA DEL ACP

De acuerdo con lo expuesto en la sección anterior, los factores principales muestrales son las autofunciones del operador covarianza muestral, es decir, las soluciones de la siguiente ecuación integral de segunda especie:

$$\hat{C}(\hat{f}(t)) = \int_0^T \hat{C}(t,s)\hat{f}(s)ds = \hat{\lambda}\hat{f}(t). \tag{3.1}$$

Ante la imposibilidad de obtener soluciones exactas de esta ecuación, nuestro objetivo es desarrollar un método eficiente para aproximar los factores principales de procesos estocásticos cuyas funciones muestrales son escalonadas, es decir, constantes en intervalos que no son, en general, iguales para todas ellas. En este caso, y en todos aquellos en los que se conoce la naturaleza de las trayectorias del proceso, es aconsejable utilizar el método de proyección ortogonal. Este procedimiento numérico consiste en aproximar los factores principales en un subespacio de dimensión finita de $L^2[0,T]$ generado por un sistema ortonormal de funciones.

3.1. Método de Proyección Ortogonal

Comenzaremos con la definición de aproximación de un espacio de Hilbert.

Definición 3.4

Dado un espacio de Hilbert E, se llama aproximación de E a una sucesión $\{E_n\}_{n=1}^{\infty}$ de subespacios de dimensión finita de E verificando:

$$\lim_{n\to\infty} P_n x = x \quad \forall x \in E,$$

siendo $\{P_n\}_{n=1}^{\infty}$ la sucesión de proyecciones ortogonales de E sobre E_n.

Dada una aproximación $\{E_n\}$ del espacio de Hilbert E=$L^2[0,T]$, aplicar el método de proyección ortogonal para resolver de forma aproximada la ecuación (3.1) consiste en aproximar los autovalores y autofunciones del operador de covarianza muestral, $\hat{C}$, mediante los del operador $\hat{C}^{(n)} : L^2[0,T] \to E_n$ definido por

$$\hat{C}^{(n)} = P_n \hat{C} P_n. \tag{3.2}$$

Es decir, se sustituye la ecuación (3.1) por la siguiente ecuación aproximada:

$$\hat{C}^{(n)} \hat{f}^{(n)} = \hat{\lambda}^{(n)} \hat{f}^{(n)}, \tag{3.3}$$

denotando por $(\hat{\lambda}^{(n)}, \hat{f}^{(n)})$ a los autovalores y autofunciones aproximados.

El siguiente lema (Riesz y Sz-Nagy (1990)) garantiza la convergencia de las soluciones aproximadas por ser $\hat{C}$ un operador compacto, autoadjunto y positivo sobre el espacio de Hilbert $L^2[0,T]$ (Deville (1973)).

Lema 3.5

Sea $K : E \to E$ un operador compacto sobre un espacio de Hilbert E con producto escalar $< | >_E$. Sea $\{E_n\}_{n=1}^{\infty}$ una aproximación de E. Entonces,

1. La sucesión de operadores de rango finito $\{K_n\}_{n=1}^{\infty}$, definida por $K_n = P_n K P_n$ converge a K respecto de la norma como operador, es decir,

$$\lim_{n\to\infty} \| K_n - K \| = 0.$$

2. Sea p un entero y $\{\lambda_i\}_{i=1}^{p}$ la sucesión de los p mayores valores propios de K en orden decreciente, que supondremos simples con vectores propios asociados $\{x_i\}_{i=1}^{p}$ unitarios. Consideremos, además, que K es un operador autoadjunto y positivo. Entonces,

(a) Para cada $i = 1 \ldots p$, existe un autovalor $\lambda_i^{(n)}$ de K_n tal que la sucesión $\{\lambda_i^{(n)}\}_{n=1}^{\infty}$ converge en norma a λ_i, es decir,

$$\lim_{n\to\infty} | \lambda_i^{(n)} - \lambda_i | = 0,$$

una vez que los autovalores $\lambda_i^{(n)}$ han sido fijados en orden decreciente.

(b) Para cada $i = 1 \ldots p$, existe un vector propio $x_i^{(n)}$ de K_n asociado a $\lambda_i^{(n)}$ tal que la sucesión $\{x_i^{(n)}\}_{i=1}^{\infty}$ converge a x_i en norma. es decir,

$$\lim_{n\to\infty} || x_i^{(n)} - x_i ||_E = 0.$$

Veamos, a continuación, que el operador $\hat{C}^{(n)}$ dado por la ecuación (3.2) es precisamente el operador de covarianza muestral de la proyección del proceso $\{X(t)\}$ sobre el subespacio E_n definida como sigue.

Definición 3.6

Sea E_n un subespacio de una aproximación del espacio de Hilbert $L^2[0,T]$. Dado el proceso $\{X(t)\}$ continuo en media cuadrática, de segundo orden y con funciones muestrales en $L^2[0,T]$, denotaremos por $\{X^{(n)}(t)\}$ al proceso aleatorio que se obtiene proyectando cada una de las funciones muestrales del proceso $\{X(t)\}$ sobre E_n. Es decir,

$$X^{(n)}(w) = P_n(X(w)) = \sum_{j=1}^{n} Y_j(w) e_j,$$

siendo $\{e_j\}_{j=1}^{n}$ una base ortonormal de E_n y denotando por Y_j a la variable aleatoria real definida para casi todo w del espacio probabilístico $(\Omega, \mathcal{A}, P)$ en la siguiente forma:

$$Y_j(w) = < X(w) | e_j > = \int_0^T X(t,w) e_j(t) dt. \tag{3.4}$$

Análogamente, denotaremos por $\{X_k^{(n)}(t) : k = 1, \ldots N\}$ a la muestra aleatoria del proceso proyectado $\{X^{(n)}(t)\}$ asociada a la muestra $\{X_k(t) : k = 1, 2, \ldots, N\}$ del proceso aleatorio $\{X(t)\}$.

El siguiente resultado (Deville (1974)) proporciona las características fundamentales del proceso proyectado.

Lema 3.7

Sea $\{E_n\}$ una aproximación de $L^2[0,T]$ y $\{X^{(n)}(t)\}$ la proyección del proceso $\{X(t)\}$ sobre E_n. Entonces se verifican las siguientes propiedades:

1. La esperanza del proceso proyectado es la proyección de la función media de $\{X(t)\}$ dada por: $\mu^{(n)}(t) = E\left[X^{(n)}(t)\right] = P_n\mu(t)$.
2. El operador de covarianza de $\{X^{(n)}(t)\}$ viene dado por $C^{(n)} = P_nCP_n$.
3. La media muestral del proceso proyectado es la proyección de la media muestral de $\{X(t)\}$ definida como: $\bar{X}^{(n)} = P_n\bar{X}$.
4. El operador covarianza muestral de $\{X^{(n)}(t)\}$ es el operador $\hat{C}^{(n)} = P_n\hat{C}P_n$.

De acuerdo con esta proposición los factores principales aproximados con el método de proyección son los factores principales del proceso que se obtiene proyectando las funciones muestrales del proceso original sobre el subespacio E_n.

Finalmente, la siguiente proposición (Aguilera *et al.* (1993)) nos da la expresión de los factores principales aproximados con este método que son las soluciones de la ecuación (3.3).

Proposición 3.8

Sea $\{E_n\}$ una aproximación de $L^2[0,T]$ y $\{e_j\}_{j=1}^n$ una base ortonormal de E_n.

Entonces, las soluciones aproximadas, $(\hat{\lambda}^{(n)}, \hat{f}^{(n)})$, de la ecuación (3.1) aplicando el método de proyección sobre E_n son de la forma:

$$\hat{f}^{(n)} = \sum_{j=1}^{n} z^j e_j$$

donde el vector columna $\underline{z}$ de coordenadas z^j es solución de la ecuación matricial

$$R\underline{z} = \hat{\lambda}^{(n)}\underline{z},$$

siendo $\mathbf{R}$ la matriz simétrica de dimensión $n \times n$ con elementos

(3.5)

$$R_{ij} = < \hat{C}e_i \mid e_j > = \int_0^T \int_0^T \hat{C}(t,s)e_i(t)e_j(s)dtds = \frac{1}{N-1}\sum_{k=1}^{N}(Y_{ki} - \bar{Y}_i)(Y_{kj} - \bar{Y}_j)$$

y definiendo,

$$Y_{kj} = \int_0^T X_k(s)e_j(s)ds, \quad \forall j = 1 \ldots n,$$

$$\bar{Y}_j = \frac{1}{N}\sum_{k=1}^{N} Y_{kj} = \int_0^T \bar{X}(s)e_j(s)ds.$$

Una vez obtenidos los factores principales aproximados $\hat{f}^{(n)}$ bajo la condición de normalización $\sum_{j=1}^{n}(z^j)^2 = 1$, las correspondientes componentes principales muestrales, a las que denotaremos $\hat{\underline{\xi}}^{(n)}$, vienen dadas por:

$$\hat{\underline{\xi}}^{(n)} = \int_0^T (X(t) - \bar{X}(t))\hat{f}^{(n)}dt = \sum_{j=1}^{n} z^j(Y_j - \bar{Y}_j).$$

Nota: Dado que $\underline{z}$ es un vector propio de la matriz de covarianzas muestral **R** del vector aleatorio Y cuyas componentes son las variables Y_j de la definición (3.6), de esta última expresión se deduce claramente que las componentes principales $\hat{\underline{\xi}}^{(n)}$ son precisamente las correspondientes componentes principales del vector aleatorio Y.

3.2. Elección del Subespacio de Aproximación para Procesos con Trayectorias Escalonadas

El problema que se plantea finalmente al aplicar el método de proyección es el de la elección del subespacio de aproximación E_n.

En primer lugar, al ser $L^2[0,T]$ un espacio de Hilbert separable, la existencia de una aproximación $\{E_n\}$, en el sentido de la definición (3.4), está garantizada sin más que tomar una base ortonormal $\{e_j\}_{j=1}^{\infty}$ de $L^2[0,T]$ y definir E_n como el subespacio generado por las funciones $\{e_j\}_{j=1}^{n}$.

En segundo lugar, el lema (3.7) demuestra que aplicar el método de proyección ortogonal sobre un subespacio E_n para aproximar el ACP de un proceso es equivalente a estimar directamente el ACP de la proyección del proceso original sobre el subespacio E_n. De acuerdo con esto, la elección del subespacio de aproximación debe estar intimamente relacionada con la naturaleza de las trayectorias del proceso. Así, en el caso de procesos con funciones muestrales regulares una elección óptima es la del subespacio generado por una base ortonormal de funciones trigonométricas en el intervalo de tiempos (Aguilera *et al.* (1993)).

Consideremos como punto de partida en este trabajo, un proceso $\{X(t)\}$ cuyas trayectorias permanecen constantes en intervalos aleatorios. Este es, por ejemplo, el

caso de los procesos puntuales y de recuento. Para procesos de este tipo aproximaremos sus funciones muestrales mediante funciones constantes en los intervalos fijos de una partición previamente elegida en [0,T].

Sea π_n una partición del intervalo [0,T] dada por los nodos:

$$0 = a_0 < a_1 < \cdots < a_n = T$$

verificando:

$$\Delta_n = \text{máx}_{j=1,\ldots,n}\{(a_j - a_{j-1})\} \to 0 \text{ cuando } n \to \infty.$$

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

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

siendo I_j la función indicadora en el intervalo $(a_{j-1}, a_j]$ definida por

$$I_j(t) = \begin{cases} 1 & t \in (a_{j-1}, a_j] \\ 0 & t \notin (a_{j-1}, a_j] \end{cases}$$

La sucesión $\{E_n\}_{n=1}^{\infty}$ definida anteriormente es una aproximación de $L^2[0,T]$ en el sentido de la definición (3.4). Efectivamente, toda función continua es el límite uniforme de una sucesión de funciones $\{f_n\}$ $(f_n \in E_n)$ y, a su vez, el conjunto de las funciones continuas es denso en $L^2[0,T]$. Por lo tanto, está garantizada la aplicación del método de proyección para calcular los factores principales de procesos de esta naturaleza.

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

$$X_k^{(n)}(t) = \sum_{j=1}^{n} < X_k | \delta_j > \delta_j(t) = \sum_{j=1}^{n} M_{kj} I_j(t),$$

siendo M_{kj} el valor medio de la trayectoria muestral k sobre el intervalo $(a_{j-1}, a_j]$ definido por:

$$M_{kj} = (a_j - a_{j-1})^{-1} \int_{a_{j-1}}^{a_j} X_k(t) dt.$$

Como consecuencia inmediata de la proposición (3.8) se tiene que los factores principales muestrales aproximados son funciones escalonadas dadas por:

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

donde $\underline{z}$ es un vector propio normalizado de la matriz $\mathbf{R}$ definida en (3.5) siendo, en este caso, los elementos Y_{kj} $(k = 1,\ldots,N; \quad j = 1,\ldots,n)$ de la forma:

$$Y_{kj} = (a_j - a_{j-1})^{-1/2} \int_{a_{j-1}}^{a_j} X_k(t)dt = (a_j - a_{j-1})^{1/2} M_{kj}.$$

4. APLICACIÓN

El objetivo de esta sección es ilustrar el comportamiento de las aproximaciones propuestas en la sección anterior mediante una aplicación con datos reales.

4.1. Estimación del ACP de la Evolución del Número de Profesores Titulares de Universidad en 1992

En lo que sigue, nos proponemos estimar los factores principales del proceso definido en cada instante t como *'Número de profesores titulares nombrados hasta dicho instante del año 1992 en las distintas universidades españolas'*. Para estimar las componentes principales de este proceso hemos elegido aleatoriamente una muestra de veinte universidades españolas y hemos contado día a día el número de nombramientos de profesores titulares durante 1992 en cada una de ellas. Los datos han sido tomados del Boletín Oficial del Estado.

Claramente, las funciones muestrales de este proceso son funciones constantes a lo largo de intervalos aleatorios cuyos extremos para cada universidad vienen dados por los días en que se producen nuevos nombramientos. Por ello, es conveniente aproximar sus factores principales mediante los de su proyección sobre el subespacio de las funciones constantes sobre los intervalos de una partición previamente fijada.

En nuestra aplicación hemos dividido el año 1992 en meses naturales, de modo que la partición π_n está definida por los nodos: $a_0 = 0$, $a_1 = 31$, $a_2 = 60$, $a_3 = 91$, $a_4 = 121$, $a_5 = 152$, $a_6 = 182$, $a_7 = 213$, $a_8 = 244$, $a_9 = 274$, $a_{10} = 305$, $a_{11} = 335$, $a_{12} = 366$.

En primer lugar, hemos calculado de forma exacta los valores medios del proceso para cada universidad en cada subintervalo. El cálculo numérico de estos valores medios es simple. Por ejemplo, si en un intervalo $(a_{i-1}, a_i]$ una trayectoria $X_k(t)$

cambia de valor una sola vez en el instante a, siendo $X(a_{i-1}) = x_{i-1}$ y $X(a_i) = x_i$, entonces

$$\int_{a_{i-1}}^{a_i} X(t)dt = x_{i-1}(a - a_{i-1}) + x_i(a_i - a).$$

En otro caso, si la función muestral $X_k(t)$ cambia de valor p veces ($p > 1$) en dicho intervalo la generalización es evidente y el cálculo resulta extremadamente rápido incluso para aplicaciones con muchos datos.

En segundo lugar, hemos calculado los autovalores y autofunciones de la matriz de covarianzas **R** definida en (3.5). El valor propio dominante y su vector propio asociado han sido calculados mediante el método de las potencias, y el resto de los autovalores y autovectores mediante el método de deflación de Hotelling y el método de la potencia inversa (ver, por ejemplo, Atkinson y Harley (1983)). El lector interesado en un estudio más completo sobre el problema del cálculo de autovalores y autofunciones puede ver el artículo de Watkins (1993).

Finalmente, el proceso puede ser reconstruido a partir de su representación en componentes principales explicando el porcentaje deseado de su variabilidad total. Una vez obtenida dicha reconstrucción estimaremos el error cometido en cada instante mediante la desviación estándar que se obtiene como la raíz cuadrada del error cuadrático medio en la siguiente forma:

$$DE(t) = (ECM(t))^{1/2} = \left(\frac{1}{N} \sum_{k=1}^{N} (X_k(t) - X_k^m(t))^2 \right)^{1/2}$$

denotando por $X_k^m(t)$ a la reconstrucción de la función muestral k mediante su ACP con las m-primeras componentes principales.

Las varianzas de las componentes principales, (autovalores de la matriz **R**), figuran en la tabla 1 junto al porcentaje de la varianza total explicado por cada una de ellas. A la vista de los resultados podemos observar que el primer factor explica un porcentaje muy elevado de la variabilidad total del proceso (más del 96 %), y que los tres primeros factores acumulan más del 99 %.

La tabla 2 recoge el error estándar mensual cometido al reconstruir el número medio de nombramientos mediante su descomposición en componentes principales truncada en el m-ésimo término (m=1, ...,12). Además, en la tabla 3 hemos representado cada universidad por tres filas. En la primera figura el número medio exacto de nombramientos mensuales acumulados, y en la segunda y tercera filas aparecen las reconstrucciones del valor del proceso en cada mes mediante su ACP con dos y cuatro factores respectivamente.

Tabla 1

Valores Propios y Varianza Explicada por las Componentes Principales

Componentes Principales	Valores Propios	Varianza Explicada	Varianza Acumulada
1	199408.796	96.142%	96.142 %
2	5439.012	2.622%	98.764 %
3	1920.881	0.926%	99.690 %
4	403.674	0.195%	99.885 %
5	149.870	0.072%	99.957 %
6	40.641	0.020%	99.977 %
7	23.759	0.011%	99.988 %
8	12.038	0.006%	99.994 %
9	5.438	0.003%	99.997 %
10	3.185	0.002%	99.998 %
11	2.383	0.001%	99.999 %
12	1.237	0.001%	100.000 %
Varianza Total = 207410.913			

Tabla 2

Error Estándar Mensual en Función del Número de Componentes Principales Utilizadas en la Reconstrucción

Nº C.P	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
1	1.761	6.566	7.275	6.656	5.816	4.268	4.406	2.726	2.531	2.564	3.387	4.088
2	0.961	2.146	1.140	1.070	2.208	4.187	4.109	2.126	1.483	1.860	3.325	3.939
3	0.932	2.099	1.109	1.004	1.782	1.492	0.629	1.291	1.481	1.052	0.869	1.472
4	0.658	1.234	0.539	0.981	0.607	0.529	0.609	0.616	0.831	0.888	0.704	1.135
5	0.380	0.393	0.382	0.334	0.434	0.366	0.485	0.604	0.522	0.588	0.670	0.597
6	0.354	0.279	0.373	0.329	0.428	0.224	0.484	0.464	0.458	0.284	0.364	0.125
7	0.324	0.135	0.211	0.236	0.182	0.217	0.327	0.452	0.280	0.207	0.201	0.124
8	0.288	0.132	0.184	0.229	0.180	0.200	0.142	0.079	0.173	0.204	0.186	0.099
9	0.219	0.126	0.142	0.132	0.162	0.198	0.113	0.064	0.135	0.113	0.080	0.044
10	0.187	0.099	0.123	0.129	0.094	0.093	0.057	0.016	0.065	0.093	0.080	0.040
11	0.036	0.025	0.099	0.088	0.010	0.043	0.026	0.013	0.065	0.091	0.068	0.032
12	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Tabla 3

Número Medio de Nombramientos de Titulares de Universidad Durante 1992 y su Reconstrucción para Varios Porcentajes de Varianza Explicada

1992	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
ALICANTE	1.61	4.79	5.00	5.00	5.61	7.97	9.65	12.00	12.00	12.87	13.33	15.36
98.03%	0.77	2.83	4.85	5.72	6.31	6.85	8.11	10.79	11.71	12.74	14.93	17.12
99.88%	1.27	4.06	5.35	5.44	5.77	7.50	9.78	12.15	12.50	12.49	13.37	15.12
AUT.MADRID	2.65	6.48	8.42	10.00	11.13	12.70	15.74	20.77	21.00	22.26	24.50	28.81
98.03%	1.38	5.45	8.29	10.16	12.43	14.22	16.17	19.82	21.01	22.19	24.79	27.58
99.88%	1.69	6.28	8.85	10.11	11.32	12.82	15.56	20.14	21.65	22.75	25.06	27.71
BARCELONA	0.77	1.41	6.13	22.17	50.32	93.93	117.71	131.48	136.00	136.00	136.00	136.00
98.03%	0.00	0.00	6.33	22.81	50.46	91.61	114.80	129.93	136.11	137.33	138.13	137.92
99.88%	0.28	0.75	6.47	22.50	50.76	93.68	117.46	131.37	136.55	136.53	135.90	135.28
CADIZ	0.00	1.24	3.71	4.60	5.42	6.00	7.97	11.65	12.27	13.00	14.43	16.39
98.03%	0.48	1.59	3.41	4.38	5.16	6.54	8.25	10.93	11.91	12.89	14.85	16.70
99.88%	0.55	1.78	3.52	4.35	4.98	6.40	8.28	11.06	12.04	12.93	14.78	16.58
CANTABRIA	1.03	2.86	3.39	4.13	5.32	7.20	9.52	14.71	16.00	16.00	17.50	21.00
98.03%	0.44	1.46	3.38	4.74	6.20	8.64	10.92	13.91	15.02	16.00	17.92	19.69
99.88%	0.75	2.28	3.96	4.70	5.02	7.04	10.13	14.16	15.66	16.63	18.34	19.98
COMP.MADRID	7.23	30.72	39.55	47.13	60.10	64.20	67.29	76.97	78.80	80.00	87.17	98.07
98.03%	7.29	30.67	40.09	47.69	59.37	63.32	66.62	76.46	78.99	81.50	88.18	97.21
99.88%	7.16	30.30	39.79	47.69	60.04	64.36	67.27	76.44	78.68	81.09	87.77	96.84
CORDOBA	0.10	1.17	2.42	4.37	6.00	6.00	6.26	8.29	9.00	9.00	9.57	10.52
98.03%	0.50	1.67	3.36	3.87	3.87	4.08	5.16	7.47	8.30	9.26	11.28	13.19
99.88%	0.46	1.49	3.00	3.73	4.92	6.38	7.16	8.06	8.05	8.36	9.78	11.57
BALEARES	0.00	0.00	0.48	1.00	1.58	2.50	3.00	3.65	4.00	4.00	4.00	6.00
98.03%	0.08	0.00	1.07	1.11	0.34	0.25	1.15	2.98	3.69	4.55	6.28	7.75
99.88%	0.12	0.00	0.81	0.94	1.25	2.53	3.32	3.74	3.58	3.66	4.61	5.90
LAGUNA	0.55	4.03	10.07	17.90	33.48	47.60	51.52	52.00	52.00	52.00	52.23	53.42
98.03%	1.83	7.50	11.96	17.75	26.75	37.38	44.13	51.07	53.52	54.88	57.61	60.56
99.88%	1.02	5.11	9.56	17.41	32.62	47.96	52.01	52.29	51.32	50.70	52.13	55.01
LEON	0.00	0.10	1.00	1.23	2.97	5.00	5.00	5.00	5.00	6.42	11.60	13.23
98.03%	0.20	0.42	1.86	2.32	2.24	2.92	4.24	6.44	7.27	8.17	9.97	11.55
99.88%	0.00	0.00	0.90	2.52	3.92	4.37	4.17	5.35	6.06	7.59	10.41	12.40
MALAGA	1.42	6.83	9.20	10.00	10.87	12.97	17.74	19.00	19.00	19.77	22.20	25.71
98.03%	1.44	5.71	8.53	10.18	12.07	13.20	14.79	18.29	19.40	20.58	23.24	26.13
99.88%	1.81	6.61	8.90	9.98	11.66	13.63	15.98	19.27	19.99	20.42	22.12	24.69
MURCIA	0.00	2.79	4.97	8.47	14.26	21.47	24.84	27.23	28.93	29.84	33.20	35.77
98.03%	0.79	3.01	5.79	8.90	13.39	19.49	23.77	28.28	29.94	31.03	33.12	35.13
99.88%	0.36	1.89	5.06	8.99	14.79	21.14	24.34	27.76	29.08	30.37	32.94	35.18
U.N.E.D.	1.10	2.17	3.00	3.43	4.00	4.00	4.45	9.00	9.00	11.97	18.57	21.19
98.03%	0.52	1.78	3.61	4.49	5.13	6.22	7.78	10.40	11.35	12.33	14.34	16.24
99.88%	0.00	0.56	3.38	4.92	4.77	3.38	4.05	8.34	10.68	13.44	17.47	19.98
OVIEDO	5.29	18.24	19.00	19.07	21.39	27.57	32.45	37.26	38.00	38.65	39.57	40.00
98.03%	3.04	12.52	17.15	20.47	25.10	27.08	29.20	34.45	35.97	37.51	41.26	45.82
99.88%	4.79	16.90	19.29	19.70	21.98	26.44	32.38	38.34	38.97	37.81	37.82	41.05
PAIS VASCO	1.07	5.10	9.55	13.57	19.10	24.47	33.55	46.07	54.20	57.52	61.33	64.41
98.03%	1.02	4.11	7.94	13.72	22.79	35.06	42.67	49.40	51.93	53.13	55.26	57.10
99.88%	1.21	4.95	9.57	14.32	18.07	24.79	33.73	46.83	53.06	57.17	61.91	64.48
LAS PALMAS	2.19	5.45	7.71	8.27	9.00	9.00	9.00	9.00	9.00	9.00	9.00	10.45
98.03%	1.40	5.45	7.79	8.09	7.63	5.46	5.27	7.65	8.32	9.46	12.15	15.11
99.88%	1.72	6.13	7.68	7.69	8.65	9.06	9.25	9.46	8.58	8.05	8.95	11.41
POL.CATALUNA	0.00	2.72	11.13	14.80	16.00	20.83	27.65	36.90	39.00	39.52	40.00	40.00
98.03%	1.22	4.87	8.21	12.03	17.69	24.65	29.40	34.59	36.45	37.64	40.02	42.46
99.88%	1.91	6.68	9.38	11.89	15.44	22.03	28.50	35.43	37.82	38.68	40.29	42.37
SALAMANCA	1.23	4.72	7.74	10.10	14.00	17.00	17.55	20.32	21.00	21.32	24.60	29.77
98.03%	1.42	5.63	8.53	10.48	12.90	14.83	16.85	20.59	21.80	22.99	25.61	28.45
99.88%	0.97	4.44	7.74	10.57	14.44	16.69	17.57	20.08	20.89	22.25	25.34	28.40
VALENCIA	0.55	5.24	13.52	22.03	26.87	27.50	28.65	39.10	45.13	53.84	62.13	65.19
98.03%	2.42	9.99	14.66	19.59	27.09	34.38	39.34	45.76	47.87	49.33	52.54	56.25
99.88%	0.64	5.79	13.54	20.90	26.92	27.23	29.00	39.50	45.42	52.10	61.43	67.02
VIGO	0.39	2.17	7.10	9.00	9.00	9.50	12.23	15.23	18.73	19.16	21.50	24.23
98.03%	1.00	3.83	6.25	7.78	9.51	11.23	13.15	16.42	17.54	18.63	20.96	23.36
99.88%	0.92	3.67	6.33	7.91	9.10	9.98	11.83	15.87	17.50	19.12	22.01	24.55

4.2. Discusión y conclusiones

En primer lugar, observemos que la primera impresión que produce la tabla 2 es de cierta sorpresa porque a pesar de que la primera componente principal explica más del 96% de la variabilidad total, el error cometido al reconstruir el proceso con dicha componente no es excesivamente pequeño especialmente en la primera mitad del año. Además, el error de reconstrucción se reduce considerablemente al ir introduciendo en el modelo componentes principales ruidosas que explican muy poco de la variabilidad total.

Estos resultados son fácilmente explicables debido a que para procesos con varianza creciente, como éste y en general para el caso de procesos con incrementos ortogonales, el ACP da más peso a los últimos instantes del intervalo [0,T] porque son las variables con mayor varianza las que contribuyen en mayor medida a la determinación de los factores. En estos casos es usual obtener una primera componente principal relativamente ligada a la evolución del proceso al final del intervalo y que explica casi la totalidad de su varianza. Una solución a este problema sería llevar a cabo un ACP del proceso tipificado que permita eliminar factores triviales.

Ahora bien, en el caso de procesos con funciones muestrales escalonadas, como ejemplo más conocido citaremos el proceso de Poisson, su tipificación llevaría a un nuevo proceso estocástico cuyas trayectorias no serían escalonadas (recordemos que la varianza del proceso de Poisson homogéneo es de la forma λt siendo λ el número medio de ocurrencias por unidad de tiempo). Por lo tanto, la tipificación altera la base del subespacio sobre el que se proyecta el proceso que tendría que elegirse en función a la nueva naturaleza de las trayectorias. Buscando una vía alternativa para la solución de este problema, hemos obtenido una representación numerable del proceso en función de las componentes principales de la tipifiación del vector aleatorio Y (definido en (3.4)) y actualmente estamos investigando su utilidad para eliminar factores triviales así como el error de truncatura.

Volviendo a las tablas 2 y 3 observemos que en las reconstrucciones con cuatro o más componentes principales el error es siempre muy pequeño. En particular, la reconstrucción mensual del proceso con los doce factores principales aproximados coincide exactamente con el valor medio mensual lo que permite concluir que la precisión de los procedimientos de cálculo de autovalores y autovectores así como la del resto de procedimientos computacionales utilizados es extremadamente alta.

Por lo tanto, el método de proyección en este caso da resultados óptimos, más aún si tenemos en cuenta que la partición elegida no ha sido demasiado fina. Así, hemos realizado un estudio simultáneo discretizando el intervalo cada quince días en el que hemos obtenido el mismo tipo de componentes principales en cuanto al porcentaje de variabilidad explicada. Sin embargo, al ser los subintervalos más pequeños el proceso

reconstruido explica mejor la naturaleza del proceso original que suele cambiar de valor en las dos quincenas de cada mes generalmente. De hecho, la elección de la partición estará en función de la densidad de cambios de valor del proceso en el tiempo observado de modo que los subintervalos no tienen que ser de la misma amplitud y serán más finos en aquellos periodos en los que se produzcan más cambios.

En el trabajo de Aguilera *et al.* (1993) ha sido aplicado el método de proyeccción para estimar el ACP del grado de ocupación hotelera en las ciudades españolas durante el año 1990. Como las trayectorias de este proceso son regulares se ha elegido como subespacio óptimo el generado por un número finito de funciones trigonométricas normalizadas en el periodo considerado. Los resultados obtenidos corroboran la eficacia del método de proyección para aproximar los factores principales muestrales de un proceso de segundo orden.

Concluimos finalmente que en el caso en que se dispone de información a priori sobre las funciones muestrales que garantice la elección óptima del subespacio de aproximación, el método de proyección puede ser idóneo para aproximar el ACP del proceso.

NOTA

Para estimar el ACP de un proceso con trayectorias escalonadas aplicando el método de proyección ortogonal, hemos desarrollado un algoritmo computacional que ha sido codificado en Turbo Pascal versión 6.0. usando programación orientada a objeto. Se trata del programa PCAP que permite, además, aproximar el ACP de un proceso estocástico a partir de observaciones discretas de sus funciones muestrales mediante tres métodos numéricos diferentes (Aguilera *et al.* (1994)). Los investigadores interesados pueden conseguir una copia del programa contactando con alguno de los autores.

AGRADECIMIENTOS

Este trabajo de investigación ha sido financiado en parte por el Proyecto Nº HF94-229 del Programa de Acciones Integradas Hispano-Francesas de la DGICYT, Ministerio de Educación y Ciencia, España.

REFERENCIAS

[1] **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 Sociedad Chilena de Estadística*, **9(2)**, 57–76.

[2] **Aguilera, A.M., Piñar, M.A.** y **Del Moral, M.J.** (1993). "On the Empirical Behaviour of a Stochastic Process". *Proceedings of the Sixth International Symposium on Applied Stochastic Models and Data Analysis*, Vol. **1**, 5–16, J. Janssen y C.H. Skiadas (eds.), World Scientific.

[3] **Aguilera, A.M., Ocaña F.A.** y **Valderrama, M.J.** (1994). "A computational Algorithm for PCA of Random Processes". *Proceedings of the Eleventh Symposium in Computational Statistics*, Software Descriptions, 39–40, R. Dutter y W. Grossmann (eds.), Physica-Verlag.

[4] **Atkinson, L.V.** y **Harley, P.J.** (1983). *An introduction to numerical methods with Pascal*. Addison-Wesley.

[5] **Baker, C.T.H.** (1977). *The Numerical Treatment of Integral Equations*. Oxford University Press.

[6] **Church, A.** (1966). "Analysis of Data When the Response Is a Curve". *Technometrics*, **8**, 229–246.

[7] **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.

[8] **Deville, J.C.** (1973). "Estimation of the Eigenvalues and of the Eigenvectors of a Covariance Operator". *Note interne de L'INSEE*.

[9] **Deville, J.C.** (1974). "Méthodes Statistiques et Numériques de l'Analyse Harmonique". *Annales de L'INSEE*, **15**, 3–101.

[10] **Fukunga, K.** (1990). *Introduction to Statistical Pattern Recognition*. Academic Press.

[11] **Gutiérrez, R., Ruiz J.C.** y **Valderrama M.J.** (1992). "On the Numerical Expansion of a Second Order Stochastic Process". *Applied Stochastic Models and Data Analysis*, **8(2)**, 67–77.

[12] **Pardoux, C.** (1989). "Apport de l'Analyse Factorielle a l'Etude d'un Processus". *Revue Statistique Appliquée*, **XXXVII(4)**, 41–60.

[13] **Riesz, F.** y **Sz-Nagy, B.** (1990). *Lecons d'Analyse Fonctionnelle*. Ediciones Jacques Gabay, reimpresión de la 3^a edición publicada por Gauthier-Villars y Akadémiai Kiadó en 1955.

[14] **Shorack, G.R.** y **Wellner, J.A.** (1986). *Empirical Processes with Applications to Statistics*. Wiley.

[15] **Watkins, D.S.** (1993). "Some Perspectives on the Eigenvalue Problem". *SIAM Rev.*, **35(3)**, 430–471.

ENGLISH SUMMARY:

PRINCIPAL COMPONENT ANALYSIS OF A STOCHASTIC PROCESS WHOSE SAMPLE PATHS ARE PIECEWISE CONSTANT FUNCTIONS

Ana Mª Aguilera Del Pino
Francisco A. Ocaña Lara
Mariano J. Valderrama Bonnet

1. INTRODUCTION

In many applied contexts, we observe a continuous function for each unit of observation rather than a vector as in the classic multivariate case. That is, the data are the sample paths of a continuous time stochastic process. Church (1966) and Deville (1974) proposed an extension of the classic principal component analysis (PCA) to reduce dimension in the case of continuous processes. Later, Dauxois *et al.* (1982) have set up the asymptotic theory for the PCA of a Gaussian random process. The sample principal factors are the solutions of a second kind integral equation whose kernel is the sample covariance of the process. The objective of this paper consists of approximating the solutions of this equation when the sample functions of the process are piecewise constant functions.

2. PCA OF A STOCHASTIC PROCESS

Let us consider a second order stochastic process $\{X(t) : t \in [0,T]\}$, mean square continuous, and whose sample functions belong to the Hilbert space $L^2[0,T]$ of square integrable functions on [0,T], with the natural inner product defined by:

$$< f \mid g >= \int_0^T f(t)g(t)dt \quad \forall f,g \in L^2[0,T].$$

2.1. Basic Theory

Under these regularity conditions, the ith principal component of $\{X(t)\}$ is a random variable given by the formula

$$\xi_i = \int_0^T f_i(t)(X(t) - \mu(t))dt,$$

where f_i, called the ith principal factor, is the normalized eigenfunction corresponding to the ith largest eigenvalue λ_i of the covariance operator C associated to the process. Then, PCA gives the following orthogonal decomposition of the process (see, e.g., Shorack and Wellner (1986), p.120):

$$X(t) - \mu(t) = \sum_i \xi_i f_i(t)$$

where μ is the mean function of the process. Moreover, this series truncated in the mth term is the best m-dimensional model of $\{X(t)\}$ in the least squares sense, being $\sum_{i=m+1}^{\infty} \lambda_i$ the minimum mean squared error.

2.1. Sample Estimation

Given N independent sample paths, $\{X_k(t) : t \in [0,T], k = 1,2,\ldots,N\}$, the eigensystem (λ_i, f_i) of C is estimated by the corresponding one $(\hat{\lambda}_i, \hat{f}_i)$ of its usual unbiased estimator $\hat{C}$. This estimator is defined as a random operator whose kernel is the sample covariance function of the process given by

$$\hat{C}(t,s) = \frac{1}{N-1} \sum_{k=1}^{N} \left(X_k(t) - \bar{X}(t)\right)\left(X_k(s) - \bar{X}(s)\right),$$

where $\bar{X}$ is the natural estimate of the mean μ. For more details about the properties of these estimators, the reader is referred to Deville (1973).

3. APPROXIMATED SOLUTION OF PCA

The sample principal factors $\hat{f}_i$ are the eigenfunctions of the sample covariance operator. That is, they are solutions of the following second kind integral equation:

$$\hat{C}\left(\hat{f}(t)\right) = \int_0^T \hat{C}(t,s)\hat{f}(s)ds = \hat{\lambda}\hat{f}(t).$$

Due to inherent difficulties in obtaining analytical solutions for such equations, our objective is to apply an efficient numerical procedure to approximate the principal factors of random processes whose sample paths are piecewise constant functions.

3.1. Orthogonal Projection Method

When we have information about the nature of the process sample paths it is suitable to apply the orthogonal projection method to solve approximately the last equation.

Let $\{E_n\}_{n=1}^{\infty}$ be a sequence of finite-dimensional subspaces of $L^2[0,T]$ such that:

$$\lim_{n\to\infty} P_n x = x \quad \forall x \in E,$$

where $\{P_n\}_{n=1}^{\infty}$ is the sequence of orthogonal projections from $L^2[0,T]$ to E_n. Then, the orthogonal projection method consists in approximating the eigensystem of the sample covariance operator $\hat{C}$ by means of the eigensystem of the operator $\hat{C}^{(n)} : L^2[0,T] \to E_n$ defined by

$$\hat{C}^{(n)} = P_n \hat{C} P_n.$$

As $\hat{C}$ is a positive, self-adjoint and compact operator on $L^2[0,T]$, the convergence in norm of the approximated solutions is warranted (Riesz and Sz-Nagy (1990)).

It can be proved (Deville (1974)) that the approximated principal factors are those of a stochastic process whose sample paths are the projection of the original sample functions over the subspace E_n.

Finally, the approximated solutions, denoted as $\hat{f}^{(n)}$, are given by (Aguilera *et al.* (1993)):

$$\hat{f}^{(n)} = \sum_{j=1}^{n} z^j e_j$$

being the vector $\underline{\mathbf{z}}$ with components z^j a solution to the matrix equation

$$R\underline{z} = \hat{\lambda}^{(n)}\underline{z},$$

where $\mathbf{R}$ is a $n \times n$ matrix with elements

$$R_{ij} = < \hat{C}e_i \mid e_j > = \int_0^T \int_0^T \hat{C}(t,s)e_i(t)e_j(s)dtds = \frac{1}{N-1}\sum_{k=1}^{N}(Y_{ki} - \bar{Y}_i)(Y_{kj} - \bar{Y}_j)$$

and defining,

$$Y_{kj} = \int_0^T X_k(s)e_j(s)ds, \quad \forall j = 1 \ldots n,$$

$$\bar{Y}_j = \frac{1}{N}\sum_{k=1}^{N} Y_{kj} = \int_0^T \bar{X}(s)e_j(s)ds.$$

3.2. Choosing the Approximative Subspace for Processes with Piecewise Constant Sample Functions

We are now going to choose the most suitable subspace E_n in $L^2[0,T]$ to approximate the principal factors of a random process whose sample paths are piecewise constant functions.

Given a partitioning π_n in the interval [0,T] with knots:

$$0 = a_0 < a_1 < \cdots < a_n = T$$

verifying:

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

we will take the subspace E_n of the piecewise constant functions over the subintervals $(a_{j-1}, a_j]$ $(j = 1,\ldots,n)$. An orthonormal basis of E_n is given by the functions:

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

defining I_j as

$$I_j(t) = \begin{cases} 1 & t \in (a_{j-1}, a_j] \\ 0 & t \notin (a_{j-1}, a_j] \end{cases}$$

4. APPLICATION

In this section we include an application with real data by means of the construction of a computational algorithm which allows to reconstruct the data after estimating their principal components.